Computer Science & Information Technology 64

Jae-Kwang Lee
Brajesh Kumar Kaushik (Eds)

Computer Science & Information Technology

Seventh International Conference on Computer Science, Engineering and Applications (CCSEA 2017)
Dubai, UAE, January 28~29, 2017

AIRCC Publishing Corporation

Volume Editors

Jae-Kwang Lee,
Hannam University, South Korea
E-mail: jklee@hnu.kr

Brajesh Kumar K,
IIT-Roorkee, India
E-mail: bkkaushik23@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-62-5
DOI : 10.5121/csit.2017.70201 - 10.5121/csit.2017.70214

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Seventh International Conference on Computer Science, Engineering and Applications (CCSEA 2017) was held in Dubai, UAE, during January 28~29, 2017. The Third International Conference on Artificial Intelligence and Applications (AIFU 2017), The Fifth International Conference on Data Mining & Knowledge Management Process (DKMP 2017), The Sixth International Conference on Cloud Computing: Services and Architecture (CLOUD 2017), The Sixth International Conference on Embedded Systems and Applications (EMSA 2017), The Sixth International Conference on Software Engineering and Applications (SEA 2017) and The Third International Conference on Signal and Image Processing (SIPRO 2017) was collocated with The Seventh International Conference on Computer Science, Engineering and Applications (CCSEA 2017). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCSEA-2017, AIFU-2017, DKMP-2017, CLOUD-2017, EMSA-2017, SEA-2017, SIPRO-2017 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CCSEA-2017, AIFU-2017, DKMP-2017, CLOUD-2017, EMSA-2017, SEA-2017, SIPRO-2017 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCSEA-2017, AIFU-2017, DKMP-2017, CLOUD-2017, EMSA-2017, SEA-2017, SIPRO-2017 .

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Jae-Kwang Lee
Brajesh Kumar Kaushik

Organization

General Chair

Natarajan Meghanathan,	Jackson State University, USA
Brajesh Kumar Kaushik,	Indian Institute of Technology - Roorkee, India

Program Committee Members

Ahmad Mani	Tarbiat modares University, Iran
Ahmad Rawashdeh	University of Central Missouri, USA
Ahmed Korichi	University of Ouargla, Algeria
Akram Abdelqader	AL-Zaytoonah University of Jordan, Jordan
Alaa Hamami	Princess Sumaya University for Technology, Jordan
Alfredo Cuzzocrea	University of Trieste, Italy
Amiya Kumar Tripathy	Edith Cowan University, Australia
Atallah Mahmoud AL-Shatnawi	Al al-Byte University, Jordan
Ayman EL-SAYED	Menoufia University, Egypt
Azeddine Chikh	University of Tlemcen, Algeria
Basar Oztaysi	Istanbul Technical University, Turkey
Bing Zhou	Sam Houston State University, USA
Christophe NICOLLE	University of Bourgogne Franche, France
Da Yan	The University of Alabama at Birmingham, USA
Dabin Ding	University of Central Missouri, USA
Elaheh Pourabbas	National Research Council, Italy
Emad Awada	Applied Science University, Jordan
Farhad pourfarzi	Ardabil University of Medical Sciences, Iran
Figen Balo	Firat University, Turkey
Fulvia Pennoni	University of Milano-Bicocca, Italy
Gábor Kiss	Óbuda University, Hungary
Gamini Wijayarathna	University of Kelaniya, Sri Lanka
Gheorghi Guzun	The University of Iowa, USA
Guoqing Xiao	Hunan University, China
Hamid Alasadi	Basra University, Iraq
Hayet Mouss	Batna Univeristy, Algeria
Hongzhi	Harbin Institute of Technology, China
Houda KHROUF	Atos Innovation Lab, France
Irena Patašienė	Kaunas University of Technology, Italy
Ishfaq Ahmad	The University of Texas at Arlington, U.S.A
Ivo Pierozzi Junior	Embrapa Agricultural Informatics, Brazil
Jafar Mansouri	Ferdowsi University of Mashhad, Iran
Jamal El Abbadi	Mohammadia V University Rabat, Morocco
Jasmine Seng K. P	Charles Sturt University, Australia
Jia Zhu	South China Normal University, China
Jun Liu	University of Michigan at Dearborn, USA
Jun Zhang	South China University of Technology, China

Karina Gibert	Universitat politecnica de catalunya, Spain
Madjid Tavana	La Salle University, USA
Mamoun Alazab	Macquarie University, Australia
Manuel Angel Serrano Martin	Universidad de Castilla, Spain
Mária Pomffyová	Matej Bel University, Slovakia
Mario Henrique Souza Pardo	University of São Paulo, Brazil
Marius CIOCA	Lucian Blaga University of Sibiu, Romania
Maryam Habibi	Humboldt-Universität zu Berlin, Germany
Maryam Hajakbari	Islamic Azad University, Iran
Mohamed BEN AOUICHA	Faculty of Sciences, Tunisia
Mohamedmaher Benismail	King saud University, Saudi Arabia
Mohammad Rawashdeh	University of Central Missouri, United States
Mohammad Shokoohi-Yekta	Stanford university, USA
Mohammed Al-Mai'itah	Al-Balqa applied university, Jordan
Mohammed Al-Sarem	Taibah University, KSA
Mohammed J. Zaki	Rensselaer Polytechnic Institute, Troy
Mostafa Ashry	Alexandria University, Egypt
Mourchid Mohammed Ibn	Tofail University Kenitra, Morocco
Nasser Thabet	Szabist University, UAE
Noura Taleb	Badji Mokhtar University, Algeria
Oded Maimon	Tel Aviv university, Israel
Ouafa Mah	Ouargla university, Algeria
Peter Kieseberg	SBA Research, Austria
Prakash Duraisamy	University of Central Missouri, United States
Razieh malekhoseini	Islamic Azad University, Iran
Roberto De Virgilio	Roma Tre University, Italy
Roberto Saia	University of Cagliari, Italy
ŞABAN GÜLCÜ	Necmettin Erbakan University, Turkey
Saoussen Bel Hadj Kacem	Carthage University , Tunisia
Sarasi Lalithsena	Wright State University, USA
Sarvenaz Choobdar	University of Lausanne, Switzerland
Sharma Chakravarthy	UT ARlington, USA
Shoeib Faraj	Institute of Higher Education of Miaad, Iran
Shouyang Wang	Chinese Academy of Sciences, China
Sugam K Sharma	Iowa State University, USA
Taeghyun Kang	University of Central Missouri, United States
Tegjyot Singh Sethi	University of Louisville, USA
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Wang Xiangyang	Liaoning Normal University, China
Xiangjie Kong	Dalian University of Technology, China
Xiaoyong Li	National University of Defense Technology, China
Xuechao Li	Auburn University, USA
Yashar Deldjoo	Politecnico di Milano, Italy
Yousef J. Al-Houmaily	Institute of Public Administration, Saudi Arabia
Yu Zhuang	Texas Tech University, USA

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Database Management Systems Community (DBMSC)

Software Engineering & Security Community (SESC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Seventh International Conference on Computer Science, Engineering and Applications (CCSEA 2017)

Routing and Tracking System for Buses…………………….………………...... 01 - 09
Ahmed Ahmed, Elshaimaa Nada and Wafaa Al-Mutiri

Predicting Venues in Location Based Social Network………………...……...... 11 - 21
Omar F.Almallah and Songül Albayrak

Checking Behavioural Compatibility in Service Composition with Graph Transformation……………………………………………...…...……...... 149 - 162
Redouane Nouara and Allaoua Chaoui

Third International Conference on Artificial Intelligence and Applications (AIFU 2017)

Model for Heuristic and AI Planning Strategies - A Path to Leadership in Sector "Telecommunications" in Bulgaria…………………….………………...... 23 - 35
Miglena Temelkova, Dimitar Radev and Strahil Sokolov

Human Emotion Estimation from EEG and Face Using Statistical Features and SVM…………………….……………………………………………………...... 37 - 47
Strahil Sokolov, Yuliyan Velchev, Svetla Radeva and Dimitar Radev

Data Sharing Taxonomy Records for Security Conservation…………...…..…... 49 - 58
Rajeswari Chandrasekaranl and Chandrasekaran Nammalwar

Fifth International Conference on Data Mining & Knowledge Management Process (DKMP 2017)

Predicting Popularity of Korean Contents in Arab Countries Using a Data Mining Technique…………………….………………………………….………...... 59 - 68
Park Young Eun, Soumaya Chaffar, Kim Myoung Sook and Ko Hye Young

Fast Algorithms for Unsupervised Learning in Large Data Sets……….…….... 69 - 76
Syed Quddus

Text Extraction from Raster Maps Using Color Space Quantization……….…. 77 - 86
Sanaz Hadipour Abkenar and Alireza Ahmadyfard

Sixth International Conference on Cloud Computing: Services and Architecture (CLOUD 2017)

EHR Attribute-Based Access Control (ABAC) for Fog Computing Environment........ 87 - 95
Aisha Mohmmed Alshiky, Seyed M. Buhari and Ahmed Barnawi

Sixth International Conference on Embedded Systems and Applications (EMSA 2017)

Modelling and Application of a Computer-Controlled Liquid Level Tank System........ 97 - 106
Hayati Mamur, Ismail Atacak, Fatih Korkmaz and M.R.A. Bhuiyan

Sixth International Conference on Software Engineering and Applications (SEA 2017)

Development of a Location-Based Approaching Notification System Using Android Platform........ 107 - 119
Hisham AlMajed and Abdelaziz Khamis

Third International Conference on Signal and Image Processing (SIPRO 2017)

Resolving Cyclic Ambiguities and Increasing Accuracy and Resolution in DOA Estimation Using Array Rotation........ 121 - 138
Abdelhamid Djouadi and NebojsaI. Jaksic

Single Image Super Resolution : A Comparative Study........ 139 - 147
Aliaa Youssef, Sameh Zarif and Amr Ghoneim

ROUTING AND TRACKING SYSTEM FOR BUSES

Ahmed Ahmed[1], Elshaimaa Nada[2] and Wafaa Al-Mutiri[3]

[1,2]Department of Computer Systems,
Zagazig UniversityUniversity, Zagazig, Egypt
[1,2]Department of Computer Engineering,
Taibahu University, Almadina, Saudi Arabia
[3]Department of Information Systems,
Taibahu University, Almadina, Saudi Arabia

ABSTRACT

This paper proposes development of an android app to improve the transportation services for bus rental companies that lift Taibah University students. It intends to reduce the waiting time for bus students, thereby to stimulate sharing of updated information between the bus drivers and students. The application can run only on android devices. It would inform the students about the exact time of arrival and departure of buses on route. This proposed app would specifically be used by students and drivers of Taibah University. Any change in the scheduled movement of the buses would be updated in the software. Regular alerts would be sent in case of delays or cancelation of buses. Bus locations and routes are shown on dynamic maps using Google maps. The application is designed and tested where the users assured that the application gives the real time service and it is very helpful for them.

KEYWORDS

Bus Routing, Bus Tracking, Google Maps, GIS

1. INTRODUCTION

Android is becoming very popular because the source code is completely free; also, Android is highly suitable for expansion as the developer see fit, so building a mobile application for Android devices is very common these days due to the mentioned reasons [1].

A geographic information system (GIS) is a machine framework intended to catch, store, control, investigate, oversee, and show different kinds of spatial or land information. The acronym GIS is in some cases utilized for geographical information science [2]. GIS platform is used to constitute a map of route net, beside the user's interface based on VB. Net to use Desktop, assisting individuals who use buses' that moves inside the city to be familiar with buses' routes and their actual sites, in addition to data about out-of-work buses[2].

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 01– 09, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70201

GIS can relate disconnected data by utilizing area as the key record variable. Areas may be recorded as x, y, and z directions. All Earth-based spatial–temporal area and degree references ought to, preferably, be relatable to each other and at last to a "true" physical area or degree. This feature of GIS has started to open new streets of exploratory. It gives a pursuing time technique of buses, assisting the management of traffic sufficiently and solving emergency situations. It is available a as web application for passengers to be familiar with routes' conditions and buses' routes [2].

Location Based Services (LBS) denotes applications integrating geographic location to provide a certain service. Examples of such applications include emergency services, car navigation systems or tourist tour planning information delivery[4]. Location services are mainly used in three areas: military and government industries, emergency services and the commercial sector[4]. Analysts and researchers classified the LBS services as the following [4]:

- Person oriented: Position the location of the user or use the position of the user to enhance a service for example, friend finder application. The located person can control the device.

- Device oriented: Used to locate objects for example a group of people or a car. The located person or object cannot control the service.

Also the location based services could be divided into [4]:

- Push services: The user receives information according to his/her location without requesting the information, for example receiving a welcoming message when entering a new town.

- Pull services: In the contrast, the user pull information from the network, for example finding the nearest ATM machine.

This paper presents an application for intelligent mobile devices, mainly GPS. Several methods are proposed to reduce the wastage of time of students waiting for a bus to arrive; we proposed GPS based Bus Tracking and Monitoring system in which the tracking is done by implementing maps with GPS facility. The Android Application is designed for students where they can access/view the daily timetable of bus, bus route, location of bus, and bus arrival and delay timing information. Our main focus is to provide the student with such a system which will for sure reduce the waiting time and will provide the student with all necessary details regarding the arrival time of the bus, its exact location and expected waiting time.

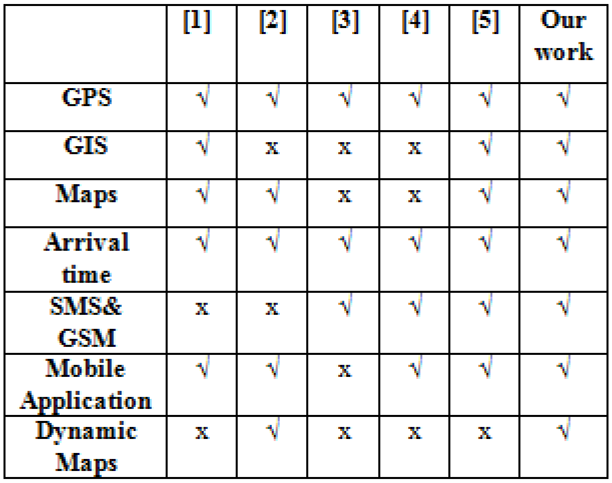
In the next section, a review of the relevant work is presented. The main characteristics of the application are presented in section 3. The results are in section 4. The conclusion and the future work is in section 5.

2. REVIEW OF RELEVANT WORK

In table 1, a comparison between related works and our proposed mobile application is presented. We compared all works against most popular features as GIS, GPS, Maps and whether or not the applications calculate the arrival time.

The proposed mobile application is used for solving many problems starting from the students waiting for the bus. The mobile application will be useful to know next stops and students can monitor the location of the bus on the map. The application is very helpful to find that either student missed the bus or it is late due to traffic. The chance of missing the bus by students will be reduced. This application will save the time and iterations of bus wait will definitely lessen by this app. GIS and GPS technologies introduced this application Taibah University is going to use this application for their students and providing them best transport services. This system stores all operations done in a day and helps the student and management of transport services. University using this app for better management and quick transport services for students.

Table 1. Comparison between related work and our work



	[1]	[2]	[3]	[4]	[5]	Our work
GPS	√	√	√	√	√	√
GIS	√	x	x	x	√	√
Maps	√	√	x	x	√	√
Arrival time	√	√	√	√	√	√
SMS& GSM	x	x	√	√	√	√
Mobile Application	√	√	x	√	√	√
Dynamic Maps	x	√	x	x	x	√

3. SYSTEM MAIN CHARACTERISTICS

The system architecture is presented in figure 1, where the main characteristics of the presented application are:

3.1 Dynamic Map

By dynamic map, we mean that the map will be updated automatically for students and drivers which will facilitate to find out the students' exact positions and also determine each bus stop.

3.2 Geodatabase

Geodatabase has many features specific to databases and the most important of the total dispensed from the other database programs and set up tables inside, tables can be presented on the map and also the student's name .

3.3 Dijkstra's algorithm

It is a solution to the single-source shortest path problem in graph theory. However, it is about as computationally expensive to calculate the shortest path from vertex u to every vertex using Dijkstra’s as it is to calculate the shortest path to some particular vertex v. Therefore, anytime we want to know the optimal path to some other vertex from a determined origin, we will use Dijkstra's algorithm feature that is in google maps platform.

3.4 Notification System

In the application, a notification system is applied which is more convenient between the server and driver and passengers. As the application is a mobile application then sending notifications is a very practical way.

The system architecture which describes the structure and overall design of a system is presented in figure 1.

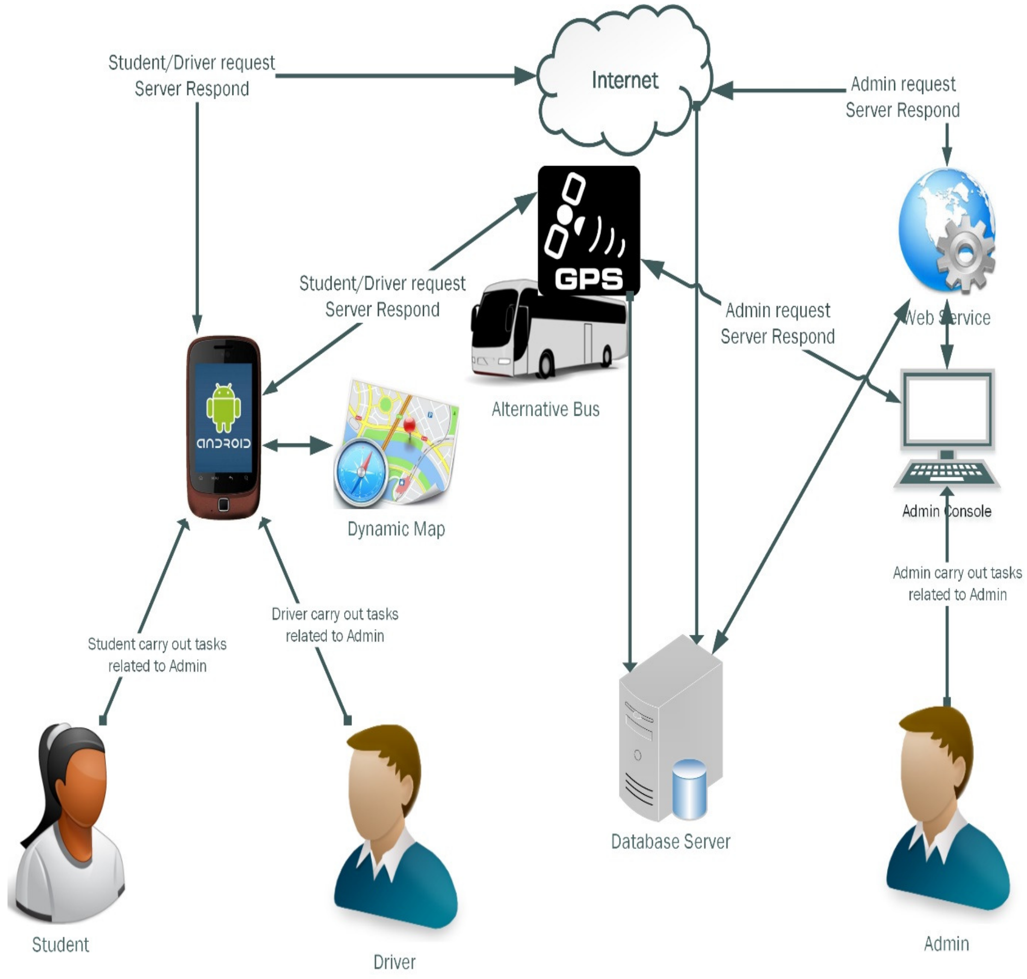


Figure.1 System Architecture

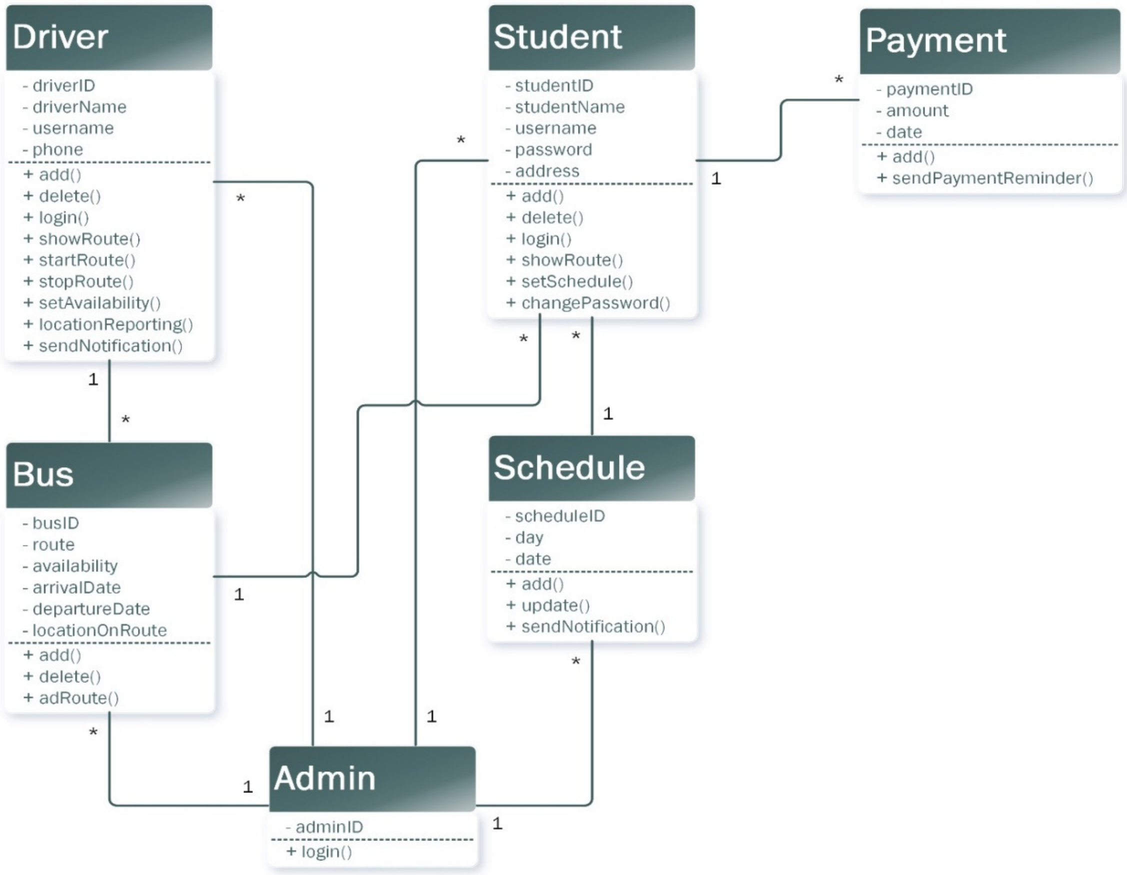


Figure 2 System Class diagram

A class diagram of the system is presented in figure 2. The admin has a login username and password to log to the web portal to carry out his tasks. The admin can control all the tasks from his screen. The driver can set the availability of the bus weather the bus is available or not, if the bus is not available a notification will be sent to the concerned students. The timetable of each student will be entered, a notification will be sent and route will be updated accordingly if there are any changes in the presented schedule.

4. RESULTS

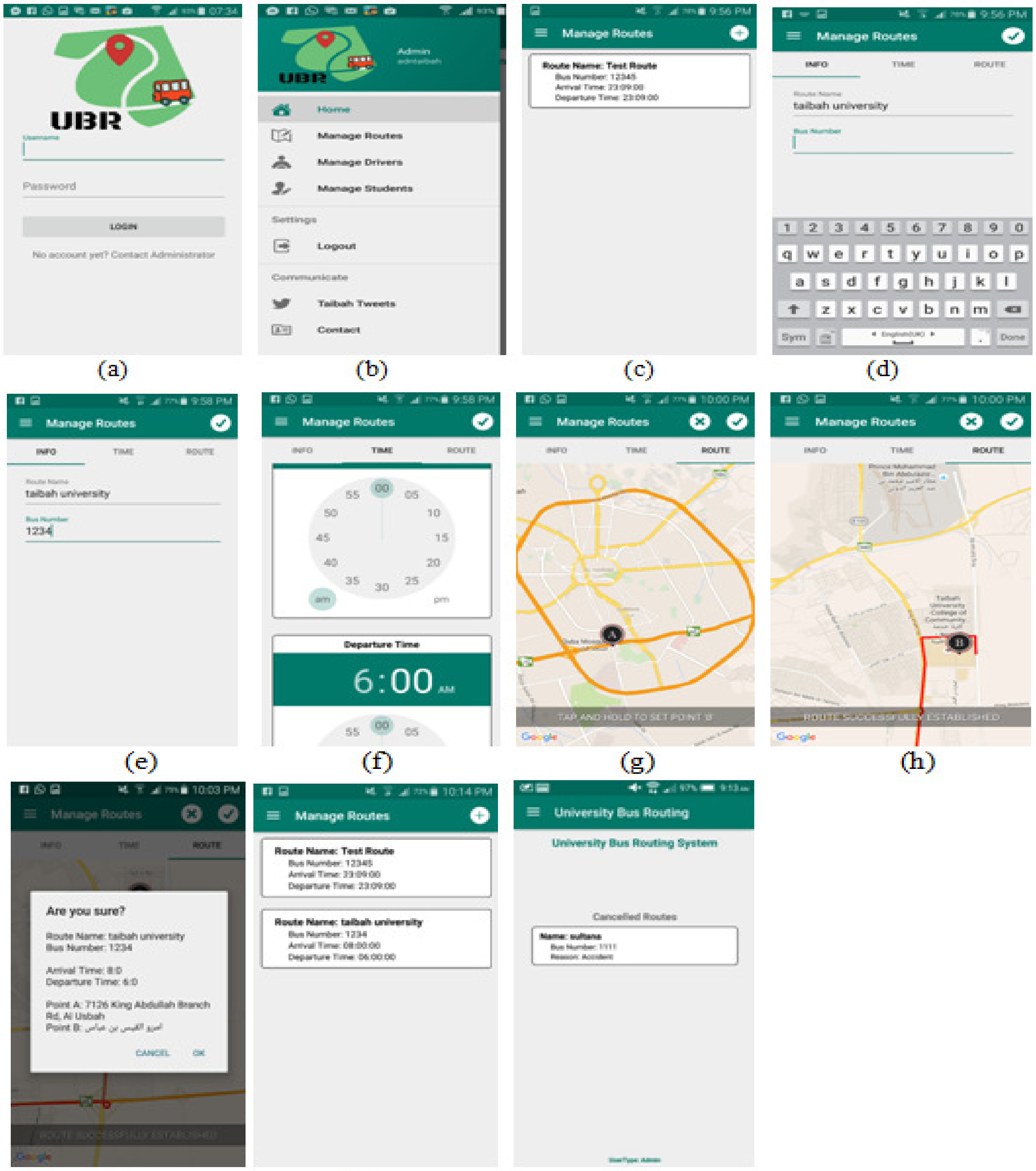
The application has 3 main users Admin, Student and Driver we will present main screen shots for each user as described below.

4.1 Admin

In Figure 3, The admin logs into the system with user name and password, the admin mages routes, students and drivers he can add students and delete them, add drivers, delete them, also the first thing the admin does is add routes that are stored in the database, also the users data is stored into the database in the users table, while adding students the admin assigns routes to student.

Admin can add routes to the system by clicking manage routes where routes screen will appear and add information about route by stating its name and bus number, adding the departure and arrival times; departure is the time the student going route to the university and arrival is the retuning route from the University.

Finally through the route tab the admin states the start and end point of the route by taping on the map to select location A then tapping to select location B. Admin also can delete a student name. Student's information can be viewed by clicking a long click on the student name where a menu appears to view credentials. Admin also can add and delete drivers' names.



(c)

Route Name: Test Route
Bus Number: 12345
Arrival Time: 23:09:00
Departure Time: 23:09:00

(d)

Route Name
taibah university
Bus Number

(e)

Route Name
taibah university
Bus Number
1234

(g)

TAP AND HOLD TO SET POINT B

(h)

ROUTE SUCCESSFULLY ESTABLISHED

(i)

Are you sure?

Route Name: taibah university
Bus Number: 1234

Arrival Time: 8:0
Departure Time: 6:0

Point A: 7126 King Abdullah Branch Rd, Al Ubbah
Point B: امرؤ القيس بن عباس

ROUTE SUCCESSFULLY ESTABLISHED

(j)

Route Name: Test Route
Bus Number: 12345
Arrival Time: 23:09:00
Departure Time: 23:09:00

Route Name: taibah university
Bus Number: 1234
Arrival Time: 08:00:00
Departure Time: 06:00:00

(k)

University Bus Routing System

Cancelled Routes

Name: sultana
Bus Number: 1111
Reason: Accident

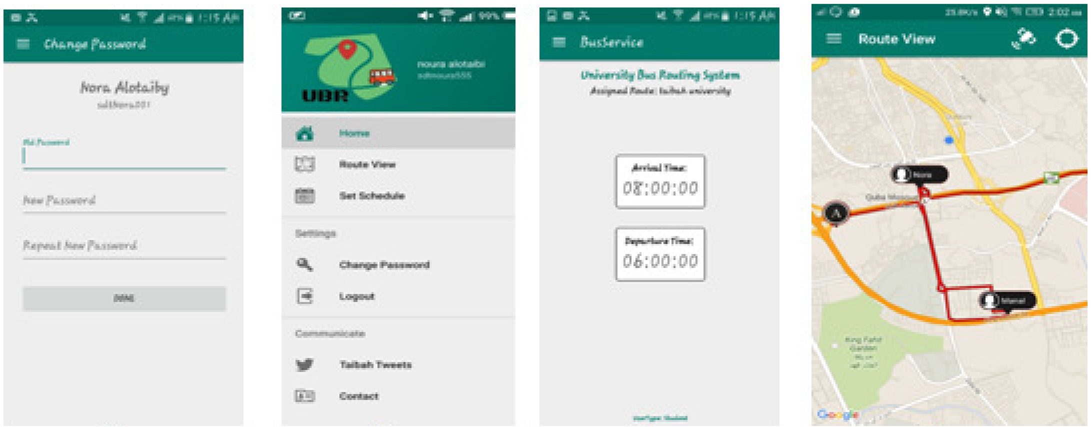
UserType: Admin

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 3 Admin screen shots' example

4.2 Students

As shown in figure 4, the student can login to the system with username and password provided by the administrator, students can view the route in real time and can locate the drivers location on the map when the driver clicks start route, the student will be able to locate the driver on the route and the speed will be displayed too.

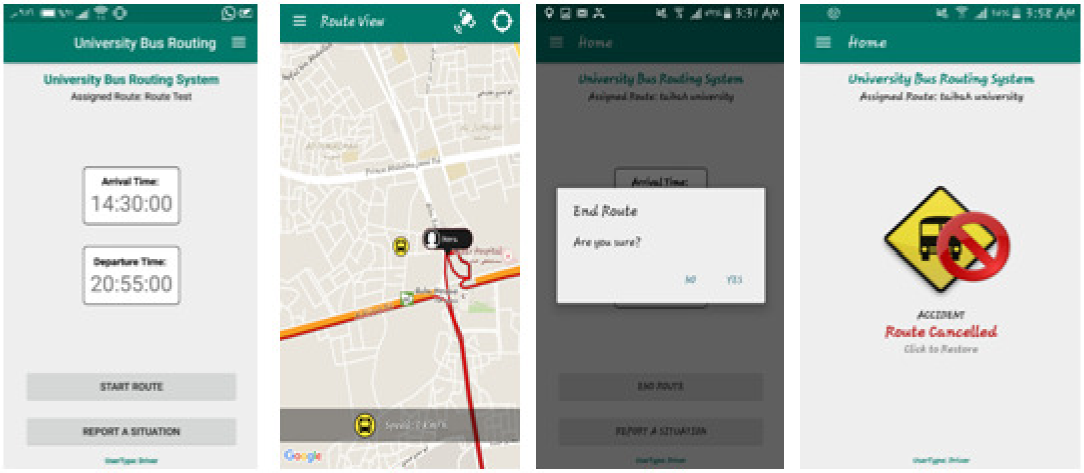


(a) (b) (c) (d)

Figure 4 Students screen shots' example

4.3 Drivers

The driver logs into the system by providing the username and password, driver can start the route by clicking start route on his home page, as soon as the driver clicks start route button the route starts and real time tracking starts on the driver and student mobiles, the driver can view the route where the students' locations are shown. The driver can choose to report a situation if an accident occurs or the bus is out of service and this will appear on the admin's home page immediately to take the ultimate solution to solve the problem as shown in figure 5.



(a) (b) (c) (d)

Figure 5 Students screen shots' example

This application was tested among students, drivers and admin in Taibah University, Almadinah ,Saudi Arabia. They assured that the application is a very useful and will help them. They assured that the user interfaces was easy and the navigation through the application was not hard. Also the application gives real time service as it was developed to be, efficiency of the application was very good.

5. CONCLUSION AND FUTURE WORK

This proposed project intends to resolve the problem with long waiting times students of the Taibah University are facing for buses. It's primarily important to upgrade the existing manual Bus Tracking and Monitoring system to improve transportation services.

For the proposed application, GPS based system is used to suffice the intended purpose. In this proposal, extensive study has been done to elaborate on the bus management and a mobile application based on the same. It also studies previous research works conducted in this field to gauge the potential challenges in future. So, users would need a mandatory GPS, reliable internet connection, and GPS enabled android phone to utilize the app.

The design entails three sections: Geo tracking, scheduling & fair query and security module. The central server would play pivotal role in storing all the information which connects the driver and the students. Considering the potential of GPS enabled systems, this app surely holds the potential to improve the existing bus transportation system for students.

We are considering adding more features and improvements such as a notification messages, calculating the remaining time, and a notification when all are on-board.

There is a possibility to modify our system by adding a 3D map at least for the famous places in the rout for example at Taibah university when the rout is Taibah university location , for example, it can appear in 3D as shown in figure 6.

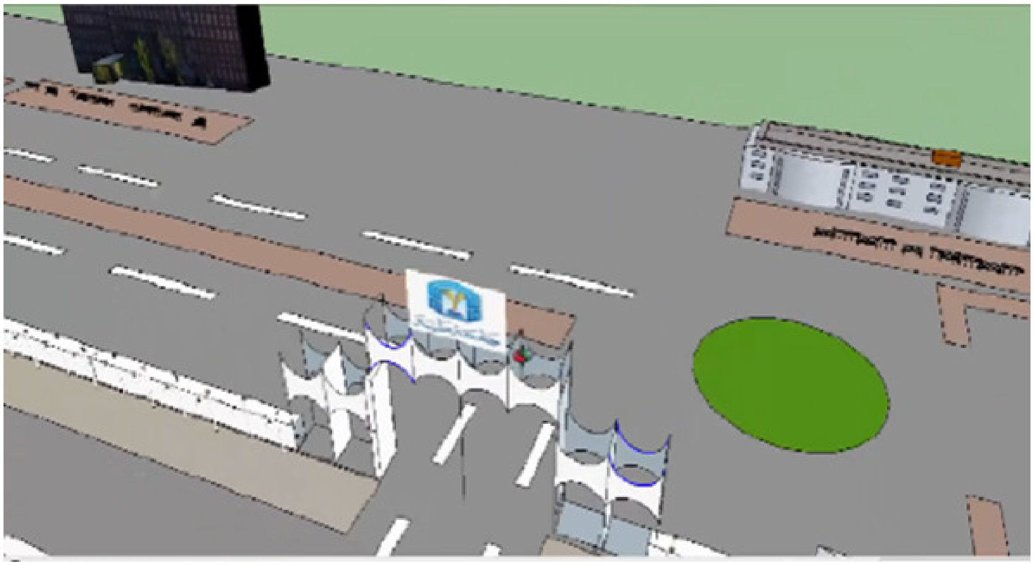


Figure 6 3D Mapping

REFERENCES

[1] Nouf, M. S. & Abdul Khader, J. S.,(2015) Smart Transportation Application using Global Positioning System", (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 6, no. 6.

[2] Tyler, I.,(2012) "Versatrans My Stop Mopile Application", Tyler Technologies.

[3] Kannaki, V. A.& Vijayalashmy, N. & Yamuna, V. &Rupavani, G.&Jeyalakshmy,(2014) " G.: GNSS Based Bus Monitoring And Sending SMS To The Passengers", International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 1.

[4] Priya, B.,(2015) "A Mobile Application for Tracking College Bus Using Google Map", International Journal Computer Science and Engeneering Communications, vol. 3, no. 3, pp. 1057-1061.

[5] Ahlam, M. A.,(2016) "Taibah Track Bus Mobile Application", Taibah University, Almadinah Almunawarra, Saudia.

AUTHORS

Ahmed Ahmed has his BS in Electronics and Communications Engineering from Zagazig University, Egypt in 1995. He has PhD in Computer Engineering and Computer Science from University of Missouri-Columbia, USA in 2005. He is working now as assistant professor at both department of Computer Engineering and systems, Faculty of Engineering, Zagazig University, Egypt and at Department of Computer Engineering, Faculty of Computer Science and Engineering, Taibah University, Elmadinah, Saudi Arabia.

Elshaimaa Nada has her BS in Electronics and Communications Engineering from Zagazig University, Egypt in 2000. She has her master in Computer Science and Control from Zagazig University, Egypt 2009. She has her PhD in Computer Science and Systems from Zagazig University, Egypt 2014. She is working now as assistant professor at department of Computer Engineering and systems, Faculty of Engineering, Zagazig University, Egypt.

Wafaa Al-Mutiri has her BS in Information Systems,Taibah University, Almadinah, Saudi Arabia in 2015.

INTENTIONAL BLANK

PREDICTING VENUES IN LOCATION BASED SOCIAL NETWORK

Omar F.Almallah and Songül Albayrak

Department of Computer Engineering,
Yildiz Technical University, Istanbul, Turkey

ABSTRACT

The circulation of the social networks and the evolution of the mobile phone devices has led to a big usage of location based social networks application such as Foursquare, Twitter, Swarm and Zomato on mobile phone devices mean that huge dataset which is containing a blend of information about users behaviour's, social society network of each users and also information about each of venues, all these information available in mobile location recommendation system .These datasets are much more different from those which is used in online recommender systems, these datasets have more information and details about the users and the venues which is allowing to have more clear result with much more higher accuracy of the analysing in the result.

In this paper we examine the users behaviour's and the popularity of the venue through a large check-ins dataset from a location based social services, Foursquare: by using large scale dataset containing both user check-in and location information .Our analysis expose across 3 different cities.On analysis of these dataset reveal a different mobility habits, preferring places and also location patterns in the user personality. This information about the users behaviour's and each of the location popularity can be used to know the recommendation systems and to predict the next move of the users depending on the categories that the users attend to visit and according to the history of each users check-ins.

KEYWORDS

Personalized Recommendations, Location based social networks.

1. INTRODUCTION

The evolution of the mobile phone has led to big usage of the location based social network application such as Twitter, Facebook, Google latitude and foursquare. The location based service (LBS) can be defined as a software service using the location information data to control features and this information service has now a big number of users who is depending on it in different majorities like health, entertainments and personal life. LBS contain services to recognize a location and give the longitude and the latitude of a person or object such as restaurant or the location of friends, LBS allow to the users to track their package and know where is it and in which time it will send or came.

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 11– 21, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70202

Location based social networks are a type of social network in which the geographic services are the main object in this system which enable the additional social network dynamics such as geo-coding and geo-tagging, according to that the users can visit any location in the world and make geo-tagging about anything the users want to show it like photos, comments and videos. In particular, the geographic services presents it by using three layers first for users, locations and the content layer [1][2]. Here it's sure that the users can exploit information from one layer to another, according to these layer we can calculate the similarity of the information between the users for instance, the places that most of the users can prefers. LBSNs improved the quality of service on: firstly, the recommendation of social events, places, friends and activities, secondly, users behaviour and community detection, finally, personalized recommendation of social events, locations and friends.

Users of location based social network application can records the location that had been visited it, referred to 'check-in'. The check-in is mostly consist of active users, date & time, places and accompanying people. This records of each check-in allows to the users to keep track the place they have been visit [3], also these records allow to the user to receive notifications about there other contact and where they made check-in and they opinion about it. On other hands the users can check the comment of the other users about the places that the user want to visit it these application allow to check everything before going to it. While the user is wondering notification can received about each area passed from directly the program is showing the rate and the comment of other users about this place, before all of these applications also allow to the user to check the place and everything about it from comment and rate and give his own one about that place. The other benefit is to tag the other user who is with the user and show it as a notification in the application. These applications allow to there users to select a list of friends and create their own list of friends as a kind of social networking system [4].

One of the most popular location based social network application is Foursquare. Foursquare is a mobile application which become too much popular as local search and discovery app which is recommending the users to find which places they prefer, also provides recommendations of the locations which is much more close to the users. Foursquare helps and let the users to search and look for different categories such as restaurants, markets, malls, sport places and so on ,also help to look and search for another places by entering the names of the place. Foursquare also displayed the personalized suggestion depending on the time like in the morning will recommend a breakfast places and so on .The recommendations in this app will depend on the history of the users and what they always search and looking for, like if the user is looking always for dinner places it will directly suggest for the user new places for dinner.

In this paper, we present an examination for recommendation system depend on the users behaviours and the popularity of the places and location in different categories that the users of the location based social networks applications are looking for. We provide a prediction system for three cities through Foursquare program to prediction the most popular city and to predict the next place that the user will look for depend on the history of the users and also the categories that the users prefer to attend to go.

2. Related Work

There are two main research related to our work the first one is about recommendations system and the second one is about the prediction system using the history of the user and the prediction system according to the categories.

There are too many number of researches about data mining algorithms that basis on the recommendation system idea [5]. These systems are taking the users performance and chooses such as rating, comments and tags as an input to predict the new move for the user and to show the popularity of the venues, all these is calculating under the term of collaborative filtering. In most of new research work has focus about the dataset which had been taken from the web site, namely movies such as Netflix [6] and music such as Yahoo Music and also another program which give a huge dataset.

Too many literature focus on the idea of using big data in there recommendation system, in [7], they trade on the check-in categories data to model implicit user motion pattern. In there research they focus on the prediction categories according to the user activities, also the prediction according to the location given by the rated category distribution.

In another review, researcher depends on there analysed about user's check-in history also social interaction pattern by using network structures of Foursquare users and venues dataset, also they focus on the venues geographical information and its effect on the users behaviours when he/she choosing the place to better understand the sensitive factors [8].

Also there's research, which evaluate a series of ways and techniques for identification of users from their own check-in information. They applied techniques to analysed the data according to users check-in over time and also the frequency of visit specific location, the techniques was depending on the users identification to analysed the dataset (Trajectory-based Identification, Frequency-based Identification and Measuring the Complexity of the Identification Task) depending on spatio-temporal trajectory emerging for their users check-in. They applied also a hybrid way to exploit both types of information [9].

On other hand, there are researches which depends on users history information to predict the next steps or the next place that the user planning to visit. The idea of these research is about when the users make a check-in by using Foursquare application and showing it by using another social network program such as Twitter. The dataset collected depending on each twit will contain any information about check-in through Foursquare or another location based social network Application [10].

The task of successive personalized point of interest (POI) in LBSNs considered by focusing on how to solve the POI recommendation or prediction by observed two prominent properties in the check-in sequence: personalized Markov chain and region localization. By submitted a matrix factorization method, namely FPMC-LR, to firm the personalized Markov qualifications and the centralized zone. The idea here is not just about personalized Markov chain in the check-in sequence, but also to look around the localized area [11].

For another point according to the interest of the user they focus on the problem of the time aware point of interest recommendation systems, which recommending to the user to visit new places at a given time according to the users interest. To explain the geographical the temporal influence in the point of interest POI recommendation systems. They suggested Geographical-Temporal influences Aware Graph (GTAG) to calculate and record the check-in also the geographic and temporal influence. In this project to make the Geographical-Temporal influences Aware Graph (GTAG) more effective and more efficient recommendation they developed the performance of it by propagate methods named Breadth-first Preference Propagation (BPP). According to this algorithm the recommendation system will returns results within at most 6 diffusion steps. The recommendation systems results were time aware because of the perception that the users looking for different places at different time [12].

3. DATASET

In this section, we describe the dataset that we are dealing with what it contain for how many check-in made by how may user and the 3 cities that we are comparing it. Then we analysed the dataset by calculating the number of check-in in each venue and find the average and the number of check-in for each location also find the average number of check-in for each user.

3.1. Data Collection

Location based social networks have been very popular subject and this kind of datasets it become very interesting point to attracted millions of people footprint and what they are looking for. The dataset had been collected from foursquare program which is represent and deal with different categories and venues that the users of this program prefer to look for it. In order to study on personalized location recommendations foursquare dataset one of the best option in this time to understand what the users interesting in. Users can access foursquare by using their own phone very easily and show their status by make check-in in any place that the users wants. Through these programs users can write any comment about there status with the rate that want to give for this place, also make share and tags for anyone from there friends which is exist in there contents, also foursquare allow for the users to create there own list about the places that they prefers which it's the best for the users according to what the user is always looking for.

The foursquare dataset that we used in our research showing dealing with different kind of categories in three different cities which they are London, Austin and Dallas in (March and April of 2011) the dataset shows check-in in too many places for different categories according to too many users. Each of cities has different amount of check-in and also check-in in different places in each of the cities for different categories and positions. Each for these cities dataset has its own properties about the number of users, check-in and the categories. For London city dataset, it contain more than 4 millions check-in for more than 100 thousands users in more than 40 thousands different positions and places for different categories and also it contain the longitude and latitude of each venues and for the check-in it's explaining and exposure the date , time for each check-in and showing also the number of the street and the building, these dataset had been taken in 2011 for March and April, according to these dataset there's some user made more than 200 check-in and other less these number until one check-in in different categories ,date and time. According to all these information we can predict the best places also speculate the rush hour and give the best recommendation for the others users.

3.2. Data Analysing

In this part we analyzed our dataset for three cities acount the number of check-in for each city .Also we calculate the average number of check-in with the number of users and with the number of places .In order to have more clear result and in order to make to make the dealing withthis much big dataset much more easily .This calculation for the average for each city according to the number of check-in with the number of users and also with number of places as shown in Table.1

Table.1 Average properties observed on Forsquare over the specific date of the data set:Total number of user(N), places (M) and check-in (C), average number of check-in per user (Cu) and per place (CP)

City	N	M	C	Cu	Cp
London	104,076	4,384	4,162,121	11.7	97
Austin	42,122	12,971	1,474,270	7.5	115
Dallas	35,593	15,751	1,637,232	13.7	105

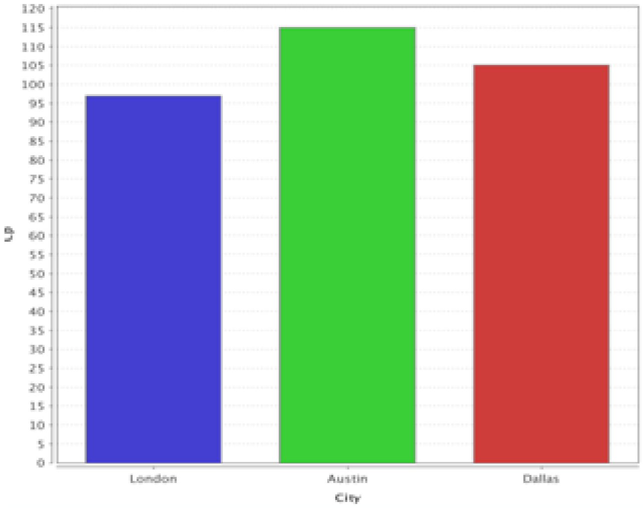


Figure1. Average number of check-in per place

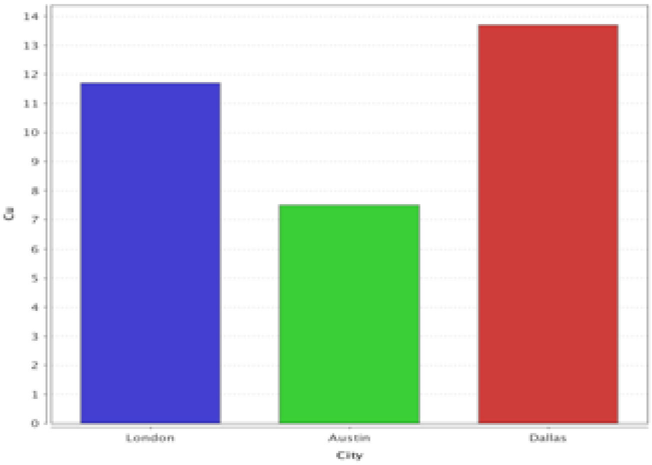


Figure 2. Average number of check-in per user

According to the table 1 we calculate the average number of check-in per place as shown in Fig.1 that the highest number is for Austin city, also the average number of check-in per user Fig.2 that the highest average is for Dallas city.

4. PREDICTION ANALYSIS

Location based social system network can improve the services on Generic recommendation of social activities, events, friends and location, another one is the personalized recommendations of activities, events, friends and users mobility. All the recommendation systems algorithms depend on the dataset in a different way such as following friends algorithms will depend on the friends what they prefer and like the algorithms which depending on the history of the users and according to that will advise the users to what they prefer to visit.

The personalized recommender system depends on the check-in histories of the users. Then, compare the history of users with each other and find the correlation of the performance for the users and suggest to the users new places, events and activities. In particular, the personalized recommender take the advantage of the time that some of the user has visited a place and give a rate or comment on that place and predict for the user unvisited place similar to the history of that user [13].

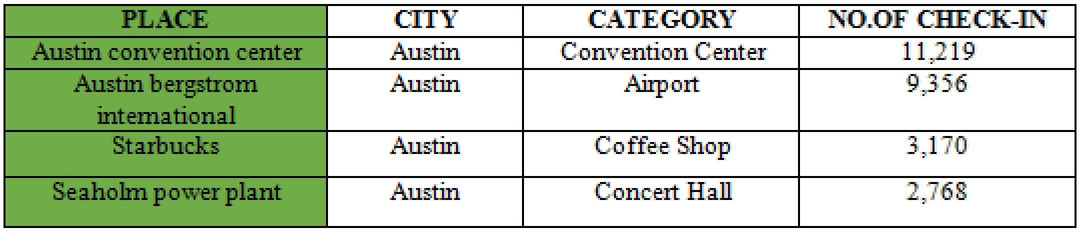
We applied some technical method on our dataset for the prediction system according to the history of the users and the number of the check-in that the user made in the three city also depending on categories and find the most popular one for the three cities In another hand predict the next place that the user will like to go according to the popularity of the place in our dataset to know the most popular city between Austin, Dallas and London by taking and counting the most high rating location and according to that we predict the popular city.

In this section, we applied a set of algorithms that we examined for the prediction system:

4.1. Visiting popular places

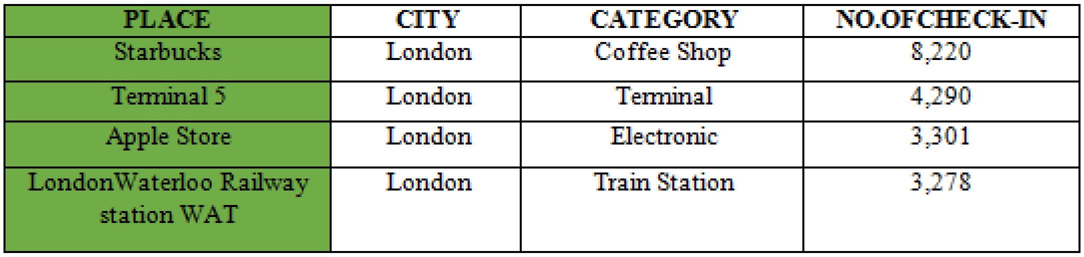
The first one is non-personalized baseline on the rank of the place also on the number of check-in that the users make in that place. We found the most popular places in the three cities:

Table.2 Austin city table for the most popular places



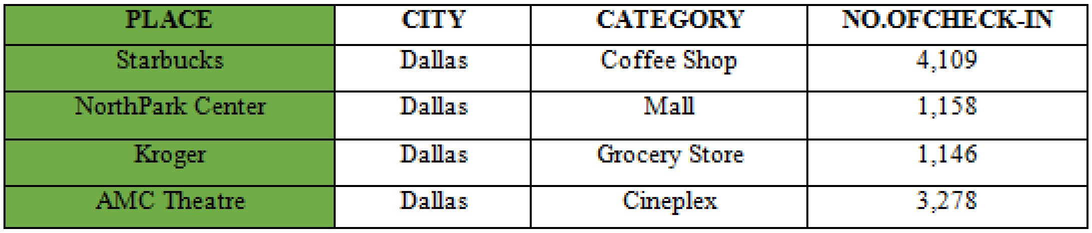
PLACE	CITY	CATEGORY	NO.OF CHECK-IN
Austin convention center	Austin	Convention Center	11,219
Austin bergstrom international	Austin	Airport	9,356
Starbucks	Austin	Coffee Shop	3,170
Seaholm power plant	Austin	Concert Hall	2,768

Table.3 London city table for the most popular places



PLACE	CITY	CATEGORY	NO.OFCHECK-IN
Starbucks	London	Coffee Shop	8,220
Terminal 5	London	Terminal	4,290
Apple Store	London	Electronic	3,301
LondonWaterloo Railway station WAT	London	Train Station	3,278

Table.4 Dallas city table for the most popular places



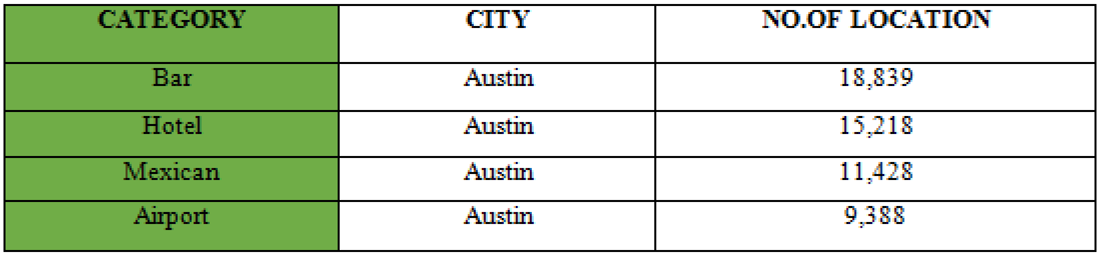
PLACE	CITY	CATEGORY	NO.OFCHECK-IN
Starbucks	Dallas	Coffee Shop	4,109
NorthPark Center	Dallas	Mall	1,158
Kroger	Dallas	Grocery Store	1,146
AMC Theatre	Dallas	Cineplex	3,278

The previous tables show the most popular places in each of the three cities depending on the number of check-ins in each of these places.

4.2. Attending places by categories

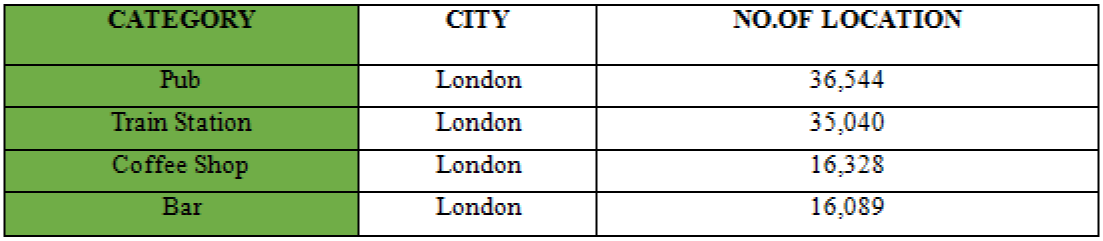
The next method is a content based filtering. The prediction system depend on the most popular categories that most of user attend to visit and the type of the categories that most of the user are looking for in all the three city.

Table.5 Austin city table for the most popular categories



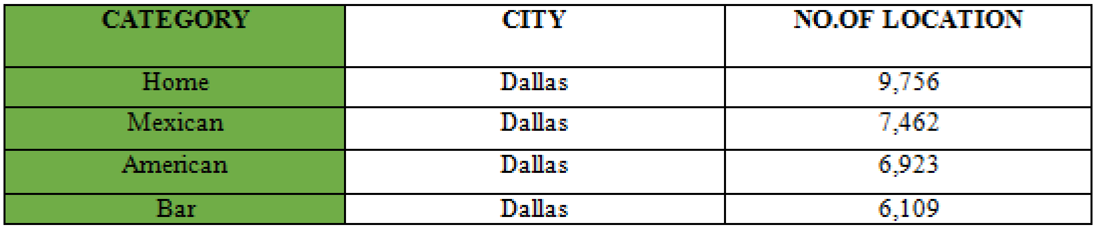
CATEGORY	CITY	NO.OF LOCATION
Bar	Austin	18,839
Hotel	Austin	15,218
Mexican	Austin	11,428
Airport	Austin	9,388

Table.6 London city table for the most popular categories



CATEGORY	CITY	NO.OF LOCATION
Pub	London	36,544
Train Station	London	35,040
Coffee Shop	London	16,328
Bar	London	16,089

Table.7 Dallas city table for the most popular categories



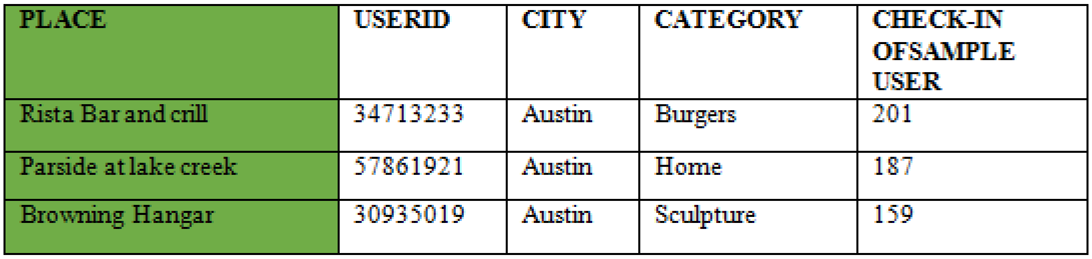
CATEGORY	CITY	NO.OF LOCATION
Home	Dallas	9,756
Mexican	Dallas	7,462
American	Dallas	6,923
Bar	Dallas	6,109

According to the previous tables that it showing the most interested categories in each of the three cities such as the most popular category in Austin city is “Bar” which means all the places under the name of Bar category is get attend more than the other categories.

4.3. Attending Places by User History

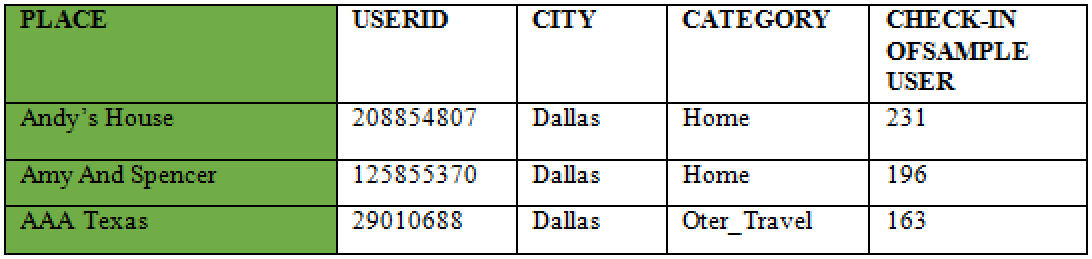
This method is depending on the history of each user and giving the next move depending on what that user is prefer to visit and attend to go. Each user has a type of venues like to visit and according to that the program will speculate the next place for the user according to the history of him/her. We applied this method on our dataset for each user in the three cities and predict the next move for each user depending on their history and the number of check-in that the user made in his/her favourite venues.

Table.8 The most popular places in Austin city according to the history of selected users



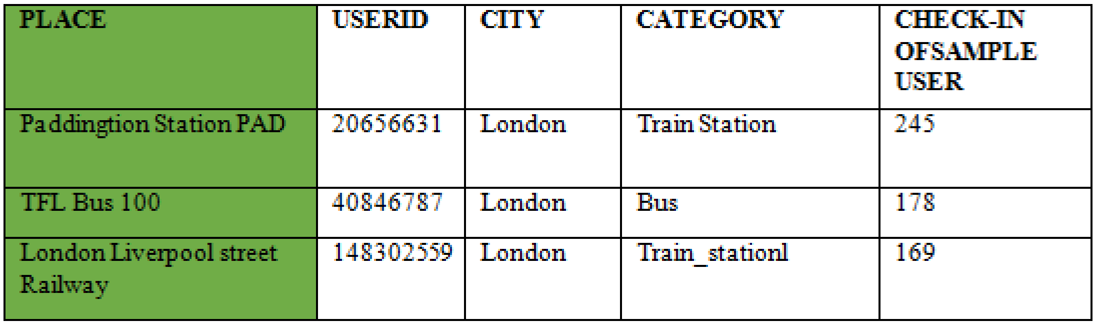
PLACE	USERID	CITY	CATEGORY	CHECK-IN OFSAMPLE USER
Rista Bar and crill	34713233	Austin	Burgers	201
Parside at lake creek	57861921	Austin	Home	187
Browning Hangar	30935019	Austin	Sculpture	159

Table.9 The most popular places in Dallas city according to the history of selected users



PLACE	USERID	CITY	CATEGORY	CHECK-IN OFSAMPLE USER
Andy’s House	208854807	Dallas	Home	231
Amy And Spencer	125855370	Dallas	Home	196
AAA Texas	29010688	Dallas	Oter_Travel	163

Table.10 The most popular places in London city according to the history of selected users.



PLACE	USERID	CITY	CATEGORY	CHECK-IN OFSAMPLE USER
Paddingtion Station PAD	20656631	London	Train Station	245
TFL Bus 100	40846787	London	Bus	178
London Liverpool street Railway	148302559	London	Train_stationl	169

In the previous tables is showing the prediction system for each city depending on the check-in of the users in all the three cities. According to the history of the users on Foursquare program to predict the next place that the user will go to it.

5. EVALUATION

We have evaluated the recommendation system algorithms result and compare it across the predictor, datasets and the three cities. In this section we describe our methodology and the metrics that we used to evaluate the recommendation system quality.

5.1 Methodology and Metrics

In this section we separate the dataset in to train and test in order to calculate accuracy result. We filter the dataset from zero and unknown check-in Ci,j value to have more clear result. We use three metrics to calculate and find the quality of these recommendation system's result that we have found.

5.2 Result

We calculate the precision and the recall for each of the three method that we analysed our dataset on for all of the three cities to major the quality of our prediction and analysis system to find the best result as shown in Tables.

Table 11. Austin city the precision, recall and accuracy

Method / Austin city	**precision**	**recall**	**accuracy**
Popular places	0.5742	0.658	0.662
Attending by categories	0.1635	0.0739	0.215
User history	0.1056	0.16	0.1667

Table 12. Dallas city the precision, recall and accuracy

Method / Dallas city	**precision**	**recall**	**accuracy**
Popular places	0.526	0.6173	0.6214
Attending by categories	0.1306	0.0691	0.1625
User history	0.0576	0.09	0.0918

Table 13. London city the precision, recall and accuracy

Method / London city	**precision**	**recall**	**accuracy**
Popular places	0.4866	0.5691	0.5711
Attending by categories	0.259	0.1206	0.223
User history	0.1257	0.24	0.2414

The user will visit the top-N places as the result on the tables .The measured of the recall and the precision were depend on true positives (tp), false positives (fp), and false negatives (fn):

$$precision = \frac{tp}{tp + fp} \quad , \quad recall = \frac{tp}{tp + fn}$$

Two formulas to calculate the prediction system quality are precision and the recall with the accuracy that we applied on it to calculate the quality of prediction system.

According all of the three cities of our dataset the highest accuracy, precision and recall where about the prediction system depend on the popularity of the places.

6. CONCLUSION

In this paper we had described the prediction system focusing on three point in the dataset that we have from Foursquare the history for each of the users and what it will be the next step of them. We focus about attending new places by their categories and which category is the most popular one in our dataset. Also we concentrate counting the most popular place in each of the cities and predict the most wanted one according to the number of check-in that get taken on it. The goal of the three kind of prediction in our paper is to give a recommendation for the other users and the same users of the Foursquare program because this number that we found it’s rating point for the users that they can depend on it for the next step of them.

In the term of the future work, we seek to evaluate a recommendation system that can calculate these number and give directly high rate for the prediction by applying another filter to have more specific result with high percentage of quality of prediction system.

REFERENCES

[1] Needleman, Rafe; Claire Cane Miller; Adrianne Jeffries (3 September 2010). "Reporters' Roundtable: Checking in with Facebook and Foursquare". CNET. Retrieved 8 October 2010.

[2] "Recommending Social Events from Mobile Phone Location Data", Daniele Quercia, et al., ICDM

[3] M. C. Gonzalez, C. A. Hidalgo, and A.-L.Barabasi, “Understanding individual human mobility patterns,” Nature, vol. 453, pp. 779–782, June 2008.

[4] C.-Y. Chow, J. Bao, and M. F. Mokbel. Towards Location-based Social Networking Services. In ACM SIGSPATIAL-LBSN, 2010.

[5] G. Adomavicius and A. Tuzhilin. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE TKDE, 17(6):734–749, June 2005.

[6] Y. Koren. Collaborative Filtering with Temporal Dynamics. In Proceedings of KDD '09, pages 89–97, Paris, France, 2009.

[7] What's Your Next Move: User Activity Prediction in Location-based Social Networks by (Jihang Ye_ Zhe Zhu_ Hong Cheng).

[8] What's Your Next Move: User Activity Prediction in Location-based Social Networks by (Jihang Ye_ Zhe Zhu_ Hong Cheng).

[9] It's the Way you Check-in: Identifying Users in Location-Based Social Networks by: Luca Rossi and Mirco Musolesi. School of Computer Science University of Birmingham, UK l.rossi@cs.bham.ac.uk . Mirco

[10] Location-based Predictions for Personalized Contextual Services using Social Network Data by: Rui Zhang, Bob Price, Maurice Chu and AlanWalendowski . Palo Alto Research Center Inc.(PARC, a Xerox Company) 3333 Coyote Hill Rd. Palo Alto CA 94304 Rui.zhang@parc.com

[11] Daniele Quercia, et al., "Recommending Social Events from Mobile Phone Location Data", ICDM 2010.

[12] Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences by: Quan Yuan, Gao Cong, Aixin SunSchool of Computer Engineering, Nanyang Technological University, Singapore 639798 {qyuan1@e., gaocong@, axsun@}ntu.edu.sg . Copyright 2014 ACM 978-1-4503-2598-1/14/11

[13] p. symeonides et al. recommender systems for location based social network springersbriefs for electrical engineering.

AUTHORS

Omar Almallah
Computer engineering
I am a master student in yildiz technical university in Istanbul turkey.
Phone :05372675182
Email: omarfiras991@gmail.com

Dr. SONGüL VARLI ALBAYRAK
Associate Professor
YILDIZ TECHNICAL UNIVERSITY
Computer Engineering Department
songul@ce.yildiz.edu.tr

INTENTIONAL BLANK

MODEL FOR HEURISTIC AND AI PLANNING STRATEGIES – A PATH TO LEADERSHIP IN SECTOR “TELECOMMUNICATIONS” IN BULGARIA

Miglena Temelkova[1], Dimitar Radev[2] and Strahil Sokolov[3]

[1]Department of Communications Management,
University of telecommunications and Post, Bulgaria
[2]Department of Telecommunications,
University of telecommunications and post, Sofia, Bulgaria
[3]Department of Information Technologies,
University of telecommunications and post, Sofia, Bulgaria

ABSTRACT

The introduction of a model of heuristic and AI planning strategies as a management instrument by the companies operating in sector “telecommunications” in Bulgaria leads to reducing the time for collecting, processing, analyzing and evaluating information necessary for the strategic planning and goal setting. At the same time several benefits are achieved, such as lower resource consumption, reduced risk, higher level of objectivity and reliability of the management process, personalization of services. Additionally, attaining optimality and higher added value not only for the management, but also for the actual production process can be ensured . These strategic competitive advantages turn the organizations into leaders in the sector of telecommunications, where digitalization and automation form a specific “ecosystem” and determine the very dynamic and innovative development of the sector over the recent years. The challenge for management is the reproduction of the dynamic processes in a model, which implements a flexible systemic architecture carrying realistic and adequate information about the structure, processes and functional fields of the studied subject, enabling the planning of its activity as a response to the strategic goal setting. The application by the Bulgarian business organizations, operating on the telecommunication market, of such a model, based on heuristic systems with artificial intelligence, would lead to achieving leadership in respect to profit and market share in the sector.

KEYWORDS

Model, heuristic and AI planning strategies, telecommunications, leadership, added value

1. INTRODUCTION

The exponential speed with which the Fourth Industrial Revolution is developing leads to the quick upgrading of the achievements of the digital revolution, to combining the multitude of innovative and intelligent technologies, to new technological breakthrough comprising areas such as artificial intelligence, robotics, the Internet of things, autonomous vehicles without a driver, 3D printing, nanotechnologies, biotechnologies, material science, energy preservation, quantum

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 23– 35, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70203

calculation. This determines both some unprecedented changes in the paradigm in economy, business and society, and a deep long-term transformation of all the public-and-economic, socio-economic, and business systems influencing the development of the traditional branches. The tectonic shifts in all the sectors of economy require the searching and finding of new business models, restructuring manufacturing, consumption, transportation, the supply systems and the entire production and commercial turnover.

The unprecedented merger of technologies in the physical, digital and biological world require from the economic entities accelerated management, production and technological transformation, which is impossible without the upgrading of the human resource potential for using a new management toolset, models and approaches based on the heuristic and AI planning strategies.

The changed conditions of the environment in which the business organizations operate ever more tangibly require the need of their intelligent management, where the role of the human factor is minimized, and the objectivity of the analysis, evaluation, planning, organization and control over the strategy is focused on the competencies of some innovative heuristic and intelligent management systems. The introduction of the heuristic and AI tools and models in the management process of the Bulgarian telecommunication companies will "push" some of them to a leadership position, since currently the market share of the three major telecommunications operators is relatively equal.

The leadership, achieved through heuristic and AI planning strategies, should be described as:

- ✓ coordinated with the strategic organizational reference points;
- ✓ realistically reflecting the changes and challenges of the environment;
- ✓ time-wise limited by the global economic, social, technological and innovation dynamics, which is accompanied by an excessive race for resources and over competition.

2. Development of The Sector "Telecommunications" in Bulgaria

In 2015 in the sector „telecommunications" in Bulgaria, the total revenue from the provision of mobile telephone services decreased by 18.5% compared to the previous year 2014. The trend is due mostly to the saturation of the traditional markets for voice services, the imposed regulatory reductions of the prices for termination, as well as a growth of 106% compared to 2014 of the services offered in a package. [1] The reported drop in the number of the active SIM cards with a possibility to use voice services for 2015 is 3.1%, where the main reason for that is the tight competition between the participants on the market, who offer increasingly advantageous plans for the users with a growing number of minutes included for making calls outside their own network, thus making useless the practice for a user to have SIM cards of more than one mobile operator.

The domestic traffic is 73.4% for the different mobile networks, which is a relatively high value, however, compared to 2014, there is drop registered of almost 7%.

The traditional telecommunication companies on the Bulgarian market are "Mobiltel" JSC, „Telenor Bulgaria" JSC and BTC JSC, where, in spite of their relative parity, the redistribution of the market shares between them is still going on.

“Mobiltel“ and BTC report some growth in the shares calculated both by the number of clients, and by the revenue from the provided mobile services at the account of a drop in the market shares of “Telenor Bulgaria”.

“Mobiltel” has a market share of 40.9% (according to the number of subscribers), which represents growth of 0,5 percentage points compared to 2014 and 40.8%, revenue-based (annual growth of 1.9 percentage points). For 2015, BTC had a share of 25.5% (according to the number of subscribers) and growth of 1.4 percentage points compared to the previous year and 30.7% revenue-based, with annual growth of 4.7 percentage points.

Unlike 2014, when “Telenor Bulgaria” had market share growth by the number of clients, in 2015 that share decreased by 1.9 percentage points down to 33.6%, while the share of the operator on the revenue basis amounted to 28.5% with yearly drop of 6.6 percentage points.

An additional factor for the redistribution is the transferability of the telephone numbers.

The total number of the transferred numbers in the mobile networks for the period 2009-2015 was above 1.2 million. In 2015 the number of the transferred numbers grew by 40% compared to 2014, with growth of 27.55% in 2014 compared to the previous year.

Towards the end of 2015, the total number of the actual providers of services for transferring data and/or Internet access was 669, where growth of 4% was reported compared to the previous year. The broadcasting of radio and television broadcasts registered a drop compared to 2014 of almost 6%.

The total number of subscribers to Internet services increased by 31% compared to 2014. The number of subscribers of fixed access to the Internet increased by 8% compared to the pervious year, and the number of the users of mobile services for Internet access – by 42.4% compared to 2014. The increase of the number of subscribers of mobile access to the Internet was due to the growth of 76.6% for one-year period of the users of package services with included mobile access to the Internet.

Towards the end of 2015 the number of the companies offering package services was 100. Compared to the previous year, there was considerable growth both of the revenue from installation fees and monthly subscription of services in a package (with 51.5%), and of the number of subscribers (with 45%).

Most popular with the users with 68.6% relative share is the package service, including mobile voice service and mobile access to the Internet, where the number of the subscribers to that service grew by 76.9% compared to 2014.

Next, in terms of the number of subscribers, are the package services for television and fixed access to the Internet with 15.6% share and mobile voice services with 10.4%.

The paper foresees growth of those services in the future as well, since they are popular with the users.

Increasing competition in the telecommunication sector has been observed on the Bulgarian market, which also leads to fostering investment in the sector and is also beneficial for the users. All of this delivers the necessary prerequisites for introducing a model of heuristic and AI planning strategies in the management of the Bulgarian telecommunication companies. Thus, those of them, which can adapt in a quick and flexible way and with appropriate goal setting to the new market conditions, will succeed in becoming leaders on the Bulgarian market in the telecommunication sector. The new challenges result from the Fourth Industrial Revolution, namely the increasingly higher requirements of the users and their expectations for fair treatment; quality and price ratio and also service accessibility.

The advantaged of the heuristic and AI planning strategies are primarily related to:

- ✓ the possibility for creating "intelligent factories", where the virtual and physical production systems will cooperate in a flexible way and on a global scale;
- ✓ the merger of technologies and their interaction in the physical, digital and biological sphere;
- ✓ total personalization of the products;
- ✓ creating new production models;
- ✓ the possibility for multiple options and application of scenarios based on the integration of the system, situational and process approach;
- ✓ lower resource consumption of the global value chains;
- ✓ optimization of activities;
- ✓ increasing the objectivity of the analytical process;
- ✓ decreasing the risk in the management process.

The low level of leadership in the telecommunication sector in Bulgaria and the insufficiently comprehended understanding of the changes occurring in all the sectors, creates favourable preconditions for achieving leadership in the sector by the economic entities, which have strategically set the goal of introducing the heuristic and AI planning strategies in their management activities. As a result from that, there will be not only an increase in the revenue and market share of those telecommunication companies, but there will be also an increase in the added value, which they have for their employees, users and investors.

3. Model For Heuristic and AI Planning Strategies in The Telecommunication Sector in Bulgaria

3.1. Nature of the planned strategy

The strategy suggests a shift in the organization from its current position to a desired but unknown future position. Since the organization has never been in that future place, the road there represents a series of related hypotheses. [5] The model for heuristic and AI planning strategies describes those hypotheses, which makes them evident and verifiable.

The strategy is a new route for the business organization – a path, which it has never followed. Nevertheless, how much it has been discussed, the strategy quite often remains unrealized. The strategic goals, set in the model, serve as starting points on the way of realizing the strategy (fig. 1.), while the strategic alternatives are preconditioned via the relation between strategic goals and specific systems of events, mutually integrated in an integral and comprehensive strategy.

There is no principle imposed in the economics literature and practice about the "right" number of strategic goals. If we assume that each strategic goal can be bound mainly to two efficiency indicators, so that its meaning could be accurately reflected, then 20 strategic goals would mean at least 40 efficiency indicators in one single organization-and-production system of indicators. Provided that such a system is multiplied in the form of a cascade structure of systems on the different hierarchical levels in the organization, it will be quick and easy to reach the impressive several hundreds of indicators – a process, which is quite difficult for management and control. In view of using the capacity of methodology as a system for simultaneous measurement and communication, it would be appropriate to keep the number of strategic goals on an acceptable and manageable level. Generally, as a recommendation and orientation for an appropriate initial first estimation, the strategic goals could be between 10 and 20.

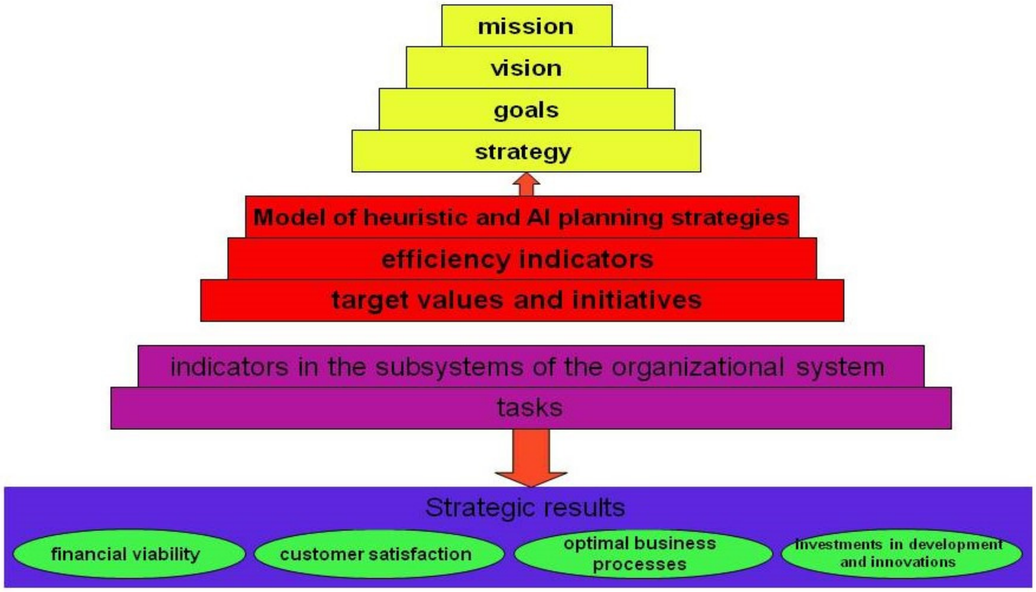

mission
vision
goals
strategy
Model of heuristic and AI planning strategies
efficiency indicators
target values and initiatives
indicators in the subsystems of the organizational system
tasks
Strategic results
financial viability
customer satisfaction
optimal business processes
investments in development and innovations

Figure 1. The place of the model for heuristic and AI planning strategies within the strategic management of the business organizations

In the elaboration of a model for heuristic and AI planning strategies, its functionality and efficiency are determined by:

- ✓ the logical completeness of the cause and effect relations therein;
- ✓ the accurateness and correctness of its elements;
- ✓ its logical theoretical appropriateness;
- ✓ the logical connection of its elements one to another;
- ✓ the efficient planning of the business organization strategy on the basis of the set goals;
- ✓ the accurate translation of the strategy from goals to efficiency indicators;
- ✓ the provision of clear understanding about what is meant with each of the goals;

- ✓ the balance between the efforts and the strategic activities in realizing the mission and vision of the organization.

3.2. Elements of the model of heuristic and AI planning strategies

The heuristic methods and planning through artificial intelligence are increasingly finding their application in business management. Decisions with them are taken on the basis of the inherent for man intuition, experience, logical thinking, included as a bulk of data into systems with artificial intelligence, providing the playing of scenarios under certain restrictive conditions. Using the model of heuristic and AI planning strategies in the telecommunication sector in Bulgaria leads to identifying problems with high degree of complexity through searching and planning decisions in huge discreet spaces. The decisions in the model of heuristic and AI planning strategies are a sequence of actions from one initial condition to one or more targeted conditions.

The way from the initial condition to the targeted one comprises a system if activities, which the model of heuristic and AI planning strategies should perform under some set:

- ✓ parameters, characterizing the particular telecommunication company and the entire telecommunication sector in Bulgaria;
- ✓ goals outlining the desired results;
- ✓ target values enabling the measurement of the goals;
- ✓ tasks issuing from the strategic goals;
- ✓ indicators measuring the fulfillment of the tasks;
- ✓ strategic alternatives generated as a result of analysis and evaluation of the environment, goals, the target values, tasks and indicators, determining their feasibility;
- ✓ strategic initiatives – components of the planned strategy, representing the tactical activities on the way of its realization;
- ✓ cause and effect relations between goals, tasks, resources, planned strategy.

The strategic goals define what needs to be done with the strategy. The strategic goals set in the model of heuristic and AI planning strategies are more specific than the content of the strategy itself, yet, they are less precise than the efficiency indicators. They translate the often too generally and vaguely expressed strategic priorities into directing and action focused descriptions of what has to be done so that the planned strategy is realized.

Through the model of heuristic and AI planning strategies there are processes and measures planned for each goal, which contribute to its achievement. Defined also are efficiency indicators, which measure the extent of reaching the goal on the basis of set target values. The responsibility for achieving a certain goal or sub-goal is personalized in the form of a strategic map.

Owing to the model for heuristic and AI planning strategies, the management of the telecommunication companies in Bulgaria can monitor the extent of implementing the planned strategy and undertake the necessary corrective activities.

The model of heuristic and AI planning strategies on the one hand is a conceptual system for targeted adaptive management, which on the basis of specific parameters systematizes and summarizes the situation in the Bulgarian telecommunication sector, and, on the other hand – multiplies the basic characteristics and directions of the activity of a particular telecommunication company. This model is an important instrument of the strategic management for evaluating the potential and efficiency of a certain company, since it transforms the mission and goals into a well-balanced complex of planned integrated indicators of the strategy [2]. In its capacity of a strategic management instrument, the model of heuristic and AI planning strategies facilitates:

- ✓ the planning of the strategy for development of the telecommunication company;
- ✓ adapting the organization development strategy on each level therein;
- ✓ multiplying the strategic goals in the operational management;
- ✓ deploying system controlling mechanisms in respect to reaching the strategic goals in a current mode.

The model of heuristic and AI planning strategies has as its main goal the accumulation of added value in a telecommunication company. Thus, showing where the growth in revenue and profits comes from, it becomes a driver for value planning and managing. The main advantages provided by the model of heuristic and AI planning strategies are:

- ✓ application of the planned strategy;
- ✓ goals defining, perception and implementation;
- ✓ formulating strategic alternatives for each of the goals, whereby monitoring efficiency in a balanced way;
- ✓ identifying the factors facilitating development, where the focus is shifted from the financial indicators and the past to the development driving indicators and the future;
- ✓ undertaking prompt and efficient initiatives adequate to the occurred and/or occurring changes in the telecommunication sector.

The equality in importance and significance of all the indicators determines the level of balancing in the model of heuristic and AI planning strategies. This balance has a multi-aspect nature and comprises, integrates and focuses the relations between the financial and non-financial indicators, the strategic and operational management, the past and future results, the internal and external aspects of the activity of a telecommunication company. On this basis, the model of heuristic and AI planning strategies should project the entire production organization through planning of strategic and operational assignments within the already defined strategic goals. Being based on balancing, comprehensiveness and the strive towards efficiency, the main idea of the model of heuristic and AI planning strategies can be summarized into:

- ✓ analysis of the environment;
- ✓ generating strategic alternatives;
- ✓ evaluation of the possibility to reach the defined strategic alternatives;

- ✓ planning the strategy on the basis of strategic goals.

A number of authors [3] quite often assume that everything in the business organization should be measurable. Thus, measurability becomes a major motive of the concept for the model of heuristic and AI planning strategies, and the impossibility to manage processes and indicators, which cannot be measured, requires that all the factors important for the management of a telecommunication company should be presented in the form of indicators (markers). The relations and interactions on all the levels of the management hierarchy in the organization (from the top management level to the auxiliary units) are established in the model of heuristic and AI planning strategies, through determining their respective functional goals and indicators.

The model of heuristic and AI planning strategies is built top down, which allows for the strategic goals of the telecommunication company to be worked out in detail logically, structurally and organizationally into operational goals. Most important is the essential determination and definition of the strategic alternatives, and the main functional goals should be derived from the strategic goals of the business organization. Upon laying out the functional goals, for each goal there should be determined also the critical success factors – i.e., a certain target function is set.
The model of heuristic and AI planning strategies is a powerful instrument for describing, planning and implementing the strategy. It considers the strategic alternatives as a result of series of cause and effect relations between the strategic and functional goals in a telecommunication company. The general interaction of those relations can be represented in the form of the so called strategic map [4].

The strategic alternatives can be defined as a key element in planning the strategy through which those entities are visualized, whose interests have substantial importance in the realization of the strategy. The tasks in the model of heuristic and AI planning strategies are determined by the formulated strategic goals, which are an integral part of the telecommunication company strategy. They define:

- ✓ the process of transforming the strategy on operational level;
- ✓ the set of tools used for the needs of the strategic and operational management;
- ✓ the parameters of the strategic and operational optimality.

The fulfillment of the tasks is reflected through the calculated indicators, which are also called measuring elements. These indicators determine the success or failure of the already formulated strategic alternatives and operational assignments. The accurate formulation of the measuring elements determines the form and scale in which the assignments on the operational and process level should be realized. Each of the strategic alternatives in the model of heuristic and AI planning strategies is determined and based on 3 to 5 adopted key indicators, which characterize it and define its variations in the course of achieving the planned strategic goal.

The numerical meanings of each indicator define the desired goals. With the establishment of the goals also defined are the accurate metrical values, which the indicators should assume in case the telecommunication company reaches its planned strategic goals. Within the model of heuristic and AI planning strategies there should be differentiated the indicators, which reflect the processes facilitating the obtaining of those results. Since reaching the one group of indicators fosters the realization of the others, there is an imperative need to bind those two groups of indicators.

The strategic initiatives are tactical measures, representing specific actions. They are an intrinsic component of the strategy planned with the assistance of the model and lead to reaching the goals.

The defining of the strategic initiatives as actions or a system of actions, leading to and facilitating the realization of the planned strategy, is a clear sign that in the course of the activity, they should implement the necessary connection between the strategic goals of the telecommunication company and its operational management. Behind each strategic initiative there should be planned not only an accurately formulated assignment and an action plan supporting its accomplishment, but also the optimum needed resources for reaching the required result.

The cause and effect relation binds all the tasks in the model of heuristic and AI planning strategies, both one to another, and also on the basis of the logical connection between the strategic and operational goals, plans and organizational capabilities. The total set of tasks and the preconditioned strategic and operational links between them form the map of the strategy. The cause and effect relation in the model requires that there should not be focusing only on some of the success elements, but covering the number of factors, which are compulsory for achieving that success.

3.3. Sequence of the activities in the model of heuristic and AI planning strategies

The planning of the strategy in a specific telecommunication company through the model of heuristic and AI planning strategies requires an accurate and adequate definition of the short-term, mid-term and long-term goals, which will be pursued. The implementation in the model of a mechanism for monitoring and maintaining the defined goals within the set target parameters and the differentiation of multi-variability of alternatives in planning the strategy of the telecommunication companies are the system-forming tasks in the model. The model of heuristic and AI planning strategies requires the choice and application of a strategic toolset, which should be adequate to the requirements of the contemporary strategic management model in Bulgarian telecommunication sector. On the basis of the already defined goals and the set situational parameters of the telecommunication sector and the particular telecommunication company, the model of heuristic and AI planning strategies should carry out:

- ✓ analysis and evaluation of the external environment;
- ✓ analysis of the company's strengths and weaknesses;
- ✓ analysis of the strategic alternatives;
- ✓ planning a strategy for achieving the already outlined goals;
- ✓ planning procedures, rules and budgets supporting the achievement of the goals.

Some analytical and forecasting information for the development of the external environment of the telecommunication company both at the stage of defining goals, and with its actual functioning, is acquired by the model of heuristic and AI planning strategies by performing permanent monitoring via:

- ✓ Political, Economic, Social, Technological Analysis;
- ✓ Industry Analysis;
- ✓ Competitive Analysis;

- Strategic analysis of Porter, Dewhurst and Burns;
- Competitor Profiling;
- Pressure Analysis;
- Key Success Factors Analysis.

The analytical techniques, which determine the planning by the model of heuristic and AI planning strategies and obtain information about the internal environment of the telecommunication company are:

- Strategic Product Analysis;
- Analysis of the Boston Consulting Group;
- Value - Chain Analysis;
- Ratio Analysis;
- Strategic Position Analysis;
- Vulnerability Analysis;
- Critical Success Factors Analysis;

The strategic research of the external and internal environment through identification and summarization by the model of heuristic and AI planning strategies of the most critical established trends, outlines the specific situation in which the telecommunication company functions. This situation requires the systematization and operationalization of the main steps through which the telecommunication company will generate and realize its advantage, i.e. to evaluate the position on which it is, and hence, plan a strategy. Such a more systematic and wide range of scanning can be done by the model of heuristic and AI planning strategies through the SWOT-analysis, which is clear, analytical, synthesized and flexible instrument for situational analysis.

The analysis of the strategic alternatives in a telecommunication company can be done through a matrix, which positions the organization in four sectors (table 1.), according to its competitive positions and its growth on the market.

Table 1. Matrix determining the positions of a telecommunication company

		COMPETITIVE POSITIONS	
		strong positions S	weak positions W
MARKET GROWTH	accelerated growth R	sector SR	sector WR
	delayed growth S	sector SS	sector WS

On the basis of the matrix and the combination of the explicit and implicit influence of internal or external operations, of related or non related actions, horizontal or vertical changes, proactive activities or lack of activities, there is a "tree of alternatives" being elaborated within the process of strategic management, which facilitates the search for the most appropriate strategic alternative (table 2.). Alternative options for planning a strategy are searched through it.

Table 2. Tree of the strategic alternatives in a telecommunication company

STRATEGIC GOALS	STRATEGIC ALTERNATIVES	FIELD OF CHANGE
FIRST STRATEGIC GOAL	increasing	size; scale
	decreasing	merger; take over
	concentrating	resources; manufacturing
	integrating	capital
	diversifying	products; markets
SECOND STRATEGIC GOAL	balancing	capacities; deliveries
	rationalization	operations; technological rules
THIRD STRATEGIC GOAL	modification	conditions; technological rules
	regrouping	working functions and/ or processes
	renewal	products; technologies
	recovery	finance; manufacturing
	reduction	expenses; manufacturing; personnel
FORTH STRATEGIC GOAL	transformation	separation; merging in
	isolating	subsystems
	liquidation	assets; participation

The model of heuristic and AI planning strategies as a particularly complex and dynamic process, integrating in itself strategic management through its main management functions, requires both planning of a strategy for achieving the set goals, and analyzing the reasons for the occurring of deviations and risks in the telecommunication company. The technology of this choice of strategy requires:

- ✓ identification of the strategic alternatives;
- ✓ choosing representative criteria or criterion;
- ✓ determining their importance;
- ✓ evaluating the usefulness of each alternative under the different criteria;
- ✓ calculating the weighed evaluation of the usefulness of each alternative;
- ✓ ordering the strategic alternatives according to their priority.

4. CONCLUSIONS

The conditions of contemporary management in the telecommunication sector require a new approach, which should get rid of the short-term concept and take management beyond the traditional planning of a strategy. This new approach needs the integration of information in respect to how well a telecommunication company copes with the challenges it is facing and how it will be positioned on the market tomorrow.

The model of heuristic and AI planning strategies plays the role of a strategic instrument for the solution of such management-challenges, where its introduction would lead to the reduction of:

- ✓ the time for gathering, processing, analyzing and evaluating the information necessary for the strategic planning and goal setting;
- ✓ the time for communication and decision making;
- ✓ the resource consumption;
- ✓ the risk.

The higher objectivity and reliability of the management process, the personalization of services, and reaching optimality and higher added value on that basis, not only in respect to management, but also in respect to the actual production process, are the other competitive advantages, which the telecommunication companies using the model of heuristic and AI planning strategies have. The arguments in favour of using this model in outlining the strategic alternatives of the telecommunication companies can be summarized primarily to the possibility to achieve a tangible competitive advantage and leadership in the sector through fast, adequate and reliable measuring, reporting and evaluating at the same time:

- ✓ the success of the business organization as a whole;
- ✓ the financial status and capacity;
- ✓ meeting clients' expectations;
- ✓ the structure of the internal processes;
- ✓ the development of the company in the future.

Through the model of heuristic and AI planning strategies in an abstract strategic model are combined the long-term and short-term goals in a single balanced evaluation of the generated strategic alternatives. This enables the obtaining of information, which is as complete and accurate as possible, concerning the strategic condition and market position of the telecommunication company, and achieving on that basis competitive advantage and leadership on the telecommunication market in Bulgaria.

In view of achieving leadership, the telecommunication company's management should be studied, analyzed and evaluated in a specific context, determined by the logic dictating the integration of the financial indicators with other aspects, illustrating the entire activity of the organization. The model of heuristic and AI planning strategies is the toolset, which providing a measuring instrument for the achievement of the goals, evaluates the most acceptable of them and plans the strategy by ensuring for the organization not only considerable competitive advantages, but also added value for it, for its users and investors.

The leadership achieved through the application of the model of heuristic and AI planning strategies identifies the telecommunication companies in the sector as such having indisputable competitive advantages, and the services offered by them are differentiated from the services of the competitors. The leadership in the sector of telecommunications is a result not only of a flexible, adaptive and creative strategy. It is related to and also results from the specific conditions of the environment in which the telecommunication sector exists and develops. On this basis, the leadership of the telecommunication companies in Bulgaria is determined by their capability to maintain their position on the market and improve their share over time by adding value not only in the financial aspect, but also in respect to their brand. This, however, is possible in the contemporary conditions only and solely if management relies on speed, adequateness, reliability and reduced resource consumption provided by the innovative toolset of the model of heuristic and AI planning strategies.

REFERENCES

[1] Annual report of the Communications Regulation Commission of Bulgaria for 2015.

[2] Temelkova, M. (2010) Controlling in the manufacturing. Color Print Inc.

[3] Temelkova, M. Korrelative Untersuchung der Faktoren, die die Unternehmensführung bestimmen. Journal L´Association 1901 "SEPIKE", Osthofen - Deutschland, Poitiers – France, Los Angeles – USA, Ausgabe 13/2016.

[4] Manoilov, G. (2008) Strategic maps – a means for synchronization between business and IT. CIO Magazine.

[5] Kaplan, R., Norton, D. (2001) The strategy-focused organization: How balanced scorecard companies thrive in the new business environment. Harvard business school publishing.

AUTHORS

Assoc. Prof. Dr. Miglena Temelkova was born in the city of Varna – Bulgaria. She graduated higher education in law and economics. She has specialized in international commercial law and international business. She has had training in Leadership at the Georgetown University in Washington. She has a Ph.D. degree in production organization and management. In 2010, she acquired the scientific title of Associate Professor. Since that same year, she has been the only certified in Bulgaria lecturer under the German GRID methodology for the training of leaders and she has established an unique for this country international Master's degree programme „Leadership in the Global Environment".

In 2014 she was the only Bulgarian representing Eastern Europe at an international forum in Washington, which was dedicated to issues of global leadership of the countries and business organizations, the crises of leadership and the tremors on a global scale. Assoc. Prof. Temelkova has more than 50 publications in Bulgarian, English, German and Russian in the field of leadership and management.

Prof. Dimitar Radev, DSc, PhD, is full professor at the Department of Telecommunications, UTP – Sofia. He is the author of numerous publications. His main research areas are modelling of communication networks, rare event simulation, quality of service, performance analysis of queuing systems and neural network modelling, BCI and artificial intelligence.

Strahil Sokolov, PhD is an assistant professor at the Department of Information technologies, UTP – Sofia. His research interests include biometric analysis, computer vision and AI, system design and innovative learning systems.

INTENTIONAL BLANK

HUMAN EMOTION ESTIMATION FROM EEG AND FACE USING STATISTICAL FEATURES AND SVM

Strahil Sokolov[1], Yuliyan Velchev[2], Svetla Radeva[3] and Dimitar Radev[4]

[1,3]Department of Information Technologies,
University of telecommunications and post, Sofia, Bulgaria
[2,4]Department of Telecommunications,
University of telecommunications and post, Sofia, Bulgaria

ABSTRACT

An approach is presented in this paper for automated estimation of human emotions from combination of multimodal data: electroencephalogram and facial images. The used EEG features are the Hjorth parameters calculated for theta, alpha, beta and gamma bands taken from pre-defined channels. For face emotion estimation PCA feature are selected. Classification is performed with support vector machines. Since the human emotions are modelled as combinations from physiological elements such as arousal, valence, dominance, liking, etc., these quantities are the classifier's outputs. The best achieved correct classification performance for EEG is about 76%. Classifier combination is used to return the final score for the particular subject.

KEYWORDS

Multimodal human emotion, EEG, arousal, valence, BCI

1. INTRODUCTION

Analysis of human emotions from multimodal data is still a challenging task for the researchers worldwide. Since the early nineties of the past century the research on human emotions has brought together clinical researchers and engineers who have tried to automate this task. Facial analysis is one of the most popular techniques for automatic estimation of human emotions.

In recent years, multimodal approaches for human emotion estimation have emerged. Researches have tried to incorporate the human electroencephalogram (EEG) signals for emotional state analysis and this represents a challenging area of modern research. Historically EEF processing was considered a traditionally medical area and now it is being used for emotion analysis. Recent advances in the area of EEG analysis are said to deliver promising emotion recognition results already. There still seem to be gaps in terms of stability and accuracy in those algorithms. This is what motivated us to research in the area to provide a framework for reliable EEG emotional state estimation with main application in the Brain-Computer Interfaces (BCI) domain. The human emotion is a highly subjective phenomenon: it has been accepted by phsycologists that multiple dimensions or scales can be used to categorize emotions. A two–dimensional model of emotion is introduced in [1] (Fig. 1). The valence axis represents the quality of an emotion ranging from

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 37– 47, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70204

unpleasant to pleasant. The arousal axis refers to the quantitative activation level ranging from calm to excited state.

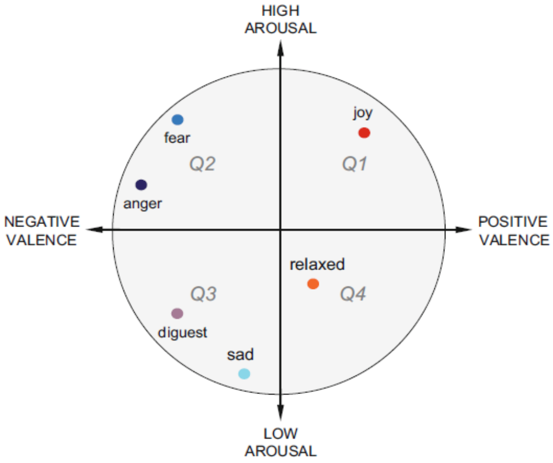


Figure 1. Two–dimensional human emotion model [2]

We have taken into consideration some of the most distinguished works in the area of EEG analysis. In [3] the authors used the EEG signal to classify two basic emotions: happiness and sadness. These emotions are evoked by showing subjects pictures that contain facial expressions of smile and cry. The authors propose a frequency band searching method to choose an optimal band into which the recorded EEG signal is filtered. They use Common Spatial Patterns (CSP) and linear Support Vector Machine (SVM) to classify these two emotions. To investigate the time resolution of classification, they explore two kinds of trials with lengths of 3 s and 1 s. Classification accuracies of 93.5% and 93% are achieved on 10 subjects for 3 s and 1 s trials, respectively. Their experimental results indicate that the gamma band (roughly 10 Hz to 30 Hz) is suitable for EEG-based emotion classification.

In [2] an approach is presented for EEG-based emotions recognition using time and frequency features. The time features are in fact some statistical quantities such as means and standard deviations of the raw signals and its first and second derivatives as well. The classified emotions are: joy, relax, sad and fear. The best accuracy is about 66% among three types of classifiers.

In [4] the achieved accuracy of the emotional valence is about 71%. The technique relies on changes in the power spectrum of short-time stationary oscillatory EEG processes within the standard EEG frequency bands. The classification stage is logistic regression with elastic-net regularization. Another interesting approach is presented in [5]. The features are extracted from very limited set of electrodes and the dimensionality is further reduced with Principal Component Analysis (PCA). The performance is evaluated for arousal, valence and modality separately. As can be expected the arousal is with highest classification accuracy (over 90%). A classification rate of 83.3% is reported in [6]. The authors have experimented with different classifiers and discrete wavelet transform as feature extraction technique. Very important results in terms of correlations between EEG waves activities from certain brain regions and the human emotions during a game play are published in [7]. These correlations are considered in the features extraction stage in our paper. In [8] is given an approach for emotions recognition using brain activity. The following types of features are used: EEG frequency band power, cross-correlation between EEG band powers, peak frequency in alpha band and Hjorth parameters.

The performances for classification in five classes is above 30%. These results are indicative for a significant variability of EEG-based features for emotion recognition among different subjects.

However, a reliable recognition of the extreme human emotions is still possible considering its distances in the two-dimensional model (Figure 1.).

Analysis of facial emotions has been in research for quite a while. There are a few works which are worth mentioning. The authors of [17] describe an approach using a combination of higher-order auto-correlation and fisher weights to classify emotions. The highest accuracy was with the combined method and reached about 97.9%. In [13] an approach is presented where EEG and peripheral physiological signals from 12 participants had been recorded. The responses to emotional videos were classified with SVM into five classes namely, joy, sadness, disgust, fear, and relax. The achieved accuracy of 41.7% using EEG signals has been reported. Feature level combination of EEG signals and peripheral physiological signals failed to improve the classification accuracy.Combined or multimodal estimation of face emotions is a fairly new area of research. The authors of [15] propose the use of EEG and face analysis from video to detect emotion. Report on the accuracy has not been published. In [16] the authors combine EEG and eye gaze data to classify emotions. Based on single modality classification of arousal and valence in three classes, the authors obtained 62.1% and 50.5% accuracy eye gaze data. The authors proposed feature-level fusion and decision-level score fusion of the results and were able to improve accuracy to about 76% for arousal and 68.5 %for valence.

The outline of the paper is organized as follows: in section 2 an overview is provided of the EEG-based emotion recognition including EEG features extraction, features selection and classification. In section 3 we provide details for the proposed approach for face data processing. In section 4 we describe the combination of classifier results. Section 5 describes the evaluation of EEG signals processing with corresponding emotional quantities, as well as the facial data processing and classifier combination. The last section discusses the results and marks out some aspects for future work.

2. EEG EMOTION ESTIMATION

The EEG-based human emotion recognition system is shown in Fig. 2. This paper does not cover the emotion model. As can be seen the dominance and liking are optional in the final emotion estimation.

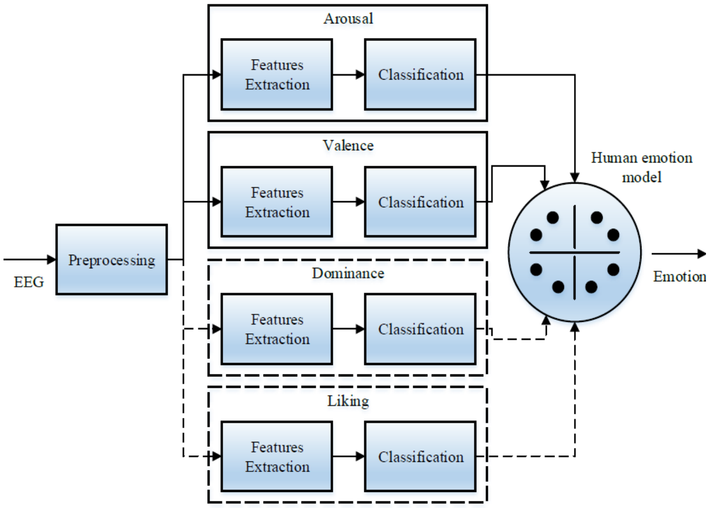


Figure 1. EEG-based human emotion recognition system

2.1. EEG Preprocessing

We have increased the topographical localization of the EEG current sources using the scalp surface Laplacian spatial filtering. We have used the spherical spline method [9] that is the human head is modelled as a sphere. The spatial electrode coordinates are as suggested in the CSD toolbox implementation [9]. The parameter for spline flexibility is set to its default value of 4.

2.2. EEG Features Extraction

Extraction of the features is realized via the EEG activities: θ (frequency range from 4 Hz to 7 Hz), α (8 Hz to 13 Hz), β (14 Hz to 29 Hz) and γ (30 Hz to 45 Hz). The used filter banks are realized as zero-phase Butterworth IIR.

EEG Channels Selection: Some researches confirm the positive valence and happy emotions are connected with an increase of frontal spectral coherence mainly in alpha band. Also the right parietal beta power increases. The higher arousal (excitation) causes an increase of coherence in parietal EEG and an increase of beta waves power. The alpha activity decreases as well. The emotion's strength is related to higher beta/alpha activity ratio in the frontal lobe. The beta activity at the parietal lobe is also increased. Considering the correlations found in [10] (Fig. 3) for arousal estimation we have used CP6, Cz, Fz and FC2 electrodes. The chosen electrodes for valence estimation are: Oz, PO4, CP1, FC6, Cz and T8.

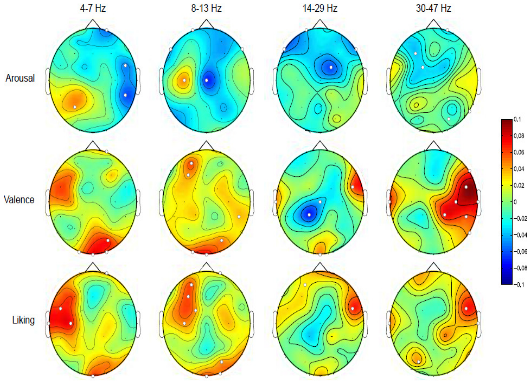


Figure 3. The mean correlations of the valence, arousal and general ratings with the power in the EEG waves [10]

1) The Feature Vectors: For a given EEG channel c and band b a feature vector is composed as follows:

$$\mathbf{f}_{c,b} = [Act_{c,b}, Mob_{c,b}, Cpl_{c,b}], \tag{1}$$

where $Act_{c,b} = var\left(j_{c,b}\left(t\right)\right)$, $Mob_{c,b} = \sqrt{\frac{var\left(\frac{dj_{c,b}}{dt}\right)}{Act_{c,b}}}$ and $Cpl_{c,b} = \sqrt{\frac{var\left(\frac{d^2 j_{c,b}}{dt^2}\right)}{var\left(\frac{dj_{c,b}}{dt}\right)} - Mob_{c,b}^2}$ are the Hjorth parameters [11] (known as activity, mobility and complexity). For arousal estimation the feature vector is organized as augmentation of $\mathbf{f}_{CP6,\theta}$, $\mathbf{f}_{Cz,\alpha}$, $\mathbf{f}_{FC2,\beta}$ and $\frac{Act_{Fz,\beta}}{Act_{Fz,\alpha}}$. For valence estimation the feature vector consists of $\mathbf{f}_{Oz,\theta}$, $\mathbf{f}_{PO4,\alpha}$, $\mathbf{f}_{CP1,\beta}$, $\mathbf{f}_{FC6,\gamma}$, $\mathbf{f}_{Oz,\beta}$, $\mathbf{f}_{Cz,\beta}$, $\mathbf{f}_{T8,\gamma}$ and $\mathbf{f}_{FC6,\gamma}$

2.3. EEG Features Selection

Among many methods for features selection, we have chosen the popular Minimum Redundancy and Maximum Relevance (mRMR) criterion [12]. The relevance RL of the set of selected features $F = \{f_1, f_2, \dots\}$ and target classes C is:

$$RL = \frac{1}{|F|} \sum_{f_i \in F} I\left(f_i, C\right), \tag{2}$$

where I denotes the mutual information. The redundancy RD of the features is:

$$RD = \frac{1}{|F|^2} \sum_{f_i, f_j \in F} I\left(f_i, f_j\right). \tag{3}$$

For incremental search $\max\left[I\left(F, C\right)\right]$ is equivalent to $\max\left[RL\left(F, C\right) - RD\left(F\right)\right]$ [12].

Our second suggestion for features is mRMR selection from all possible sets of ratios $\frac{Act_{c,b}}{Act_{c,b}}, c \neq k$ and $\frac{Mob_{c,b}}{Mob_{k,b}}, c \neq k$ where c and k denote the EEG channel and b is the activity (θ, α or β).

2.4. Classification

Since the features for arousal and valence are very different, we have trained two classifiers of type SVM. We suggest classifying only the prime emotional quantities (arousal, valence, etc.). The emotions can be further inferred using and appropriate emotion model (two or three-dimensional) (Figure 2.).

3. Facial Emotion Estimation

3.1. Data acquisition and preprocessing

The proposed approach for face detection and validation is based on our previous research [19]. It utilizes the OpenCV face detection algorithm [18] and a convolutional neural network.

The cascade structure of the face is intended for higher speeds of image processing due to the fast background rejection and focus on face-like regions. But in comparison with the monolithic classifiers, the cascade classifier (for example, Haar-like features' cascade of weak classifiers [18]) increases the detection error and false alarms rate. Therefore, it is necessary to join the rapid cascaded classifier with the accurate monolithic one within the two-level combined cascade of classifiers instead of using them independently. This is realized in order to achieve higher detection and lower false alarm rates. The two-level cascade of classifiers is called "combined" since it combines different types of classifiers, which have been proved in the course of time: the first level is represented by the Haar-like features' cascade of weak classifiers, which is responsible for the face-like objects detection, and the second level is a convolutional neural network for the objects' verification (Figure 4)

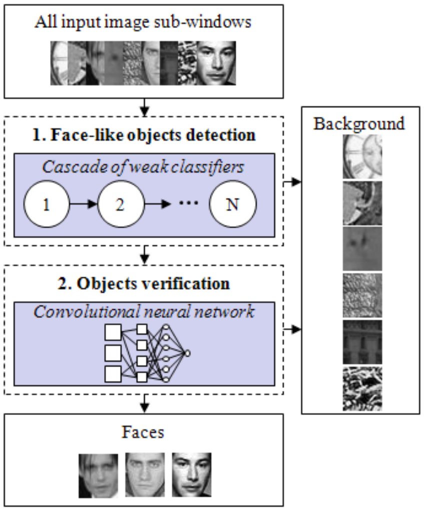

All input image sub-windows
1. Face-like objects detection
Cascade of weak classifiers
1
2
N
2. Objects verification
Convolutional neural network
Faces
Background

Figure 4. Face detection process using combined cascade of neural network classifiers

In this phase the fairly fast face detector is also able to deliver faces in frontal pose. This depends on the training set of images for the CNN. In our approach this has proven to be useful since we are using short-length videos of the subjects' faces may have slight fluctuations off the frontal pose.

3.2. Facial features extraction

The usage of PCA is proposed for facial features extraction. PCA has the property of packing the greatest energy into the least number of principal components. It computes the basis of a space which is represented by the training images as vectors. These basis vectors are computed by solution of an eigenproblem, and as such the basis vectors are eigenvectors. These eigenvectors are defined in the image space. They can be treated as images and indeed resemble faces. Hence they are usually referred to as eigenfaces. The components of the projected images in the eigenspace are characterised by statistical independence.

PCA is a linear transformation of type $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$. $\mathbf{X}$ is the matrix of the normalized faces, extracted from the video of the person; each face image is represented as a row vector. The major step in PCA is determining the basis $\mathbf{W}$. This basis is calculated by solving the problem of eigenvalues and eigenvectors of covariance matrix of $\mathbf{X}$:

$$\mathbf{C}_{\mathbf{X}} = E\{(\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}}\} \tag{4}$$

$\overline{\mathbf{X}}$ is a matrix with each row representing the average of all faces, the so called *mean face*. The solution of the eigenvalue problem is stated as:

$$\mathbf{C}_{\mathbf{X}}\mathbf{W} = \mathbf{W}\Lambda \tag{5}$$

Each of the columns of $\mathbf{W}$ represents a normalized eigenvector of $\mathbf{C}_{\mathbf{X}}$ and Λ is a diagonal matrix with the eigenvalues on the main diagonal. Solving (5) can be done by finding roots of characteristic polynomials or by Jacobi rotation. Dimensionality reduction can be achieved by first rearranging eigenvalues in descending order and then rearranging $\mathbf{W}$ to match Λ. Next, the dimension of the eigenspace is calculated by the criteria normalized Residual Mean Square Error:

$$RMSE(b) = \sum_{i=1}^{b} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i > T \tag{6}$$

where T is a threshold representing the fraction of the power of the signal, that must be preserved in the output space. Dimensionality reduction is an intuitive approach, since there exists a lot of statistical redundancy in natural images. Using the reduced covariance matrix is another possibility for reducing dimensionality even before resorting to (6).

4. COMBINING CLASSIFIER RESULTS

The combined classification that we propose is expected to increase accuracy and reduce noise using two parallel modules (distributed system). The output of the combined-modality biometric systems of type Multi-Sample-Multi-Source is the average score form the outputs of both modalities according to [20]:

$$y_{combined} = \frac{1}{M} \sum_{j=1}^{M} \left[\frac{1}{N} \sum_{i=1}^{N} \left(y_{i,j} \right) \right]. \tag{7}$$

where $y_{combined}$ is the combined output score, M is the number of modalities used (EEG and Face), N is the number of samples for each biometric modality, $y_{i,j}$ is the output of the system for the *i-th* sample from *j-th* modality and $y_{i,j} \in [0,...,1]$.

5. EXPERIMENTAL RESULTS

5.1. The Used Dataset

For our experiments we have used the DEAP dataset [10]. It consists of multimodal data physiological including EEG signals taken from 32 leads. Each subject participates with 40 trials with duration of 60 s. The EEG signals were downsampled to 128 Hz, bandpass filtered (4 Hz to 45 Hz) and the EOG artefacts were removed as suggested in [10]. The data was averaged to the common reference.

5.2. Performance Evaluation

We have used the accuracy as criterion for classification performance:

$$Acc = \frac{TP + TN}{P + N} \tag{8}$$

where TP denotes the number of true positives, TN is the number of true negatives. The denominator in (8) denotes the total population of positives P and negatives N .

The performance is validated and evaluated using the k-fold technique. That is, a testing part is extracted from the whole dataset. The rest of the dataset is used to train the classifier. This procedure repeats (10 times in our case) and the accuracy is calculated as an average of the accuracies in the iterations. The arousal and valence scores in dataset are given as fractional numbers ranging from 1 to 9. We have quantized these scores to 3, 5 and 7 levels and the testing was performed for each case. The accuracies of the estimated arousal are given in Table 1. Respectively the results for valence are summarized in Table. II.

Table 1. Classification Performance for Arousal with Feature Vectors Built According to (8)

Quantization Levels	**Accuracy, %**
3	77.5
5	63.2
7	34.5

Table 2. Classification Performance for Valence with Feature Vectors Built According to (1)

Quantization Levels	**Accuracy, %**
3	75.2
5	57.2
7	25.5

As an alternative some experiments have been performed using the ratios for activities and mobilities for theta, alpha and beta waves. In this case the features have been selected using the mRMR method. The first 20 features for arousal and valence are given in Table. III. In Fig. 4 can be seen the calculated classification accuracies versus dimensionality of the feature vectors. Not all possible sets of features have been tested since the number of combinations is too large for an average computational machine.

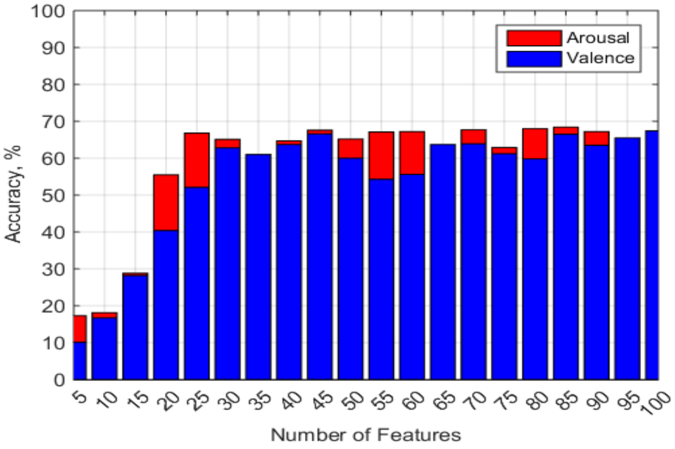


Figure 4. Classification accuracies for arousal and valence versus dimensionality of the feature vectors. The feature selection is according to mRMR method

Table 3. The 20 Most Discriminative Features for Arousal and Valence

Index	**Arousal**	**Valence**
1	$\frac{Act_{Fp2,\theta}}{Act_{C4,\beta}}$	$\frac{Act_{Fp2,\theta}}{Act_{C4,\beta}}$
2	$\frac{Mob_{Fp1,\theta}}{Mob_{Fp1,\alpha}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{Fp1,\alpha}}$
3	$\frac{Mob_{Fp1,\theta}}{Mob_{Fp1,\beta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{Fp1,\beta}}$
4	$\frac{Mob_{Fp1,\theta}}{Mob_{AF3,\alpha}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{AF3,\alpha}}$
5	$\frac{Mob_{T7,\alpha}}{Mob_{AF4,\theta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{AF3,\beta}}$
6	$\frac{Mob_{Fp1,\theta}}{Mob_{AF3,\beta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{F3,\alpha}}$
7	$\frac{Act_{C4,\beta}}{Act_{P4,\theta}}$	$\frac{Mob_{F7,\alpha}}{Mob_{Oz,\theta}}$
8	$\frac{Mob_{F7,\alpha}}{Mob_{T7,\theta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{F3,\beta}}$
9	$\frac{Mob_{Fp1,\theta}}{Mob_{F3,\alpha}}$	$\frac{Act_{Oz,\theta}}{Act_{AF4,\theta}}$

10	$\frac{Act_{CP5,\theta}}{Act_{Fp2,\theta}}$	$\frac{Mob_{Fp1,\alpha}}{Mob_{P8,\theta}}$
11	$\frac{Mob_{Fp1,\theta}}{Mob_{F3,\beta}}$	$\frac{Mob_{T7,\alpha}}{Mob_{P4,\theta}}$
12	$\frac{Mob_{Fp1,\theta}}{Mob_{F7,\alpha}}$	$\frac{Act_{CP5,\theta}}{Act_{Fp2,\beta}}$
13	$\frac{Act_{FC1,\alpha}}{Act_{T8,\theta}}$	$\frac{Mob_{Fp2,\alpha}}{Mob_{F4,\theta}}$
14	$\frac{Act_{PO3,\beta}}{Act_{Pz,\theta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{F7,\alpha}}$
15	$\frac{Mob_{Fp1,\theta}}{Mob_{F7,\beta}}$	$\frac{Act_{FC1,\beta}}{Act_{P4,\theta}}$
16	$\frac{Mob_{Cz,\alpha}}{Mob_{T8,\theta}}$	$\frac{Mob_{F7,\alpha}}{Mob_{C3,\theta}}$
17	$\frac{Act_{AF3,\theta}}{Act_{Pz,\beta}}$	$\frac{Mob_{P3,\beta}}{Mob_{Pz,\theta}}$
18	$\frac{Mob_{Fp1,\theta}}{Mob_{FC5,\alpha}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{F7,\beta}}$
19	$\frac{Mob_{CP1,\alpha}}{Mob_{F4,\theta}}$	$\frac{Act_{Fp1,\theta}}{Act_{Pz,\beta}}$
20	$\frac{Mob_{Fp1,\theta}}{Mob_{FC5,\beta}}$	$\frac{Mob_{Fp1,\theta}}{Mob_{FC5,\alpha}}$

An exact comparison of accuracies between our approach and those achieved by other researchers is not possible because of the different testing dataset. Also the chosen outputs of the classifiers are not the same (different levels of quantization for arousal and valence). Nevertheless, an approximate comparison can be seen in Table 4.

Considering the accuracies in this study and those published by other researchers, the practical use of EEG-based emotion recognition is limited in terms of distinguishing between some very basic emotional states such as joy, anger, sadness and relax. The channels selected using mRMR method in general do not match those suggested by the other studies. That confirms the poor subject–independence of the EEG features for emotions recognitions. Face emotion estimation runs in parallel and contributes to the improvement of the scores generated by the EEG analysis module. The proposed PCA-based face emotion estimation provides an additional improvement for the EEG signal analysis.

Table 4.Classification Performance of the Presented Approach Compared with Results of Related Works

Approach	**Average accuracy for arousal and valence, %**
In [5]	96.2
In [8]	36.5
In [2]	66.5
In [4]	71.3
In this paper	76.4

6. CONCLUSIONS

This work presented an approach for automated multimodal EEG and face-based estimation of human emotions. The accuracy was investigated for different levels of EEG arousal and valence. For EEG, the maximal accuracy of 76.4% is achieved when the arousal and valence is classified in only three levels. The improvement of the classification is delivered via parallel face emotion analysis system based on PCA. The classifiers that we have used are SVM and we propose score-level decision fusion. In our future work we will seek to implement future improvement of the BCI which we are developing. We are considering the usage of Active Appearance Models as well as 3D facial emotion recognition as further improvement in order to achieve our objective and create HCI for people with motor disabilities.

ACKNOWLEDGEMENTS

This work is part of a research project funded by the National Scientific Research Fund at the Bulgarian Ministry of Education and Science DFNI I02/2014 “Human-computer interface for medical assisting systems for life improving of people with propelling problems”.

REFERENCES

[1] R. Davidson, G. Schwartz, C. Saron, J. Bennett, and D. Goleman, “Frontal versus parietal eeg asymmetry during positive and negative affect,” Psychophysiology, vol. 16 (2), pp. 202–203, 1979.

[2] X.-W. Wang, D. Nie, and B.-L. Lu, “EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines,” ICONIP’11 Proceedings of the 18th international conference on Neural Information Processing, vol. Part I, pp. 734–743, 2011.

[3] M. Li and B.-L. Lu, “Emotion classification based on gamma-band EEG,” Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pp. 1223–1226, 2009.

[4] C. Kothe, S. Makeig, M. Soleymani, and J. Onton, “Emotion Recognition from EEG During Self-Paced Emotional Imagery,” 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 855–858, 2013.

[5] D. Oude Bos, “EEG-based emotion recognition-The Influence of Visual and Auditory Stimuli,” Capita Selecta (MSc course), 2006.

[6] M. Murugappan, N. Ramachandran, and Y. Sazali, “Classification of human emotion from EEG using discrete wavelet transform,” J. Biomedical Science and Engineering, vol. 3, pp. 390–396, 2010.

[7] B. Reuderink, C. Mühl, and M. Poel, “Valence, arousal and dominance in the EEG during game play,” International Journal of Autonomous and Adaptive Communications Systems, vol. 6 (1), pp. 44–62, 2013.

[8] R. Horlings, D. Datcu, and L. J. M. Rothkrantz, “Emotion Recognition using Brain Activity,” International Conference on Computer Systems and Technologies-CompSysTech’08, vol. II, pp. 1–6, 2008.

[9] J. Kayser and C. Tenke, “Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks,” Clinical Neurophysiology, vol. 117 (2), pp. 348–368, 2006.

[10] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “DEAP: A Database for Emotion Analysis using Physiological Signals,” IEEE Transactions on Affectiv Computing, vol. 3 (1), pp. 18–31, 2012.

[11] B. Hjorth, “EEG analysis based on time domain properties,” Electroencephalography and Clinical Neurophysiology, vol. 29 (3), p. 306–310, 1970.

[12] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27 (8), pp. 1226–1238, 2005.

[13] Takahashi, Kazuhiko. "Remarks on SVM-based emotion recognition from multi-modal bio-potential signals." In Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on, pp. 95-100. IEEE, 2004.

[14] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In Proceedings of the fourth IEEE International conference on automatic face and gesture recognition (FG'00), pp. 46–53, Grenoble, France,2000.

[15] SAVRAN, Arman, et al. Emotion Detection in the Loop from Brain Signals and Facial Images. In: Proceedings of the eNTERFACE 2006 Workshop. 2006.

[16] Soleymani, M., Pantic, M. and Pun, T., 2012. Multimodal emotion recognition in response to videos. IEEE transactions on affective computing, 3(2), pp.211-223.

[17] Shinohara, Yusuke, and N. Otsuf. "Facial expression recognition using fisher weight maps." In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pp. 499-504. IEEE, 2004

[18] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57, no. 2 (2004): 137-154.

[19] Paliy, Ihor, Anatoly Sachenko, Yuriy Kurylyak, Ognian Boumbarov, and Strahil Sokolov. "Combined approach to face detection for biometric identification systems." In Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009. IDAACS 2009. IEEE International Workshop on, pp. 425-429. IEEE, 2009.

[20] Poh, N., Bengio, S., Korczak, J., A Multi-sample Multi-source Model for Biometric Authentication. Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing, 375-384.

AUTHORS

Strahil Sokolov, PhD is an assistant professor at the Department of Information technologies, University of Telecommunications and Post –Sofia

Yuliyan Velchev, is PhD at the Department of Telecommunications, University of Telecommunications and Post –Sofia.

Prof. Svetla Radeva, DSc, PhD, is full professor at the Department of Information technologies, University of Telecommunications and Post –Sofia.

Prof. Dimitar Radev, DSc, PhD, is full professor at the Department of Telecommunications, University of Telecommunications and Post –Sofia.

INTENTIONAL BLANK

DATA SHARING TAXONOMY RECORDS FOR SECURITY CONSERVATION

Rajeswari Chandrasekaran[1] and Chandrasekaran Nammalwar[2]

[1,2]Faculty of Computing,
Botho University, Gaborone, Botswana

ABSTRACT

Here, we discuss the Classification is a fundamental problem in data analysis. Training a classifier requires accessing a large collection of data. Releasing person-specific data, such as customer data or patient records, may pose a threat to an individual's privacy. Even after removing explicit identifying information such as Name and SSN, it is still possible to link released records back to their identities by matching some combination of non identifying attributes such as {Sex,Zip,Birthdate}. A useful approach to combat such linking attacks, called k-anonymization is anonymizing the linking attributes so that at least k released records match each value combination of the linking attributes. Our goal is to find a k-anonymization which preserves the classification structure. Experiments of real-life data show that the quality of classification can be preserved even for highly restrictive anonymity requirements.

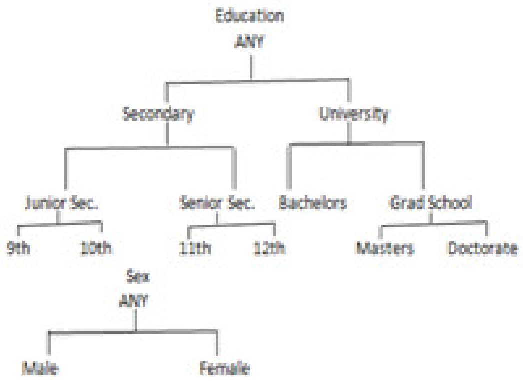
KEYWORDS

Privacy protection, Anonymity, Security integrity, Data mining, classification, Data sharing.

1. INTRODUCTION

DATA sharing in today's globally networked systems poses a threat to individual privacy and organizational confidentiality. An example by Samarati [2] shows that linking medication records with a voter list can uniquely. Identify a person's name and medical information. New privacy acts and legislations are recently enforced in many countries. In 2001, Canada launched the Personal Information Protection and Electronic Document Act [3] to protect a wide spectrum of information, such as age, race, income, evaluations, and even intentions to acquire goods or services. This information spans a considerable portion of many databases. Government agencies and companies have to revise their systems and practices to fully comply with this act in three years. Consider a table T about a patient's information on Birthplace, Birth year, Sex and Diagnosis. If a description on fBirthplace; Birth year; Sexg is so specific that not many people match it, releasing the table may lead to linking a unique record to an external record with explicit identity, thus identifying the medical condition and compromising the privacy rights of the individual [2]. Suppose that the attributes Birthplace, Birth year, Sex and Diagnosis must be released (say, to some health research institute for research purposes). One way to prevent such linking is masking the detailed information of these attributes as follows:

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 49– 58, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70205



1. If there is a taxonomical description for a categorical attribute (for example, Birthplace), we can generalize a specific value description into a less specific but semantically consistent description. For example, we can generalize the cities San Francisco, San Diego, and Berkeley into the corresponding state California.

2. If there is not taxonomical description for a categorical attribute, we can suppress a value description to a “null value” demoted? For example, we can suppress San Francisco and San Diego to the null value? While keeping Berleley.

3. If the attribute is a continuous attribute (for example, Birth year), we can discredited the range of the attribute into a small number of intervals. For example, we can replace specific Birth year values from 1961 to 1965 with an interval [1961-1966].

By applying such masking operations, the information on fBirthplace; Birth year; Sexg is made less specific, and a person tends to match more records. For example, a male born in San Francisco in 1962 will match all records that have the values HCA; {1961-1966}; Mi; clearly, not all matched records correspond to the person. Thus, the masking operation makes it more difficult to tell whether an individual actually has the diagnosis in the matched records. Protecting privacy is one goal. Making the released data useful to data analysis is another goal. In this paper, we consider classification analysis [4].

2. PROBLEM STATEMENT

A data provider wants to release a person-specific table T (d1;……….Dm; Class) to the public for modeling the class label Class. Each Di is either a categorical or a continuous attribute. A record has the form hv1… vm;clsi, where vi is a domain value for Di and cls is a class for Class. Att(v) denotes the attribute of a value v. The data provider also wants to protect against linking an individual to sensitive information either within or outside T through some identifying attributes, called QID. A sensitive linking occurs if some value of the QID identifies a “small” number of records in T. This requirement is formally defined below. Definition 1 (anonymity requirement). Consider p QIDs QID1;……;QIDp on T. a(qidi) denotes the number of data records in T that share the value qidi of QIDi. The anonymity of QIDi, denoted A(QIDi), is the smallest a(qidi) for any value qidi on QIDi. A table T satisfies the anonymity requirement fhQID1; k1i…hQIDp; kpig if A(QIDi) _ki for 1_i_p, where ki is the anonymity threshold on QIDi specified by the data provider. It is not hard to see that if QIDj is a subset of QIDi, A(QIDi)_A(QIDj). Therefore, if kj_ki, A(QIDi)_ki implies A(QIDj)_kj, and hQIDj; kji can be removed in the presence of hQIDi; kii. Following a similar argument, to prevent a linking through any QID, that is, any subset of attributes in QID1 [----] QIDp, the single QID [---]QIDp and $k^{1/4}$ maxfkjg, can be specified. However, a table satisfying fhQID1; k1i…hQIDp; kpig does not have to satisfy hQID; ki.

2.1 Masking Operations

A. Generalize Dj if Dj is a categorical attribute with a taxonomy tree. A leaf node represents a domain value and a parent node represents a less specific value. Fig. 2 shows a taxonomy tree for Education. A generalized Dj can be viewed as a “cut” through its taxonomy tree. A cut of a tree is a subset of values in the tree, denoted Cutj, which contains exactly one value on each root-to-leaf path. This type of generalization does not suffer from the interpretation difficulty discussed in Section 1.

B. Suppress Dj if Dj is a categorical attribute with not taxonomy tree. The suppression of a value on Dj means replacing all occurrences of the value with the special value?j. All suppressed values on Dj are represented by the same?j, which is treated as a new Supj to denote the set of values suppressed by ?j. This type of suppression is at the value level in that Supj is in general, a subset of the values in the attribute Dj.

C. Discredited Dj if Dj is a continuous attribute. The discretization of a value v on Dj means replacing all occurrences of v with an interval containing the value. Our algorithm dynamically grows a taxonomy tree for intervals at runtime, where each node represents an interval, and each nonleaf node has two child nodes representing some “optional” binary split of the parent interval. More details will be discussed in Section 3. A discredited Dj can be represented by the set of intervals, denoted Intj, corresponding to the leaf nodes in the dynamically grown taxonomy tree of Dj.

Definition 2 (Anonymity for Classification). Given a table T, an anonymity requirement fhQID1; k1i…hQIDp; kpig and an optional taxonomy tree for each categorical attribute contained in [QIDi mask T on the attributes [QIDi to satisfy the anonymity requirement while preserving the classification structure in the data (that is, the masked table remains useful for classifying the Class column). A masked table T can be represented by h[Cutj; [Supj;] Inti; where Cutj, Supj, and Intj are defined as above. If the masked T satisfied the anonymity requirement, h[Cutj; [supj; [Intji is called a solution set.

3. SEARCH METHODS

A table T can be masked by a sequence of refinements starting from the most masked state in which each attribute is either generalized to the topmost value, suppressed to the special value? Our method iteratively refines a masked value selected from the current set of cuts, suppressed values, and intervals, until violating the anonymity requirement. Each refinement increases the information and decreases the anonymity since records with specific values are more distinguishable. The key is selecting the “best” refinement at each step with both impacts considered.

3.1. Modifications

Below, we formally describe the notion of refinement on different types of attributes Dj 2[QIDi and define a selection criterion for a single refinement.

3.1.1 Refinement for Generalization

Consider a categorical attribute Dj with a user-specified taxonomy tree. Let child(v) be the set of child values of v in a user-specified taxonomy tree. A refinement, written v! Child (v) replaces the parent value v with the child value in child (v) that generalized the domain values in each (generalized) record.

3.1.2 Refinement for Suppression

For a categorical attribute Dj without taxonomy tree, a refinement! fv;? Jg refers to disclosing one value v from the set of suppressed values Supj. Let R?j denotes the set of suppressed records that currently contain? j. Disclosing v means replacing? j with v in all records in R?j that originally contain v.

3.1.3 Refinement for Discretization

For a continuous attribute, refinement is similar to that for generalization except that no prior taxonomy tree is given and the taxonomy tree has to be grown dynamically in the process of refinement. Initially, the interval that covers the full range of the attribute forms the root. The refinement on an interval v, which is written as v! Child (v) refers to the optimal split of v into two child intervals child (v) that maximizes the information gain. The anonymity is not used for finding a slit good for classification. This is similar to defining a taxonomy best describes the application. Due to this extra step of identifying the optimal split of the parent interval, we treat continuous attributes separately from categorical attributes with taxonomy trees.

A refinement is valid (with respect to T) if T satisfied the anonymity requirement after the refinement. A refinement is beneficial (with respect to T) if more than one class is involved in the refined records. A refinement is performed only if it is both valid and beneficial. Therefore, a refinement guarantees that every newly generated qid has a (qid) _k.

3.2 Selection Criterion

We propose a selection criterion for guiding our TDR process to heuristically maximize the classification goal. Consider a refinement v ! Child (v), where v 2 Dj and Dj is a categorical attribute with a user-specified taxonomy tree or Dj is a continuous attribute with a dynamically grown taxonomy tree. The refinement has two effects: it increases the information of the refined records with respect to classification, and it decreases the anonymity of the refined records with respect to privacy. These effects are measured by “information gain”, denoted AnonyLoss(v). v is a good candidate for refinement if InfoGain(v) is large and AnonyLoss(v) is small. Our selection criterion is choosing the candidate v, for the next refinement, that has the maximum informationgain/ anonymity-loss trade-off. To avoid division by zero, 1 is added to AnonyLoss(v). Each choice of InfoGain(v) and anonyLoss(v) gives a trade-off between classification and anonymixation. It should be noted that Score is not a goodness metric of k-anonymixation. In fact, it is difficult to have a closer-form metric to capture the classification goal (on future data). We achieve this goal through this heuristic selection criterion. For concreteness, we borrow Shannon’s information theory to measure information gain [26]. Let Rv denote the set of records masked to the value v, and let Rc denote the set of records masked to a child value c in child(v) after refining v. Let jxj be the number of elements in a set of x.

AnonyLoss(v): Defined as AnonyLoss(v) ¼ avgfA(QIDj)_AV(QIDj)g; (4) where A(QIDj) and AV(QIDj) represent the anonymity before and after refining v. avgfA(QIDj)_AV(QIDj)g is the average loss of anonymity for all QIDj that contain the attribute of v if Dj is a categorical attribute without taxonomy tree, the refinement ?j ! fv;? jg means refining R?j into Rv and R0 ?j, where

R?j denotes the set of records containing?j before the refinement. Rv and R0? J denote the set of records contain v and? j after the refinement, respectively. We employ the same Score(v) function to measure the goodness of the refinement?j ! fv;?jg.

3.3 InfoGain versus Score

An alternative to Score is using InfoGain alone, that is, maximizing the information gain produced by a refinement without considering the loss of anonymity. This alternative may pick a candidate that has a large reduction in anonymity, which may lead to a quick violation of the anonymity requirement, thereby, prohibiting refining the data to a lower granularity. Table 2b shows the calculated InfoGain, AnonyLoss, and Score of the three candidate refinements. According to the InfoGain criterion, ANY Edu will be first refined because it has the highest InfoGain. The result is shown in Table 2c with A(QID)1/4 4. After that, there is no further valid refinement because refining either ANY Sex or [1-99] will result in a violation of 4-anonymity. Note that the first 24 records in the table fail to separate the 4N from the other 20Y. In contrast, according to the Score criterion, ANY Sex will be first refined. The result is shown in Table 2d, and A(QID)1/4 14. Subsequently, further refinement on ANY Edu is invalid because it will result in a (h9th;M;$^{1/2}$1_99)i)1/4 2<k, but the refinement on [1-99] is valid because it will result in A(QID) ¼ 6_k. The final masked table is shown in Table 2e where the information for separating the two classes is preserved. Thus, by considering the information / anonymity trade-off, the Score criterion produces a more desirable sequence of refinements for classification.

4. TOP DOWN REFINEMENT

4.1 The Algorithm

We present our algorithm TDR. In a preprocessing step, we compress the given table T by removing all attributes not in [QIDi and collapsing duplicates into a single row with the Class column storing the class frequency as in Table 1. The compressed table is typically much smaller than the original table. Below, the term “data records” refers to data records in this compressed form. There exists a masked table satisfying the anonymity requirement if and only if the most masked table does that is, jTj_k. this condition is checked in the preprocessing step as well. To focus on main ideas, we assume that jTj_k and the compressed table first in the memory. In Section 4.5, we will discuss the modification needed if the compressed table does not fit in the memory.

Algorithm: [Top-down Refinement (TDR)]

1. Initialize every value of Dj to the topmost value, suppress every value of dj to ?j, or include every continuous value of Dj into the full-range interval, where Dj 2[QIDi].

2. Initialize cut j of Dj to include the topmost value, Supj of Dj to include all domain values of Dj, and Intj of Dj to include the full-range interval, where Dj 2[QIDi].

3. While some x 2 h [cut j]; [Supj]; [Intj is valid and beneficial].

4. Find the Best refinement from h[cutj]; [Supj]; [Intji].

5. Perform Best on T and update h[cutj]; [supj]; [Intji].

6. Update Score(x) and validity for x 2 h[cutj]; [supj]; [Intji].

7. End while

8. Return Masked T and h[cutj; [Supj; [Intji].

High level description of our algorithm. Algorithm summarizes the conceptual algorithm. Initially, cutj contains only the topmost value for a categorical attribute Dj with a taxonomy tree, Supj contains all domain values of a categorical attribute Dj without a taxonomy tree, and Intj contains the full-range interval for a continuous attribute Dj. The valid beneficial refinements in h[cutj; [Supj; [Intji form the set of candidates. At each iteration, we find the candidate of the highest Score, denoted Best (Line 4), apply Best to T, update h[cutj; [Supj; [Intji (Line 5), and update Score and the validity of the candidates in h[cutj; [Supj; [Intji (Line 6). The algorithm terminates when there is no more candidate in h[cutj; [Supj; [Intji, in which case it returns the masked table together with the solution set h[cutj; [supj; [Intji. Our algorithm obtains the masked T by iteratively refining the table form the most masked state. An important property of TDR is that the anonymity requirement is antimonotone with respect to the TDR. If it s violated before a refinement, it remains violated after the refinement. This is because a refinement never equates distinct values; therefore it never increased the count of duplicates a(qid). Hence, the hierarchically state at the top is separated by a border above which lie all satisfying states and below which lie all violating states. The TDR finds a state on the border, and this state is maximally refined in that any further refinement of it would cross the border and violate the anonymity requirements. Note that there may be more than one maximally refined state on the border. Our algorithm finds the one based on the heuristic selection criterion of maximizing Score at each step. Samarati[2] presents some results related to ntimonotonicity, but the results are based on a different masking model that generalizes all values in an attribute to the same level and suppresses data at the record level. Theorem 1. Algorithm 1 finds a maximally refined table that satisfied the given anonymity requirement. Algorithm 1 makes no claim on efficiency. In fact, in a straightforward implementation, Lines 4,5 and 6 require scanning all data records and recomputing Score for all candidates in h[cutj; [Supj; [Intji. Obviously, this is not scalable. The key to efficiency of our algorithms is directly accessing the data records to be refined and updating Score based on some statistics maintained for candidate in h[cutj; [Supj; [Intji. In the rest of the section, we explain a scalable implementation of Lines 4, 5, and 6.

4.2 Find the Best Refinement (Line 4)

This step makes use of computed InfoGain(x) and Ax(QIDi) for all candidates x in h[cutj; [Supji; [Intji and computed A(QIDi) for each QIDi. Before the first iteration, such information is computed in an initialization step for every topmost value, every suppressed value, and every full-range interval. For each subsequent iteration, such information comes from the update in the previous iteration (Line 6). Findings the best refinement Best involves at almost j [cutjj) j [supjj) j [Intjj computations of Score without accessing data records. Updating InfoGain(x) and Ax(QIDi) will be considered in section 4.4

4.3 Perform the Best Refinement(Line 5)

We consider two cases of performing the Best refinement, corresponding to whether a taxonomy tree is available for the attribute Dj for Best. Case 1: Dj has a taxonomy tree. Consider the refinement Best ! child(Best) where Best 2 Dj and Dj is either a categorical attribute with a specified taxonomy tree or a continuous attribute with a dynamically grown taxonomy tree. First, we replace Best with child(Best) in h[cutj; [Intji. Then, we need to retrieve RBest, the set of data records masked to Best, to tell the child value in child(Best) for each individual data records. We present a data structure Taxonomy Indexed PartitionS (TIPS) to facilitate this operations. This data structure is also crucial for updating InfoGain(x) and Ax(QIDi) for candidate x. the general idea is to group data records according to their masked records on [QIDi. Definitions 3 (TIPS). TIPS is a tree structure with each node representing a masked record over [QIDi and each child node representing a refinement of the parent node on exactly one attribute. Stored with each leaf node is the set of (compresses) data record having the same masked record, called a leaf partition. For each candidate refinement x, Px denotes a leaf partition whose masked record contains x, and Linkx denotes the link of all such Px. The head of Linx is stored with x. the masked table is represented by the leaf partitions of TIPS. Linkx provides a direct access to Rx, the set of (original) data records masked by the value x. initially, TIPS has only one leaf partition containing all data records, masked by the topmost value or interval on every attribute in [QIDi. In each iteration, we perform the best refinement Best by refining the leaf partition on LinkBest. Refine Best in TIPS. We refine each leaf partition PBest found on LinkBest as follows:

For each value c in child(Best), a child portion Pc is created under PBest and data record in PBest are split among the child partitions. Pc contains the data records in PBest if a categorical value c generalized the corresponding domain value in the record or if an interval c contains the corresponding domain value in the record, an empty Pc is removed. Linkc is created to link up all Pcs for same c. Also, link Pc to every Linkx to which PBest was previously linked, except for LinkBestFinally, mark c as “beneficial” if Rc has more than one class, where Rc denotes the set of data records masked to c. This is the only operation that actually accesses data records in the whole algorithm. The overhead is maintaining Linkx. For each attribute in [QIDi and each leaf partition on LinkBest, there are at most jchild(Best)j “relinking.” Therefore, there are at most j QIDjj _ jLinkBestj _ jchild(Best)j”relinkings” for applying Best.

A TIPS has several useful properties:

1) All data records in the same leaf partition have the same masked record, although they may have different refined values.

2) Every data record appears in exactly one leaf partition.

3) Each leaf partition Px has exactly one masked qidj on QIDj and contributes the count jPxj towards a(qidj). Later, we use the last property to extract a(qidj) from TIPS.

4.4 Update Score and Validity (Line 6)

This step updates Score(x) and validity for candidates x in h[cutj; [Supj; [Intji to reflect the impact of the Best refinement. The key is computing Score(x) from the count statistics maintained in Section 4.3 without accessing data records. We update InfoGain(x) and Ax(QIDi) separately. Note that the updated A(IDi) is obtained from ABest(QIDi).

4.4.1 Update InfoGain(x)

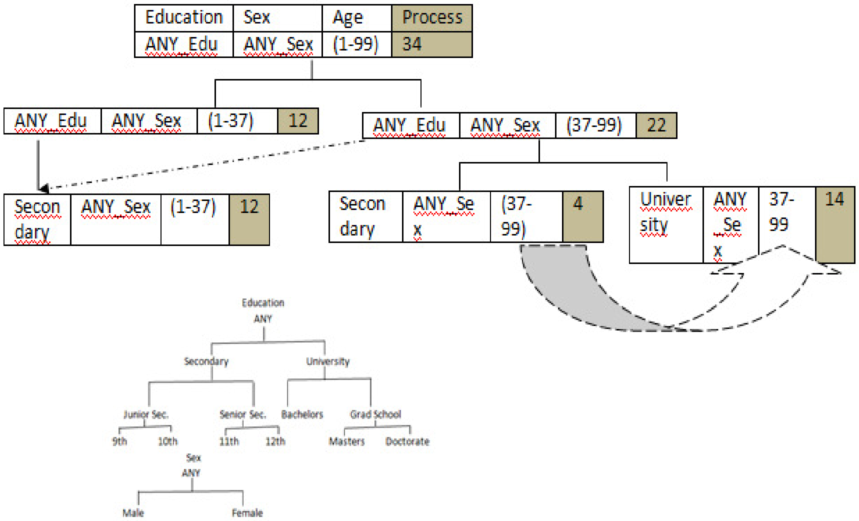
An observation is that InfoGain(x) is not affected by Best ! child(Best), except that we need to compute InfoGain(c) for each newly added value c in child(Best). InfoGain(c) can be computed while collecting the count statistics for c in Case 1 of section 4.3. in case the refined attribute has not taxonomy tree, InfoGain(x) can be computed from the count statistics for x in Case 2 of Section 4.3.

4.4.2 Update AnonyLoss(x)

Again, we consider the two cases:

Case1: Dj has a taxonomy tree. Unlike information gain, it is not enough to compute Ac(QIDi) only for the new values c in child(Best). Recall that Ax(QIDi) is equal to the minimum a(qidi) after refining x. if both att(x) and att(Best) are contained in QIDi, the refinement on Best may affect this minimum hence, Ax(QIDi).

Consider VID_1 = {Education, Sex}, VID_2 ={Sex, Age}



The above Fig presents the TIPS data structure presents the data structure Quasi-Identifier TreeS (QITS) to extract a(qidi) efficiently from TIPS for updating Ax(QIDi). Definition 4 (QITS) QITi for DIDi ¼ fD1;...;Dwg is a tree of w levels. The level p > 0 represents the masked values for Dp. Each root-to-leaf path represents an existing qidi on DIDi in the masked data, with a(qidi) stored at the leaf node. A branch is trimmed if it's a(qidi)$^{1/4}$ 0. A(QIDi) is equal to the minimum a(qidi) in QITi. In other words, QITi provides an index of a(qidi) by qidi. Unlike TIPS, QITS does not maintain data records. On applying Best ! child(Best), we update every QITi such that QIDi contains the attribute att(Best). Update QITi, for each occurrence of Best in QITi, create a separate branch for each c in child(Best). The procedure in algorithm 2 computes a(qidi) for the newly created qidis on such branches. The general idea is to loop through each Pc on Linkc in TIPS, increment a(qidi) by jPcj. This step does not access data records because jPcj was part of the count statistics of Best. Let r be the number of QIDi containing att(Best). The number of a(qidi) to be computed is at most r_jLinkBestj_jchild(Best)j.

5. SUMMARY

Our experiments verified several claims about the proposed TDR method. First, TDR masks a given table to satisfy a broad range of anonymity requirements without sacrificing significantly the usefulness to classification. Second, while producing a comparable accuracy, TDR is much more efficient than previously reported approaches, particularly, the genetic algorithm in [12]. Third, the previous optimal k-anonymization [7], [16] does not necessarily translate into the optimality of classification. The proposed TDR finds a better anonymization solution for classification. Fourth, the proposed TDR scales well with large data sets and complex anonymity requirements. These performances together with the features discussed in Section 1 make TDR a practical technique for privacy protection while sharing information.

6. CONCLUSION

We considered the problem of ensuring an individual's anonymity while releasing person-specific data for classification analysis. We pointed out that the previous optimal k-anonymization based on a closed-form cost metric does not address the classification requirement. Our approach is based on two observations specific to classification: Information specific to individuals tends to be over fitting, thus of little utility, to classification; even if a masking operation eliminates some useful classification structures, alternative structures in the data emerge to help. Therefore, not all data items are equally useful for classification and less useful data items provide the room for anonymizing the data without compromising the utility. With these observations, we presented a top-down approach to iteratively refine the data from a general state into a special state, guided by maximizing the trade-off between information and anonymity. This top-down approach serves a natural and efficient structure for handling categorical and continuous attributes and multiple anonymity requirements. Experiments showed that our approach effectively preserves both information utility and individual's privacy and scales well for large data sets.

REFERENCES

[1] P. Samarati and L.Sweeney, "Generalizing Data to provide Anonymity when Disclosing Information," Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems (PODS '98), p. 188, June 1998.

[2] P.Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge Eng., vol. 13, no.6, pp.1010-1027, Nov/Dec. 2001

[3] The House of Commons in Canada, "The Personal Information Protection and Electronic Documents Act," 1991, http://www.privcom.gc.ca/.

[4] S.M. Weiss and C.A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Machine Learning, and Expert Systems. Morgan Kaufmann, 1991.

[5] T.Dalenius, "Finding a Needle in a Haystack or Identifying Anonymous Census Record," J. Official Statistics, vol.2, no.3, pp.329-336, 1986.

[6] L.Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol.10, no.5 pp.571-588, 2002.

[7] R.J. Bayardo and R.Agrawal, "Data Privacy through Optimal k-Anonymization," Proc.21st Int'l Conf. Data Eng. (ICDE '05), pp.217-228. April 2005.

[8] G.Aggarwal, T.Feder, K.Kenthapadi, R. Motwani, R.Pamigraphy, D. Thomas, and A. Zhu, "Approximation Algorithms for k-Anonymity," J.Privacy Technology, no. 2005``1000`, Nov. 2005.

[9] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," Proc. 23rd ACM Symp. Principles of database Systems (PODS'04), pp. 223-228, 2004.

[10] L. Sweeney, "Datafly: A System for providing anonymity in Medical Data," Proc. Int'l conf. Database Security, pp. 356-381, 1998.

[11] A. Hundepool and L. Willenborg," and Argus : Software for Statistical Disclosure Control," Proc. Third Int'l Seminar on Statistical Confidentiality, 1996.

[12] V.S. Iyengar, “Transforming Data to Satisfy Privacy Constraints,” Proc. Eighth ACM SIGKDD Int’l conf. Knowledge Discovery and Data Mining, pp.279-288, July 2002.

[13] K.Wang, P. Yu, and S. chakraborty, “Bottom-Up Generalization: A Data Mining Solution to Privacy Protection,” Proc. Fourth IEEE Int’l conf. Data Mining (ICDM ’04), Nov. 2004.

[14] B.C.M. Fung, K.Wang, and P.S. Yu, “Top-Down Specialization for Information and Privacy Preservation,” Proc. 21st Int’l Conf. Data Eng. (ICDE ’05), pp.205-216, April 2005.

PREDICTING POPULARITY OF KOREAN CONTENTS IN ARAB COUNTRIES USING A DATA MINING TECHNIQUE

Park Young Eun[1], Soumaya Chaffar[2], Kim Myoung Sook[3] and
Ko Hye Young[4]

[1]College of Business Administration, Prince Sultan University
[2]Department of Computer Science,
Prince Sultan University, Riyadh, Saudi Arabia
[3]Department of Business Administration,
Future Convergence Industry College, Seoul Women's University
[4]Department of Digital Media, Future Convergence Industry College,
Seoul Women's University

ABSTRACT

Recently, many people in the Middle East and North Africa enjoy watching a variety of Korean contents such as Korean dramas, films, broadcasting programs and listening to Korean Pops. The Korean wave refers to the phenomenon of Korean entertainment and popular culture rolling over the world with TV dramas, films and pop music. Also it is known as "Hallyu " literally meaning 'flow from Korea' in Korean. This study examines the analysis of pattern on Arab countries (Middle East and North Africa) Consumers' consumption of the Korean Contents using social media, Facebook data. Then we focus on developing Predictive System using a Data Mining Technique.

KEYWORDS

Korean Wave (Hallyu in Korean), Social Media, Data Mining, Predictive Analysis

1. INTRODUCTION

The Korean Wave (K-wave), or Hallyu literally meaning 'flow from Korea' in Korean, referred as "the growing popularity of Korean pop culture, such as TV dramas, films, pop music, fashion, beauty, and online games being widely embraced and shared among the people of Asian countries in the late 1990s. In addition, with the rapid spread of social media like Facebook, YouTube and Twitter, K-wave has expanded its fandom outside of Asia to the West. That is, currently, this Korean wave has become the phenomenon of Korean popular culture rolling over the world not only Asian countries but also North and Latin America, Europe, even Middle Eastern countries and North Africa with Korean entertainment contents. The world-wide success of Korean pop culture contributed to improve the 'Korea' image and make a positive impact on Korean economy (Ahn et al., 2013).

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 59– 68, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70206

In the late 1990s, a few Korean TV dramas (hereafter, K-dramas), such as *What is Love All About?* (1997) and *Stars in My Heart* (1997), became popular in East and Southeast Asia and provided a wide range of Asian audiences with glimpses of Korean pop culture. The initial Korean wave was followed by the megahits of three K-dramas—*Autumn Fairy Tale* (2000), *Winter Sonata* (2002), and *Dae Jang Geum* (2003)—in Japan, Thailand, Singapore, and Hong Kong between 2002 and 2006. In the early 2000s, Hallyu was also led by the success of K-pop artists, such as BoA, Big Bang, and Dong Bang Shin Ki(TVXQ), in several Asian countries. In recent years, K-pop fandom has been evident even outside of Asia (Hong Mercier, 2013; Lansky, 2012). Finally, Psy and his song *Gangnam Style* were not only a world-wide phenomenon, but also a great significant turning point in the history of K-Pop and Korean pop culture. This is explicit just from Psy's achievements in the year 2012 alone. *Gangnam Style* has reached up to 1.7 billion views on YouTube, becoming the most-liked video in the site's history and topped the charts in over forty-one countries. What is more, the mere fact that small towns in South America, Southeast Asia, and the Middle East know the "horse dance" of *Gangnam Style*is an unexplainable achievement—not done justice by words (Park, 2015).The global dissemination of Korean popular culture such as *Gangnam Style* would not have been possible without global social media or social network service (SNS) sites (Park, 2013).

However, while it has been more than 15 years since the Korean pop culture phenomenon has emerged, academic analyses have not sufficiently addressed its consumption of Middle Eastern and North African area from a global perspective. The existing literatures on the Korean Wave focus their attention on the Asian market and tend to still define it primarily as an intra-Asian flow of particular forms of content without sufficiently addressing its dimension of social media and its technology and effect on Korean Wave from a global perspective (Jin and Yoon, 2016). Thus, it look over how a wide range of Western, Middle Eastern and African fans of Korean pop culture engage with social media and are networked with other fans (Jin and Yoon, 2016).

In addition, there has no analytical research and forecasting system to find the key factors affecting consumers' demand in these regions. In this regard, drawing on consumption pattern with these regions' fans of the recent Korean wave, this study explores how the Hallyu phenomenon is integrated into Middle East and North Africa through a social media. Finally, this paper aims to develop the pattern modeling of consumption behavior, systemically, in Middle East and North Africa using Data Mining techniques to build Korean contents management system and enhance marketing performance expectations by applying the results to the company's prediction system.

2. Related Work

Data mining is an essential tool in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Predicting is one of the most interesting and challenging tasks where to develop data mining applications. The use of computers with automated tools, large volumes of data are being collected and made available to the research groups (Shweta, 2012). As a result, many researches were carried out on various datasets using the data mining techniques to enhance forecasting in the business and medical fields.

Shweta (2012) has discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. The most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy (Shweta, 2012). As a result, data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets (Delen et al., 2005; Sarvestan et al., 2010 ; Shweta, 2012).

In addition, Der-Chiang Li et all.(2012) thought the overall electricity consumption, treated as a primary guideline for electricity system planning, is a major measurement to indicate the degree of a nation's development and The electricity consumption forecast is especially important with regard to policy making in developing countries. However, it is difficult to obtain accurate predictions using long-term data, and thus forecasting with limited (short-term) data is more effective and of considerable interest (Der-Chiang Li et all., 2012). Khana et al. (2013) also approached with similar idea. They explore three different data mining techniques for detecting abnormal lighting energy consumption using hourly recorded energy consumption and peak demand (maximum power) data. Two outliers' detection methods are applied to each class and cluster for detecting abnormal consumption in the same data set. In each class and cluster with anomalous consumption the amount of variation from normal is determined using modified standard scores (Khana, I., Capozzolia,A.,Corgnatia, S. P. and Cerquitellib,T.,2013).This study gives valuable implications for building energy management systems to reduce operating cost and time by not having to detect faults manually or diagnose false warnings. It means that will be useful for developing fault detection and diagnosis model for the whole building energy consumption. Besides, Fan et al.(2014) present data mining based approach to developing ensemble models for predicting next-day energy consumption and peak power demand, with the aim of improving the prediction accuracy(Fan, Xiao and Wang, 2014). Their ensemble model is developed and the weights of the eight predictive models are optimized using genetic algorithm. This approach shows that we can analyze the large consumption data and based on this analysis, we can apply the predictive models to know the consumption and peak demand for next season and develop strategies of fault detection, diagnosis, operation optimization and interactions between consumers and providers in the business field.

Predictive analytics has been useful in predicting the frequency of trading and the stock price for the next day, based on data from social media such as Twitter. This model was developed by Riverside and other researchers. A trading strategy supported by this model was created by three researcher, Ruiz et al.(2012). Their trading strategy performed 1.4%-11% better and other baseline strategies. In addition, during a 4-month simulation, this strategy outperformed the Dow Jones Industrial Average. Hristidis's study looked further than the effect of negative and positive sentiment on stock prices, it looked at the quantity of tweets and the interrelationship between tweets, users and topics.

Recently, O'Connor (2015) has found the relationship between Facebook popularity and consumer brand stock prices. The premise of his study aimed at finding out whether popularity (measured in Facebook likes) affected performance (share prices). He identified 30 brands that had the maximum number of followers, and tracked the likes these companies got for a period of one year as well as their share price on a daily basis. He found that 99.95% of the changes in share prices could be explained by the change in the number of fans. He did not see a direct relationship between likes and an increase or decrease in share prices, however, stock market trends were affected by the appreciation a company got on social media. The majority of a change in a particular company's stock price was linked with the likes that brand received for that period or day. Finally, the data of social media sentiment has become most applicable to produce other predictive models.

Jin and Yoon(2014) examine how Hallyu fans engage with a social media-saturated environment, drawing on qualitative interviews with North American fans of Korean pop culture. In comparison to the existing cultural analyses of the Korean wave, which focus, at best, on the content of particular genres or texts and their consumption, they map out transnational pop cultural flows with reference to the media environment through which the participatory culture of media users is spread (Jin& Yoon, 2014).

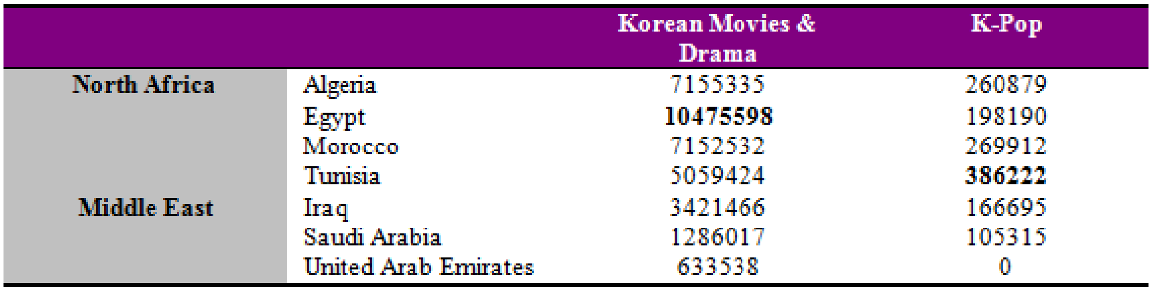
However, most studies tend to focus on either the role of digital technology of social media or predictive analytics using data mining techniques without fully consideration their conjunction between social media and data mining technique. Based on these existing studies and our awareness of these limitations, we approach to explore predictive models with social media using data mining technique for application of this concept into the consumption of Korean pop cultures. From this study, we, therefore, engage with the notion of spreadable social media in disseminating Korean pop cultures, and developing predictive models of K-contents' consumption and demand for the future, especially in the Middle East and North Africa.

3. METHODOLOGY

3.1. Data Collection

In order to understand popularity growth of Korean contents in Arabic countries, we collected data from two popular Facebook pages: the first one about 'Korean movies and drama' and the second one about 'K-pop'. Numbers of likes recorded on a daily basis by different countries over a period of two years (October 10th 2014 to October 10th 2016) have been collected. Then, we selected only those of Middle East and North Africa. Four North African (Algeria, Egypt, Morocco and Tunisia) and three Middle Eastern countries (Iraq, Saudi Arabia and United Arab Emirates) liked the Korean Movies and Drama. These countries except the United Arab Emirates also liked the K-pop page (see table 1).

Table 1. Number of likes of Korean pages by North African and Middle Eastern Countries over a Period of 2 Years



		Korean Movies & Drama	K-Pop
North Africa	Algeria	7155335	260879
	Egypt	**10475598**	198190
	Morocco	7152532	269912
	Tunisia	5059424	**386222**
Middle East	Iraq	3421466	166695
	Saudi Arabia	1286017	105315
	United Arab Emirates	633538	0

During the last two years and compared to other Arab countries (Algeria, Egypt, Morocco, Iraq and Saudi Arabia), Tunisia has the smallest population (around 11.317 million people), nevertheless it has the largest number of fans of K-pop (about386.222 likes). On the other hand, Egypt (about 10.475.598likes) has the largest numbers of likes of Korean movies and drama compared to other Arab countries.

3.2. Predictive Data Mining

Aiming to analyze the popularity of Korean contents in Middle Eastern and North African countries, we adopted a data-driven approach based on Data Mining techniques. The collected data as described above represent a time series composed of a chronological sequence of observations on number of likes by country associated to public Korean pages (movies/drama and k-pop). Predict future likes trends of Korean contents for Middle Eastern and North African countries can look ahead to future consumption behavior in order to maximize the success and profitability of Korean contents in these regions. This task falls under time series forecasting which is performed with different techniques including statistical and machine learning ones. The latest are often more powerful than the classical statistical techniques such as ARMA and ARIMA(Saigal and Mehrotra, 2012).Different Machine Learning algorithms have been used for time series forecasting such as Linear Regression, Robust Regression, Gaussian Processes, Support Vector Machine (SVM), etc. SVM algorithms have been used with a considerable success and often outperformed other methods (Ristanoski, Liu and Bailey, 2013). In this research work we will use SVM for regression (namely SMOreg in Weka Software) in order to predict future likes trends of Korean contents for different Middle Eastern (Iraq, Saudi Arabia and United Arab Emirates) and North African countries (Algeria, Egypt, Morocco and Tunisia). We will use the Mean Absolute Error in order to evaluate our model. The Mean Absolute Error (MAE) is used to measure how close predictions are to the eventual outcomes. MAE is given by the following equation:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|$$

where f_i is the predicted value and y_i is the true value.

4. RESULTS & DISCUSSION

In order to forecast the future trend for the number of likes of Korean music, movies and dramas, we employed WEKA forecasting plugin (Saigal and Mehrotra, 2012), which is a time series analysis and forecasting model. The target variable to be predicted is the number of likes that will be collected in 1 day until 10 days into the future.

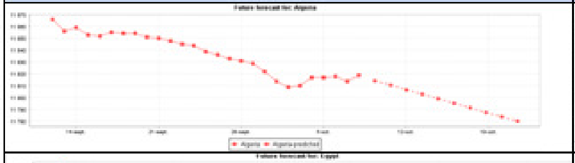
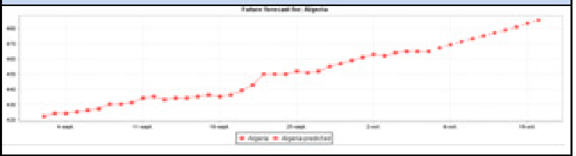
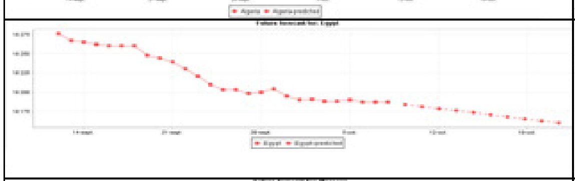
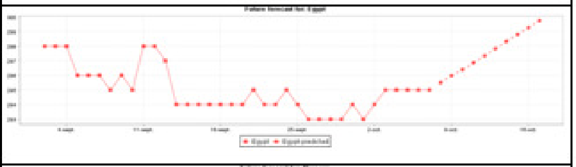
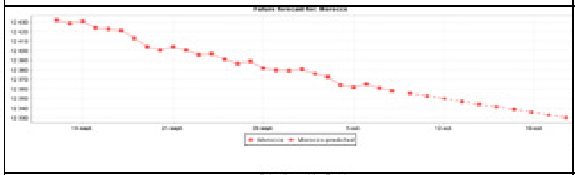
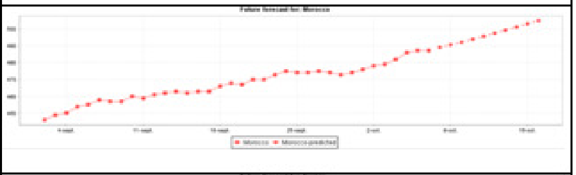
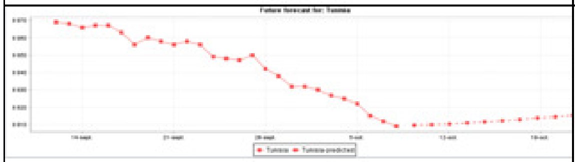
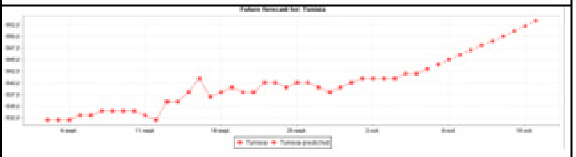
The collected data was divided into training data (used to estimate the model) and test data (used to evaluate the forecasts).The size of the test data set depends on the size of the whole data and number of time units to forecast and should ideally be at least as large as the maximum forecast horizon required (Hyndman and Khandakar, 2008). In this research 5% of the total sample used to evaluate the forecasts for a period of 10 days. The performance results of the algorithms are based on the Mean Absolute Error (MAE) and are reported in Table 2.

Table 2. Evaluation of SMOreg for predicting the number of likes of Korean pages within a period of 10 days using MAE

MAE	Korean Movies and Drama										K-pop									
	1 day	2 days	3 days	4 days	5 days	6 days	7 days	8 days	9 days	10 days	1 day	2 days	3 days	4 days	5 days	6 days	7 days	8 days	9 days	10 days
Algeria	3.65	5.87	8.11	10.19	12.02	13.68	15.13	16.58	18.03	19.97	1.29	2.05	2.51	3.01	3.28	3.42	3.64	3.8	3.95	4.06
Egypt	3.82	6.03	8.03	9.84	11.43	12.73	15.38	17.94	20.48	23.24	0.82	1.25	1.68	2.04	2.36	3.24	2.76	3.66	4.14	4.59
Morocco	2.76	3.75	3.82	4.21	4.35	4.53	5.35	5.77	5.19	5.78	1.34	1.74	2.11	2.7	2.95	3.45	3.65	3.95	4.46	4.84
Tunisia	3.35	5.61	8.11	10.7	13.37	16.16	18.88	21.40	24.37	27.15	1.12	1.61	2.08	2.58	2.94	3.34	3.83	4.24	4.77	5.39
Iraq	1.98	2.56	3.74	4.86	6.21	7.57	9.22	10.79	12.51	14	0.66	0.97	1.18	1.34	1.61	1.83	2.12	2.4	2.71	3.09
Saudi Arabia	2.89	4.80	6.52	7.94	9.06	10.91	12.95	15.08	16.85	18.78	0.54	0.82	1.13	1.37	1.6	1.72	1.85	1.9	2.05	2.03
UAE	2.16	4.06	5.96	7.75	9.49	11.44	13.23	14.86	16.47	17.88										

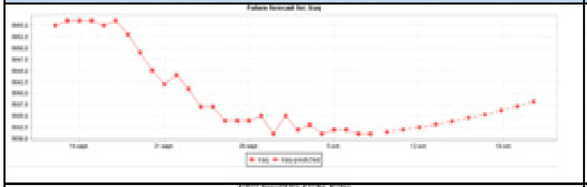
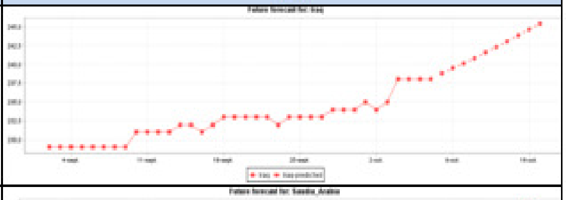
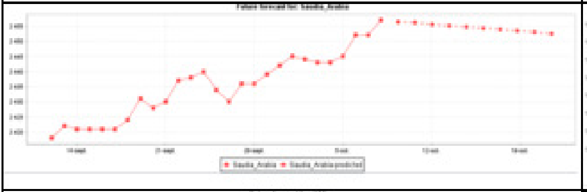
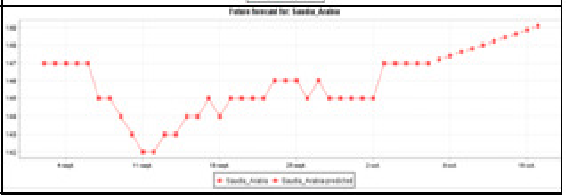
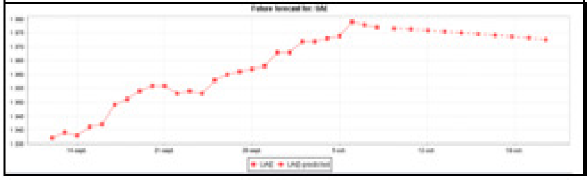
Popularity trends of Korean Movies and K-pop in the North African and Middle Eastern countries are respectively presented in table 3 and 4.On the Y axis are the number of likes and on the X axis are the weekly dates in the last month.

Table 3. Actual and predicted values of likes by North African Countries for Korean Movies and K-pop using SMOreg

Korean Movies and Drama	K-pop
	
	
	
	

For the next ten days, it is obvious that the number of likes for K-pop will increase for all North African countries, however concerning Korean Movies and Drama except Tunisia it is decreasing for Algeria, Egypt and Morocco. Popularity trends of Korean Movies and K-pop in the Middle Eastern countries are presented in the table 4 below

Table 1. Actual and predicted values of likes by Middle Eastern Countries for Korean Movies and K-pop using SMOreg

Korean Movies and Drama	K-pop
	
	
	

Similarly to North African, Middle Eastern countries recognize an increase of the number of likes for K-pop. However, concerning Saudi Arabia and United Arab Emirates, the number of likes will decrease for Korean Movies and Drama which is not the case for Iraq.

From the results, we found some interesting findings as follows. First of all, the trend for the number of likes of Korean music, movies and drama shows differently between genres. The number of likes for Korean movies and dramas is decreasing while the number of likes for K-pop is increasing. This result can be explained with time difference. In the beginning, people in Arab countries started to enjoy Korean contents with drama and movie. And then over time they also enjoyed Korean drama's songs (OST : Open Source Track) after they got used to Korean contents by dramas. Second, from the actual and predictive trend, we found that the trend and also speed for the number of likes of Korean contents is totally decreasing while Korean contents became more popular in Arab countries. This result can be explained with the change in consumer behavior. At the beginning with the emergence of Social Media, users were interested to share their experiences about Korean drama and movie through Facebook, but more and more after many websites and applications (such as 'www.myasiantv.se'; www.viki.com ; www.kissasian.com ; www.baykorean.net ; www.dramafever.com; www.dramayou.com ; www.dardarkom.com ; http://kshowonline.com ; http://aradrama.com) appeared to make people enjoy Korean contents directly, users access several websites related to Korean contents straightly without using social networks like Facebook. Lastly, as we noticed in the results, the number of likes of Korean products by Gulf countries are very small comparing to other countries, this is because we believe that Twitter is more popular than Facebook in these countries. In addition, some countries started to broadcast Korean drama through main public channels, for example MBC, main channel of Saudi Arabia spread Korean dramas with Arabic subtitles. This also affects the future trend of Korean contents in Arab countries.

5. CONCLUSION & FUTURE WORK

Based on our analysis and results on predicting popularity of Korean contents in Arab countries, we can find some valuable implications as follows. First, K-contents such as drama, movie and music are sometimes a gateway to a wider interest in Korean culture, food and brands. Korean brands can be inserted into dramas and music videos. South Korea's government long ago embraced pop cultures as a way to transform itself into global market's trendsetter and fuel its economy. Product placement is huge in K-drama. Korean companies' products such as Samsung phones and Hyundai cars make frequent appearances. In 2016, market observers who forecast that one Korean drama "*Descendants of the Sun*" alone will boost the Korean economy by $261 million, partly by driving demand for tourism and products. Considering this huge effect, prediction on K-contents' consumption in the attractive, emerging markets such as Middle east and North Africa will be various. Increasing the awareness of Korean brand by promoting Korean cultural content will remove entry barrier and create some opportunities for Korean companies who seek to operate a business in Arab countries. Second, Korea entertainment companies can take into consideration the cultural characteristics of the Middle East. There was a long queue to taste Halal certified traditional Korean food, such as bibimbap, and bulgogi and beauty or fashion tips for Arabic women. Considering and depicting these unique points, contents companies can produce distinctive cultural contents to integrate with Arabic culture. Third, this study gives valuable implications for building Korean contents management systems to reduce operating cost and time by not having to detect faults manually or diagnose false warnings. It will be useful for developing predictive model for the whole building Korean contents consumption and also for overcoming liability of foreignness in the global marketplace.

Even though our findings were significant, this study has some limitations. Such limitations and the future direction of the research are as follows. First, we need to study many other features such as demography (age, sex, gender, etc.) comments in order to make more individualized, customized , localized marketing strategies. Second we need to collect comments entered by users and apply sentiment analysis techniques (Chaffar & Inkpen, 2016) in order to analyze their satisfaction about some features of Korean products. Companies, through a business intelligence process, aim to analyze customers' feelings about the products, services, agents and organization. This can lead to the development of new strategies for customers' satisfaction and can provide the company with a competitive advantage in the market. Technologies that automatically recognize unhappy customers can be extremely useful to companies. Third, in this research only Facebook is used to collect the data, we need also to use another social network and specifically Twitter. Facebook is more popular in North African countries however Twitter is more popular in the gulf region. Given this market difference, this study cannot describe the general market situation, and this should be considered in future studies.

ACKNOWLEDGEMENTS

This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) (R7719-16-1002) supervised by the IITP(Institute for Information & communications Technology Promotion).

REFERENCES

[1] Ahn, J., Oh, S. and Kim, H. (2013). Korean pop takes off! Social media strategy of Korean entertainment industry. In: Proceedings on the 10th international conference on service systems and service management, Hong Kong, 17–19 July, pp. 774–777. New York: IEEE.

[2] Chaffar, S., Inkpen, D. (2016). Using a Generic Text-based Approach for Emotion Prediction. Accepted in International Conference on Computer and Applications, Dubai, UAE.

[3] Dursun, D. Glenn, W. and Amit, K. (2005) "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127.

[4] Der-Chiang, L., Che-Jung, C., Chien-Chih, C., Wen-Chih, C. (2012). Forecasting short-term electricity consumption using the adaptive grey-based approach—An Asian case, Special Issue on Forecasting in Management Science, Volume 40, Issue 6, pp. 767–773.

[5] Fan, C., Xiao, F. and Wang S. (2014), Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy, 127 (2014), 1-10.

[6] Ruiz, E.J., Hristidis, V, Castillo, C., Gionis, A., Jaimes, A. (2012)Correlating Financial Time Series with Micro-Blogging Activity. ACM International Conference on Web Search and Data Mining (WSDM).

[7] Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting : the forecast package for R. Journal of Statistical Software 26(3), 1–22.

[8] Jin, D.Y. and Yoon, K. (2016), The social mediascape of transnational Korean pop culture : Hallyu 2.0 as spreadable media practice, New media & society, 18(7), 2016.

[9] Khana,I., Capozzolia,A.,Corgnatia,S. P., Cerquitellib,T. (2013),Fault Detection Analysis of Building Energy Consumption Using Data Mining Technique, Energy Procedia 42, pp. 557 – 566.

[10] Park, G.S. (2013), Manufacturing Creativity : Production, Performance, and Dissemination of K-pop, Korean Journal 53(4) : 14-33.

[11] Park, Y.E. (2015), YG Family, We are One and Number One : based on Eclectic Paradigm (OLI) by J. Dunning, Korean Academy of International Business.

[12] Saigal, S., Mehrotra, D., (2012). Performance comparison of time series data using predictive data mining techniques. Advances in Information Mining. 4 (1), pp. 57–66.

[13] Ristanoski, G., Liu, W. and Bailey, J. (2013) A time-dependent enhanced support vector machine for time series regression, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 946-954.

[14] Saigal, S., Mehrotra, D., (2012). Performance comparison of time series data using predictive data mining techniques. Advances in Information Mining. 4 (1), pp. 57–66.

[15] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M. (2010) Predicting Breast Cancer Survivability using data mining technique, Software Technology and Engineering (ICSTE), 2nd International Conference, vol.2, pp.227-231.

[16] Shweta, K. (2012). Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, No. 2.

FAST ALGORITHMS FOR UNSUPERVISED LEARNING IN LARGE DATA SETS

Syed Quddus

Faculty of Science and Technology,
Federation University Australia, Victoria, Australia.

ABSTRACT

The ability to mine and extract useful information automatically, from large datasets, is a common concern for organizations (having large datasets), over the last few decades. Over the internet, data is vastly increasing gradually and consequently the capacity to collect and store very large data is significantly increasing.

Existing clustering algorithms are not always efficient and accurate in solving clustering problems for large datasets.

However, the development of accurate and fast data classification algorithms for very large scale datasets is still a challenge. In this paper, various algorithms and techniques especially, approach using non-smooth optimization formulation of the clustering problem, are proposed for solving the minimum sum-of-squares clustering problems in very large datasets. This research also develops accurate and real time L2-DC algorithm based with the incremental approach to solve the minimum sum-of-squared clustering problems in very large datasets, in a reasonable time.

GENERAL TERMS **:** Data Mining.

KEYWORDS **:**

Clustering analysis, k-means algorithm, Squared-error criterion, Large-data sets.

1. INTRODUCTION

Data classification by unsupervised techniques is a fundamental form of data analysis which is being used in all aspects of life, ranging from Astronomy to Zoology. There have been a rapid and massive increase in amount of data accumulated in recent years, due to this, the use of clustering has also expanded further, in its applications such as personalization and targeted advertising. Clustering is now a key component of interactive- systems which gather information on millions of users on everyday basis [1-10, 20]. A process of dividing, classifying or grouping a dataset into meaningful similar partitions or subclasses based on some criteria, normally a distance function between objects, called clusters.

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 69– 76, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70207

Existing clustering algorithms are not always efficient & accurate in solving clustering problems for large datasets. The accurate and real time clustering is essential and important for making informed policy, planning and management decisions. Recent developments in computer hardware allows to store in RAM and repeatedly read data sets with hundreds of thousands and even millions of data points. However, existing clustering algorithms require much larger computational- time and fail to produce an accurate solution [16,17,18,19].

In this paper, we present an overview of various algorithms and approaches which are recently being used for Clustering of large data and E-document. In this paper we will discuss widely used evolutionary techniques and present results of DC-based clustering methodology in very large & big datasets.

1.1 Heuristics Approaches:

The k-means clustering technique and its modifications are representatives of such heuristics approaches. The global k-means and modified global k-means are representatives of incremental based heuristic algorithms. The within-cluster point scatter should be symmetric and it should attain its minimum value. The distance measure within cluster scatter is known as metric, we can measure this distance by the different methods such as Minkowski and Euclidean distance measure [6, 8, 9].

1.1.1 Minkowski Distance Measure

The distance between two data instances can be calculated using the Minkowski Metric as below[22]:

$D(x, y) = (|x_{i1} - x_{ji}|)g + |x_{i2} - x_{j1}|g + \ldots\ldots\ldots + |x_{in} - x_{jn}|g)1/g$

1.1.2 Euclidean Distance Measure

It is the most commonly used method to measure the distance between two objects when g = 2.when g = 1, the sum of absolute paraxial distance is obtained and when g = Infinity one gets the greatest of the paraxial distance. If the variable is assigned with a weight according to its importance then weighted distance should be measure [22].

1.1.3 Parallel k-means

The concept is to distribute processing of k-means on k machines which result in a satisfactory time complexity. Due to memory limitation it may not be efficient for massive data set.

1.1.4 Partial/Merge k-Means

Partial/merge k-means re-runs the k-means several times to get better result in each partition. However this algorithm is sensitive to the size of partitioning in massive data sets.

Different heuristics have been developed to tackle clustering problems. These heuristics include k-means algorithms and their variations such as h-means and j-means. However, these algorithms are very sensitive to the choice of initial solutions, they can find only local solutions and such solutions in large data sets may significantly differ from global ones [2-9]. However, the success of local methods depends on starting points.

1.2 Heuristics Based on the Incremental Approach

These algorithms start with the computation of the centre of the whole data set A and attempt to optimally add one new cluster centre at each stage. In order to solve Problem (2.5) for k > 1 these algorithms start from an initial state with the k-1 centres for the (k-1)-clustering problem and the remaining k-th centre is placed in an appropriate position. The global k-means and modified global k-means, a single pass incremental clustering algorithm, CURE, DC-based algorithm with the incremental approach are representatives of these algorithms [4, 5, 6, 7].

Usually the massive data set cannot fit into the available main memory, therefore the entire data matrix is stored in a secondary memory and data items are transferred to the main memory one at a time for clustering. Only the cluster representations are stored in the main memory to alleviate the space limitations. DC based algorithm with the incremental approach, is used to solve optimization problems in these massive and very large data sets in a reasonable time [7,8,9].

1.3 Population based evolutionary algorithms

Population based evolutionary algorithms are suitable to generate starting cluster centres as They can generate points from the whole search space. By using before mentioned five Evolutionary algorithms (Genetic algorithm, Particle swarm optimization, Ant colony Optimization, Artificial bee colony and Cuckoo search) in combination with the incremental Algorithm, a new algorithms will be designed to generate starting cluster centres.

Over the last several years different incremental algorithms have been proposed to solve clustering problems. These algorithms attempt to optimally add one new cluster centre at each stage. In order to compute k-partition of a set these algorithms start from an initial state with the k-1 centres for the (k-1)-clustering problem and the remaining k-th centre is placed in an appropriate position. In this paper, our aim is to discuss various clustering techniques for very large datasets and to present how smooth optimization algorithms to solve clustering problems. We propose the L2-DC based algorithm which is based on the combination of smoothing techniques and the incremental approach. In order to find starting points for cluster centres we introduce and solve the auxiliary cluster problem which is non-smooth and non-convex. The hyperbolic smoothing technique is applied to approximate both the cluster and auxiliary cluster functions. Then we apply the Quasi-Newton method with the BFGS update to minimize them. We present results of numerical experiments on five real-world data sets [8, 9].

2. EXPERIMENTAL RESULTS

Algorithms were implemented in Fortran95and compiled using the gfortran compiler. Computational results were obtained on a Laptop with the Intel(R) Core(TM) i3-3110M CPU @ 2.4GHz and RAM 4 GB (Toshiba). Five real-life data sets have been used in numerical experiments [21].The brief description of these data sets is given below in table.1. All data sets contain only numeric features and they do not have missing values. To get as more comprehensive picture about the performance of the algorithms as possible the datasets were chosen so that:(i) the number of attributes is ranging from very few (3) to large (128); (ii) the number of data points is ranging from tens of thousands(smallest13,910) to hundreds of thousands (largest434,874). We computed upto24clusters in all data sets. The CPU time used by algorithms is limited to 20h. Since the L2-DC based algorithm compute clusters incrementally we present results with the maximum number of clusters obtained by an algorithm during this time.

2.1 Tables

Table: 1 the brief description of datasets.

N	**Data Sets**	**Number of instances**	**Number of attributes**
1	Gas Sensor Array Drift Dataset	13910	128
2	Bank Marketing	45211	17
3	Shuttle Landing Control	58000	10
4	Educational Process Mining (EPM): A Learning Analytics Data Set	230318	9
5	3D Road Network (North Jutland, Denmark)	434874	3

We run experiments on these real-life data sets to compute the Cluster function values obtained by algorithms, CPU time and the total number of distance function evaluations for all these five datasets. For numerical results: k - is the number of clusters;

f - is the optimal value of the clustering function obtained by the algorithm; N - is the total number of distance function evaluations; t- is the CPU time.

Table: 2. Results for data set 1

k	**f**	**N**	**t**
2	7.91E+13	6.42E+07	88.2969
4	4.16E+13	3.53E+08	444.0938
6	2.74E+13	7.09E+08	878.5781
8	2.03E+13	1.34E+09	1651.594
10	1.66E+13	1.79E+09	2254.563
12	1.41E+13	2.46E+09	3068.984
14	1.21E+13	3.26E+09	4108.953
16	1.06E+13	3.87E+09	4906.828
18	9.65E+12	4.62E+09	5848.375
20	8.85E+12	5.55E+09	7027.031
22	8.14E+12	6.21E+09	7862.984
24	7.55E+12	6.97E+09	8842.969

Table: 3. Results for data set 2

k	f	N	t
2	2.02E+11	2.96E+08	11.1385
4	7.33E+10	2.85E+09	116.4079
6	3.62E+10	5.86E+09	239.8671
8	2.41E+10	9.34E+09	379.2696
10	1.64E+10	1.45E+10	605.1747
12	1.23E+10	1.77E+10	736.5119
14	1.05E+10	2.07E+10	852.3739
16	8.48E+09	2.65E+10	1116.811
18	7.16E+09	3.38E+10	1449.125
20	6.43E+09	3.99E+10	1741.174
22	5.66E+09	5.17E+10	2349.64
24	5.13E+09	6.85E+10	3165.276

Table: 4. Results for data set 3

k	F	N	t
2	2.13E+09	59566001	8.3773
4	8.88E+08	5.45E+08	67.6108
6	5.67E+08	2.30E+09	291.9091
8	3.73E+08	4.20E+09	538.6871
10	2.85E+08	6.29E+09	808.4908
12	2.21E+08	1.23E+10	1648.26
14	1.78E+08	2.04E+10	2249.597
16	1.46E+08	2.59E+10	2573.205
18	1.20E+08	3.37E+10	3048.447
20	1.06E+08	3.77E+10	3333.289
22	95703872	4.65E+10	3849.216
24	84889772	5.58E+10	4687.924

Table: 5. Results for data set 4.

K	F	N	t
2	2.19E+19	1.83E+09	93.117
4	4.10E+17	8.34E+09	396.6793
6	1.54E+17	2.08E+10	999.28
8	8.41E+16	3.62E+10	1765.261
10	5.78E+16	4.99E+10	2399.046
12	3.79E+16	7.85E+10	3889.526
14	2.79E+16	2.96E+11	15253.17
16	2.09E+16	3.36E+11	17159.49
18	1.49E+16	4.01E+11	20600.2
20	1.09E+16	5.38E+11	28124.38
22	7.80E+15	5.92E+11	30885.5
24	6.40E+15	6.70E+11	34813.82

Table: 6. Results for data set 5.

k	f	N	t
2	4.91E+07	8.18E+10	1443.181
4	1.35E+07	2.74E+11	4860.913
6	6.38E+06	4.68E+11	8231.611
8	3.78E+06	6.63E+11	11673.59
10	2.57E+06	8.63E+11	15169.38
12	1.85E+06	1.07E+12	18800.85
14	1424129	1.29E+12	22818.89
16	1139559	1.5E+12	26609.68
18	948040.9	1.72E+12	30518.32
20	808708.8	1.94E+12	34452.17
22	703308.7	2.17E+12	38592.34
24	638434.2	2.41E+12	42971.4

The results of implementation of the L2-DC based algorithm are shown, respectively, in Table 2 for vowel dataset, Table 3. Table 4, Table 5, and Table 6 for five real time datasets.

All five data sets can be divided into two groups. The first group contains data sets with small number of attributes (3or 9). 3D-Road Network data set and Educational Process Mining (EPM): A Learning Analytics Data Set belongs to this group. The number of points in these datasets ranges from 2072862 to 1304622. Results presented in Tables5 and 6 demonstrate that in these datasets the performance of algorithm is similar in the sense of accuracy. All algorithms can find at least near best known solutions in these datasets.

The second group contains data sets with relatively large number of attributes. Gas Sensor Array

Drift, Shuttle Control data and Bank Marketing data sets belong to this group. The number of attributes in these data sets ranges from 10 to128. Results show that the algorithm is very efficient to find (near) best known solutions. The dependence of the number of distance function evaluations on the number of clusters in group1 of datasets is similar and the dependence of the number of distance function evaluations on the number of clusters in group2 of datasets is also similar.

The dependence of the CPU-time on the number of clusters for all datasets in group1 is similar. As the number of clusters increase, the dependence of CPU time monotonically increases. It is obvious that as the size (the number of data points) of a data set increase this algorithm requires more CPU time. The dependence of the CPU-time on the number of clusters for all datasets in group1 is similar in a sense, as the number of clusters increase, the dependence of CPU time monotonically increases. But the algorithm takes almost similar time pattern in clustering datasets: Shuttle Control data and Bank Marketing data sets but in case of Gas Sensor Array Drift dataset, the algorithm requires much more CPU time.

3. CONCLUSION

In this paper the minimum sum-of-squares clustering problems are studied using L2-DC based approach. An incremental algorithm based on DC representation is designed to solve the minimum sum-of-squares clustering problems. A special algorithm is designed to solve non-smooth optimization problems at each iteration of the incremental algorithm. It is proved that this algorithm converges to inf-stationary points of the clustering problems.

4. FUTURE WORK

As we know very large data set clustering is an emerging field. In this research, different evolutionary methods, their features and their applications have been discussed. We have implemented one of the techniques and we may implement more in future. The expectation is to implement the best technique which can efficiently solve the minimum sum-of-squares clustering problems and find the best solution in real time.

REFERENCES

[1] Yasin, H., JilaniT. A., and Danish, M. 2011. Hepatitis-C Classification using Data Mining Techniques. International Journal of Computer Applications.Vol 24– No.3.

[2] K.S. Al-Sultan, A tabu search approach to the clustering problem, {\em Pattern Recognition}, 28(9)(1995) 1443-1451.

[3] A.M. Bagirov, Modified global k-means algorithm for sum-of-squares clustering problems, {\em Pattern Recognition,} 41(10), 2008, 3192--3199.

[4] A.M. Bagirov, A.M. Rubinov, J. Yearwood, A global optimisation approach to classification, {\em Optimization and Engineering,} 3(2) (2002) 129-155.

[5] A.M. Bagirov, A.M. Rubinov, N.V. Soukhoroukova, J. Yearwood, Supervised and unsupervised data classification via nonsmooth and global optimization, {\em TOP: Spanish Operations Research Journal,} 11(1)(2003) 1-93.

[6] A.M. Bagirov and J. Ugon, An algorithm for minimizing clustering functions, \emph{Optimization,} 54(4-5), 2005, 351-368.

[7] A.M. Bagirov, J. Ugon and D. Webb, Fast modified global k-means algorithm for sum-of-squares clustering problems, {\em Pattern Recognition,} 44, 2011, 866--876.

[8] A.M. Bagirov, J. Yearwood, A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems, {\em European Journal of Operational Research,} 170(2006) 578-596.

[9] A.M. Bagirov, A. Al Nuaimat and N. Sultanova, Hyperbolic smoothing method for minimax problems, \emph{Optimization,} accepted.

[10] H.H. Bock, Clustering and neural networks, in: A. Rizzi, M. Vichi, H.H. Bock (eds), {\em Advances in Data Science and Classification}, Springer-Verlag, Berlin, 1998, pp. 265-277.

[11] D.E. Brown, C.L. Entail, A practical application of simulated annealing to the clustering problem, {\em Pattern Recognition}, 25(1992) 401-412.

[12] G. Diehr, Evaluation of a branch and bound algorithm for clustering, {\em SIAM J. Scientific and Statistical Computing}, 6(1985) 268-284.

[13] R. Dubes, A.K. Jain, Clustering techniques: the user's dilemma, {\em Pattern Recognition}, 8(1976) 247-260.

[14] P. Hanjoul, D. Peeters, A comparison of two dual-based procedures for solving the p-median problem, {\em European Journal of Operational Research,} 20(1985) 387-396.

[15] P. Hansen, B. Jaumard, Cluster analysis and mathematical programming, {\em Mathematical Programming,} 79(1-3)(1997) 191-215.

[16] A. Likas, M. Vlassis, J. Verbeek, The global k-means clustering algorithm, {\em Pattern Recognition}, 36(2003) 451-461.

[17] O. du Merle, P. Hansen, B. Jaumard, N. Mladenovic, An interior point method for minimum sum-of-squares clustering, {\em SIAM J. on Scientific Computing,} 21(2001) 1485-1505.

[18] H. Spath, {\em Cluster Analysis Algorithms}, Ellis Horwood Limited, Chichester, 1980.

[19] L.X. Sun, Y.L. Xie, X.H. Song, J.H. Wang, R.Q. Yu, Cluster analysis by simulated annealing, {\em Computers and Chemistry,} 18(1994) 103-108.

[20] A.E. Xavier, The hyperbolic smoothing clustering method, \emph{Pattern Recognition}, 43(3), 2010, 731-737.

[21] http://www.ics.uci.edu/mlearn/MLRepository.html, UCI repository of machine learning databases.

[22] Neha Khan, Mohd Shahid, Mohd Rizwan, ‘Big data classification using evolutionary techniques:A survey’,(2015), IEEE International Conference (ICETECH), India.

TEXT EXTRACTION FROM RASTER MAPS USING COLOR SPACE QUANTIZATION

Sanaz Hadipour Abkenar and Alireza Ahmadyfard

Department of Electronic and Robatic Engineering,
Shahrood University of Technology, Shahrood, Iran

ABSTRACT

Maps convey valuable information by relating names to their positions. In this paper we present a new method for text extraction from raster maps using color space quantization. Previously, most researches in this field were focused on Latin texts and the results for Persian or Arabic texts were poor. In our proposed method we use a Mean-Shift algorithm with proper parameter adjustment and consequently, we apply color transformation to make the maps ready for K-Means algorithm which quantizes the colors in maps to six levels. By comparing to a threshold the text layer candidates are then limited to three. The best layer can afterwards be chosen by user. This method is independent of font size, direction and the color of the text and can find both Latin and Persian/Arabic texts in maps. Experimental results show a significant improvement in Persian text extraction.

KEYWORDS

Color space conversion, K-Means clustering, Mean-Shift algorithm, Quantization, Text extraction.

1. INTRODUCTION

Images are one of the most important media for transferring data. An image can have much higher impression than hundred lines of documents. Hence, image understanding and data extraction from images could be used for other tasks such as Machine learning [1].

Many organizations routinely use large sets of hard-copy graphic documents, including maps, engineering drawings, electrical schematics, and technical illustrations. Recent advances in computer technology allow graphic information to be stored and accessed more conveniently and cost-effectively in electronic form than on paper [2].

Maps are easily accessible compared to other geospatial data, such as vector data, satellite imagers, gazetteers, etc. Due to the availability of high quality scanners and existence of Internet, we can now obtain various maps in raster format for areas around the globe. By converting the text labels in a raster maps to machine editable texts, we can produce geospatial knowledge for understanding the region on map while its other geospatial data are not available. Moreover, a raster map can be registered to other geospatial data (e.g., imagers) and recognized texts from the map can be exploited for indexing and retrieval of the other geospatial data [3].

Text labels in raster maps link place names to their geographic locations. Texts in maps contain very important information, as converting the text labels in a raster map to machine editable text, helps produce geospatial knowledge for understanding a map region [4].

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 77– 86, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70208

Finding a special place in a city map needs a strong map reader system. Therefore, improving existed map reader systems helps tourism industry. Today with the advent of auto-pilot cars, users prefer just to enter a name as an input to their cars GPS to be there. If we could extract texts from a map it could be useful in this case as well.

As stated earlier the most important use of text extraction is in GIS systems, but it could also be used in data mining, tourism and auto-pilot cars. Presenting a perfect map reader system which can improve precision (find the most texts with the least road sections) in minimum time and with minimum user interaction is the goal of almost all the researches in this field.

When we discuss about the streets of a city, color has an important role. In these kinds of raster maps, text, road lines, important building such as hospitals, schools, churches, mosques, etc. have been shown in different colors. Thus, color segmentation and quantization can help different layers separation and also text extraction.

In this paper we tried to improve the algorithm for extraction of Persian/Arabic and English text from geographical maps. Existence of points, subscripts, superscripts and some special parts of words which are in a lower or in a higher level from the words, discriminate Persian structure from English. Hence, existed methods on Latin texts are not applicable on Persian maps. Existing Persian methods work on grey scale maps. As most of the maps are colorful nowadays, they do not use color`s abilities. Moreover, their precision is low. In this paper we explain a method which can solve these issues. An algorithm is proposed in which we can extract text, specially Persian text, from colorful raster maps with the lowest error. We have tried to find the most possible words in the maps but least road lines and graphic symbols. This method is applicable for both English and Persian texts on the maps and is independent of font, size, direction and color of the texts.

The paper is organized as follows: In the next section we will review related works to text extraction. In the third section implementation of the algorithm will be described. Section four shows experimental results and last section shows conclusion of the proposed method.

2. RELATED WORKS

Fletcher and Kasturi described development and implementation of a new algorithm for automated text string separation which is relatively independent of changes in text font style and organizes individual characters. In their work, first connected components are produced. Then they use an area/ratio filter. Collinear component grouping and logical grouping of strings into words and phrases are respectively, their lateral steps in the proposed approach to separate text strings. The algorithm produces two images; One for texts and the other for graphics [5]. In their work Hough transform is used for character grouping and then text strings are extracted. As Hough transform only detect straight lines, their method cannot be applied to curved strings.

Chen and Wang presented a complete algorithm for extracting and recognizing numeral string on maps. Character extraction algorithm can segment slant and touching characters to their unique elements. Recognizing algorithm based on properties can also detect numeral characters with any size, position and direction. Discrimination property which is used here is simply detectable. In their proposed approach at first characters are extracted. After that recognition operation is applied. In this recognition holes, intended points, symmetric shapes and crossing points are recognized. Hough transform and a set of font and size dependent properties are also utilized for numeral strings detection [6]. However, this algorithm is not useful for alphabetic characters.

Velazquez and Levachkine proposed a method for separating and recognizing alphanumeric characters. In their method the map is segmented first, therefore all text strings, which contains touching symbols, strokes and characters, are extracted. Second, OCR-based recognition with artificial neural network (ANN) is applied to define coordinates, size and orientation of alphanumeric character strings in each case presented in map. Third, four straight lines or a number of curvatures which computed as a function of primarily recognized by ANN characters are extrapolated to separate those symbols that are attached. Finally, the separated characters are used as inputs into an ANN again to be finally identified [7].Velazquez and Levachkine's technique is presented for text detection in multi direction and curved strings. They divided their input documents into two equal columns. Each column is divided to some blocks based on connected components sizes to calculate linearity of local connected components and extract existing text strings.

In [8] Roy et. al. proposed a new approach for extracting unique text lines include pages of documents and also presented methods based on foreground and background textual character information. In their proposed approach, elements were recognized uniquely at first and were grouped to three clusters. Considering graph concept, the first three characters united to shape groups. Using background information between characters, the direction of added characters of a larger group is determined and based on these directions two candidate regions of different clusters formed. Finally, with the help of these candidate regions unique lines are extracted.

Lee at. el. in [9] introduced a technique for text recognition from binary maps. These kinds of techniques which works on binary maps cannot process scanned maps easily. Scanned maps usually are destroyed by the compression and the noise caused by scanning processes, producing a binary map will have some problems and its processing is a time consuming process.

Pouderoux et. al. [10] presented an automatic approach for extraction and detection of places' names that is based on image segmentation and connected component analysis. Different filtering steps ensure strings and possible characters' stability. Recognized text regions are used as OCR input and then detected words are analysed and improved. The advantage of their work is that there is no initial assumption about font, size and direction.

Roy et. al. [11] detected text strings in the multiple directions, straight or curved. But their method only can cluster strings in one of the four main directions.

In [14] Chiang and Knoblock, in contrary to previous researches, which had worked only on special cases, proposed an approach for non-homogeneous raster maps with multi-direction, curved and multi-sized text. This approach is the most powerful existed algorithm for English text extraction from raster maps.

In the field of separating Persian texts from scanned images of city maps, Kabir et. al. [12,13] made some researches. At the first step, they converted a grey level image to binary and with the usage of area size filter and vision ratio filter divided connected components into two clusters, one for text and the other for graphics [12]. In [13] they used a distance transform based method for clarifying text in city maps. Again, they transformed a grey level image to a binary image. In the obtained binary image, the number of text pixels and graphical lines are less than background pixels. With this property the foreground image is separated from the background. To calculate representation, background pixels are labeled with zero and foreground pixels are labeled with one. Then, they use Euclidean distance measurement to obtain connected components of 8 nearest neighbors. They used the smallest distance for image representation. In both of their method they didn`t use colors' properties, therefore both algorithm could not work properly on color texts and their precision percentages on text extraction are low.

3. IMPLEMENTATION OF THE ALGORITHM

Figure (1) shows a part of Tehran map that we used in this work to report our results. The steps used to obtain the results are described in the following sub-sections.

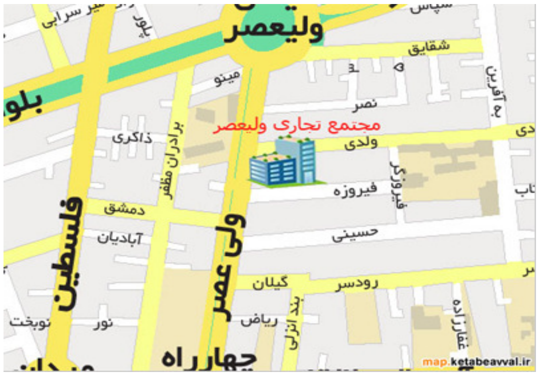


Figure 1. A part of Tehran's map obtained from map.ketabeavval.ir

3.1. Mean-Shift algorithm

Initially, we applied Mean-Shift algorithm to reduce noise and to smooth the color of each region on the map. Mean-Shift algorithm considers the relation between colors in an image (pixel's position) and also considers the color of each pixel. It tries to change a cluster pixels' color to the mean of that cluster. As HSI color space provides a proper human understanding, we used this color space in our method. P(x, y) is the coordinates of pixel P in the image and H, S and I are the color of P.

To reduce noise in a map Mean-shift algorithm starts calculating the mean node for pixel P from Nth neighboring node M(x_m, y_m, h_m, s_m, i_m). The position of the mean node contains mean value on each of the x, y, h, s and I axis of the Nth neighboring node to a local region. As explained in [14] if the distance between M and N become greater than a small threshold, Mean-Shift move N to M and calculate the mean node in the local area again. After convergence, the Mean-shift algorithm considers the values of h, s and i as the color of P(x, y) pixel. The result of applying mean-shift to Figure (1) is illustrated in Figure (2).

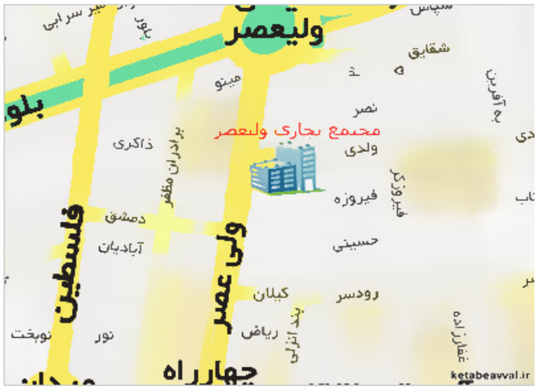


Figure 2. Applying Mean-Shift algorithm to the map shown in Figure 1.

3.2. Changing color space to Lab

Lab color space is a colorproof space with dimension of L for light, and a and b for colorproof dimensions based on non-linear compression coordinates. This color space contains all perceptible colors. This means that its expanse is larger than RGB and CMYK. One of the most important properties of Lab is that it is device independent and this senses that all of the colors are defined depend to their producing natures and the device in which they are shown. Figure (3) demonstrates changing color space process from RGB to Lab.

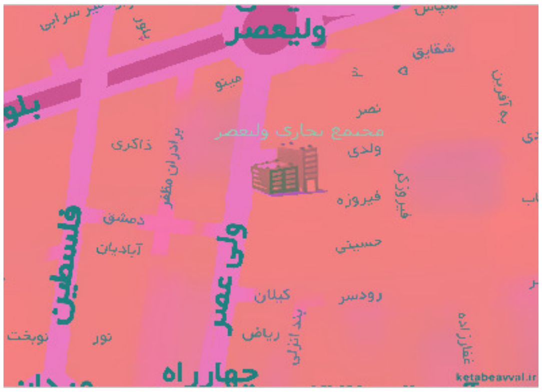


Figure 3. Changing color space from RGB to Lab for perception of colors and preparing it for K-Means.

3.3. K-Means algorithm

K-Means algorithm considers some points haphazardly as the mean points. Then with a distance measurement (usually it is Euclidean measure) each node distance to the nearest chosen node is calculated as the mean point and the new centre of gravity is found. The process is repeated again and again until mean points converge. The purpose of this algorithm is that i observations are classified in to k categories. We applied K-Means algorithm to generate an image that has a maximum of K colors. K-Means algorithm significantly reduces the number of colors in a map with maximizing the variance between classes.

Figure (4) indicates the results of applying K-Means algorithm on figure (3) which was obtained from color space transformation.

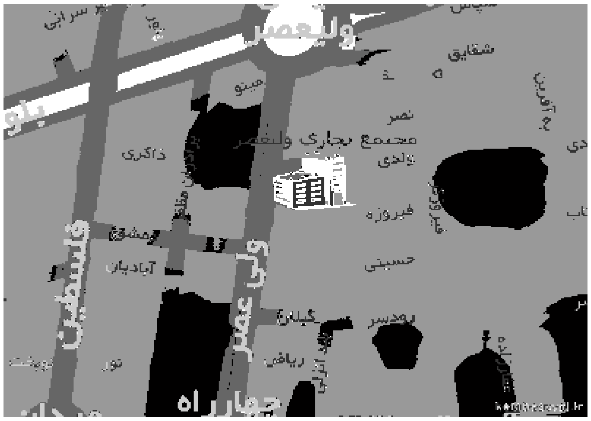


Figure 4. Applying K-Means algorithm on figure (3).

As can be seen from figure (4) all of the colors in the image are converted to K class and each of the regions is labeled to one of these K class. In our experiments K=6 was a good number for K and it had good results on other maps as well.

3.4. Calculating each region's area

As mentioned in previous section, a label is assigned to each region on the map. We consider each region which has a special label as a layer of the map and calculate the area of each region. Then we sort the areas in ascending format and find the median of the area's sizes. With the assumption that if the size of an area is smaller than the median it contains text part, the system chooses three layers as the possible text layers. In the last step the user choses the best layer from these three layers. This step needs user interference because K-Means is a supervised technique and its results may be different in each repetition. In other word, each layer may show different labels in each repetition.

3.5. Removing the largest connected component

In most of the maps, a long road line or a large non-text element is still in our best layer. We used the connected components analysis to find the largest connected component and remove this possible non-text region. System will ask the user whether or not to remove the connected component. If the user chooses yes, it will be omitted and if the user choose no it will be remained on the map. Figure (5) shows the results of choosing best layer between the possible layers.

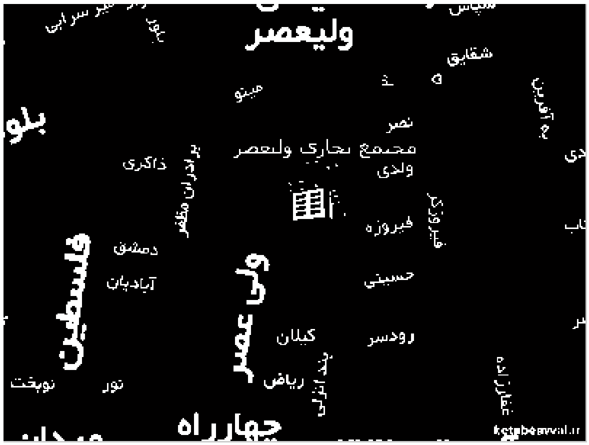


Figure 5. The best textual layer

4. Experimental Results

We have done our experiments on 40 maps from Google, Yahoo, Tehranmaps, Ketabeavval, a number of maps from Tehran municipality's site and some other maps from different sources. Figures (6-10) show another example in which all steps of our algorithm are applied to an English map. This shows that the method is applicable to maps with Latin texts as well.

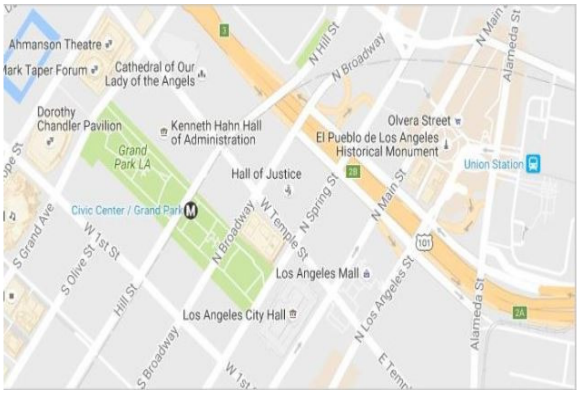


Figure 6. A map with English text obtained from maps.google.com

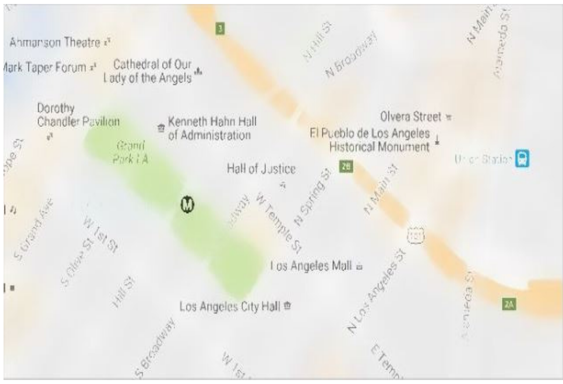


Figure 7. Applying Mean-Shift algorithm to map in figure (6).

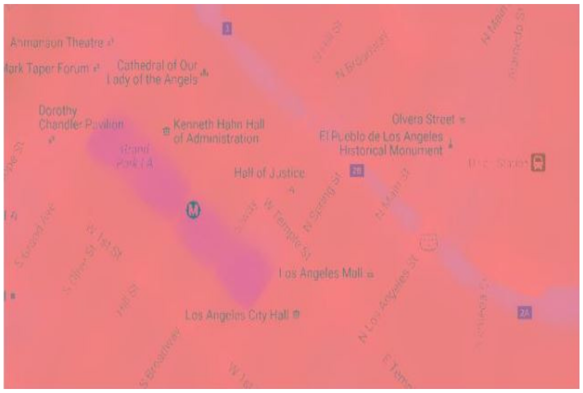


Figure 8. Map of figure (7) after color transformation.

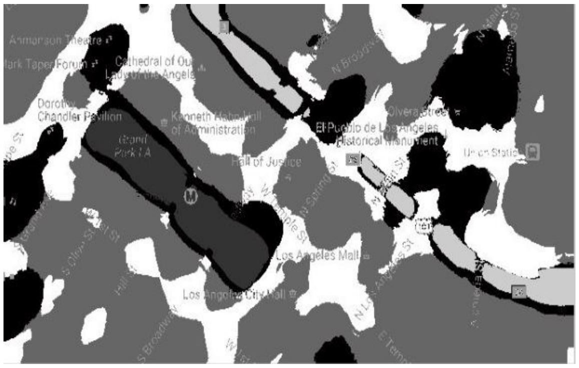

Anmarson Theatre
lark Taper Forum
Cathedral of Our Lady of the Angels
Dorothy Chandler Pavilion
Kenneth Hahn Hall of Administration
Grand Park
Hall of Justice
El Pueblo de Los Angeles Historical Monument
Olvera Street
Union Station
Los Angeles Mall
Los Angeles City Hall
N Broadway

Figure 9. Results after applying K-Means algorithm

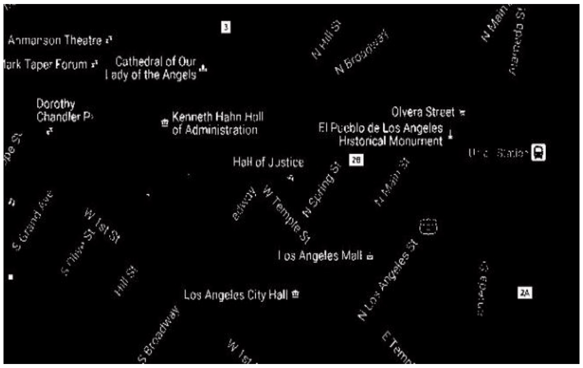


Figure 10. The best textual layer

As researches on Persian texts are very limited and the results of the previous works were poor, it was not possible to make a comparison. However, we calculated precision and recall percentages with the following formulas:

$$\text{Precision}= TP/ (TP+FP) \quad (1)$$

$$\text{Recall}= TP/ (TP+FN) \quad (2)$$

In these formulas, TP (True Positive) refers to text areas which our system could recognize as text correctly. FP (False Positive) refers to non-text areas which our system detected them as text wrongly and FN (False Negative) refers to text areas which our system could not recognize them, wrongly. We counted each word as a candidate for text areas and considered each road line or each graphical symbol as a non-text area. Experimental results show %95.06 precision with %85.72 Recall. This confirms the applicability of the proposed approach for maps with Persian text with good precision. It should be noted that the proposed method is not limited to Persian/Arabic texts and can also work on maps with Latin text.

We are currently working on our approach to make the last step, which can find the best textual layer, automatic and make the texts ready for OCR software.

5. CONCLUSION

This paper presents an algorithm for extracting both Persian and English texts from raster maps. The proposed method is independent of font, size, color and direction of texts. We use a

Mean-Shift algorithm with proper parameter adjustment and apply color transformation to make the maps ready for K-Means algorithm, which quantizes the colors in maps to six levels. By comparing to a defined threshold the text layer candidates are then limited to three. The best layer is finally selected by user. Experimental results show %95.06 precision with %85.72 Recall for the proposed approach.

REFERENCES

[1] M. Tabassum and M. Shorif Uddin, (2011) "Extraction of ROI in Geographical Map Image," Journal of Emerging Trends in Computing and Information Sciences, Vol 2, No. 5, pp. 237-242.

[2] G. K. Myers and P. G. Mulgaonkar, (1996) "Verification-Based Approach for automated Text and Feature Extraction from Raster-Scanned Maps", Springer, Vol. 1072, pp. 190-203.

[3] Y. Y. Chiang and C.A. knoblock, (2010) ,"An approach for recognizing text labels in raster maps," International Conference on pattern recognition, pp. 3199-3202.

[4] Y-Y. Chiang and C.A. Knoblock, (2014) " Recognizing text in raster maps", GeoInfomatica, Vol 19, Issue 1, pp. 1-27.

[5] LA. Fletcher and R. Kasturi, (1988) "A robust algorithm for text string separation from mixed text/graphics images". IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 10, Issue 6, pp. 910-918.

[6] L-H. Chen, J-Y. Wang, (1997) "A system for extracting and recognizing numeral strings on maps", Proceedings of the 4th international conference on document analysis and recognition, Vol 1, pp. 337–341.

[7] Vel´ azquez A, Levachkine S, (2004) "Text/graphics separation and recognition in raster-scanned color cartographic maps", Graphics recognition. Recent Advances and perspective, Springer, Vol 3088, pp. 63-74.

[8] Roy PP, LIados J, Pal U, (2007) "Text/graphics separation in maps", International Conference on computing Theory and Application, pp. 545-551.

[9] L.Li, G. Nagy, A.Samal, SC.Seth and Y.Xu, (2000) "Integrated text and line-art extraction from a topographic map", IJDAR, Vol 2, Issue 4, pp. 177-185.

[10] J . Pouderoux, JC . Gonzato, A . Pereira and P. Guitton, (2007) "Toponym recognition in scanned color topographic maps", 9th international conference on document analysis and recognition, Vol 1, pp. 531–535.

[11] PP. Roy, U. Pal, J. Llad´ os and F. Kimura, (2008) "Multi-oriented English text line extraction using background and foreground information", The eighth IAPR international workshop on document analysis systems, pp. 315–322.

[12] A. Kabir, A. Ghaffari, K. Kangarloo (2010), "Sepration of Persian text from scanned metropolitan maps", 17th Iranain Conference on Image processing and machine learning, pp 1-4.

[13] A. Kabir, A. Ghaffari, K. Kangarloo (2011), "A method based on distance conversion for revealing text in metropolitan map images", 20th Iranian Conference on Electrical Engineering, pp. 1-4.

[14] Y-Y. Chiang and CA. Knoblock, (2011) "A general approach for extracting road vector data from raster maps," IJDAR, Vol. 16. pp. 55-8.

AUTHORS

Sanaz Hadipour Abkenar studied electronic engineering in Guilan university (2011) for her bachelor degree and is now a Master student of communication Engineering in Shahrood University of Technology. Her main interest is digital image processing.

Alireza Ahmadyfard received Ph.D. in image processing and Computer vision from CVSSP (Center for Vision Speech and Signal Processing) at University of Surrey in 2002. He is director of Electrical Engineering Department in Shahrood university of technology. His research interests are digital signal processing, object recognition, image based inspection and human identification using biometrics.

EHR ATTRIBUTE-BASED ACCESS CONTROL (ABAC) FOR FOG COMPUTING ENVIRONMENT

Aisha Mohmmed Alshiky, Seyed M. Buhari and Ahmed Barnawi

King Abdulaziz University, KSA, Jeddah

ABSTRACT

Cisco recently proposed a new computing environment called fog computing to support latency-sensitive and real time applications. It is a connection of billions of devices nearest to the network edge. This computing will be appropriate for Electronic Medical Record (EMR) systems that are latency-sensitive in nature. In this paper, we aim to achieve two goals: (1) Managing and sharing Electronic Health Records (EHRs) between multiple fog nodes and cloud, (2) Focusing on security of EHR, which contains highly confidential information. So, we will secure access into EHR on Fog computing without effecting the performance of fog nodes. We will cater different users based on their attributes and thus providing Attribute Based Access Control ABAC into the EHR in fog to prevent unauthorized access. We focus on reducing the storing and processes in fog nodes to support low capabilities of storage and computing of fog nodes and improve its performance.

KEYWORDS

Fog computing, Electronic Medical Record (EMR), Electronic Health Record (HER)

1. INTRODUCTION

The explosive increase in the use of sensors and sensing information leads to the scope of producing plenty of future applications. The most important requirement in these applications is low-latency processing and as known centralizing of services at the core of the Internet in the cloud computing may lead to high latency which is rejected. While there are numerous economic advantages of cloud, there is a problem for latency-sensitive applications due to frequent movements of huge data from the source to the server/cloud [1].

The latency-sensitive and real time applications require nodes in the vicinity to provide fast responses. A new platform is needed to achieve these requirements; [2] Cisco recently proposed a new computing environment called fog computing, call it “Fog”, simply because fog is a cloud close to the ground. It is a connection of billions of devices (called as fog nodes) around the globe. It's different from Cloud Computing: distribution of processing in distributed nodes with mobility. In fog computing environment, the generic application runs logic on resources throughout the network, including dedicated computing nodes and routers [3]. "The emerging

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 87– 95, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70209

Fog Computing architecture is a highly virtualized platform that provides compute, storage, and networking services between end devices and traditional Cloud Computing data centers, typically, but not exclusively located at the edge of the network" [4].

However, developing applications using fog computing resources is critical because it includes heterogeneous resources at different levels of network hierarchy to provide low latency and scalability requirement for new applications [3]. In this research, the Fog environment is considered to be an appropriate platform to implement Electronic Health Records (EHR). Nowadays, in modern healthcare environments, healthcare providers are shifting their electronic medical record systems to clouds [5]. But as the cloud is not good choice for real time and latency sensitive applications, we propose that the Fog computing is appropriate for EHR. EHR contains private and sensitive patient health information which are needed to be secured and the privacy of the patient must be ensured. Security in Fog Computing Environment will eventually become an issue; with security embedded into the Fog Computing environment, we envision, in this research, to provide appropriate security solutions without effecting on performance. With the proposal of Attribute Based Access Control which is a flexible and logical mechanism [6], we will cater different users based on their attributes, object (information and resources) attributes and environment conditions (time and location). Thus, providing secure access mechanism into the EHR fog to prevent unauthorized access to fog and also prevent leaks of information; user-based attributes might be related to a targeted application.

In this paper, we introduce an innovative ABAC architecture for EHR in fog computing environment as an alternative that provides inherent advantages that will improve the security measures related to EHR. In addition to that, we exhibit that the introduction of fog computing will outperform the cloud based alternatives. This paper is organized as follows, in Section 2, we review the literature and present related work in cloud computing access control architecture. In Section 3, we list few applications for Fog Computing with emphasis on Healthcare sector. In Section 4, we describe our ABAC Fog computing architecture. In Section 5, we propose a location based Fog Computing ABAC architecture and analyze the proposed system against security threats and risks. In Section 6, we conclude and summarize our future work.

2. LITERATURE REVIEW

Instead of cannibalizing Cloud Computing, Fog Computing allows a new type of applications and services, and that there is a rich interplay between the Cloud and the Fog, mainly when it comes to data management and analytics. This review is mostly related to work and deals with the potential risks of privacy exposure to the healthcare system and implement electronic health record (EHR) in fog computing [1]. Security in Fog Computing Environment will eventually become an issue; this issue is not being investigated yet and it seems to be completely absent in the literature. For that, this section discusses a number of related and similar researches that provide security of cloud system especially for EHR.

One of studies [7] explains that patients' records must be accessible only by authorized users and they justified that patients should have the opportunity to exert the control over their own data. For that, they proposed a cryptographic access control scheme allowing patients to grant medical teams authorizations to access their medical data. They proposed a schema consists of decentralized hierarchical key agreement protocol to securely establish a hierarchy of crypto keys in agreement with the privilege levels of the team members. The scheme provides data confidentiality, but it must be guaranteed that hierarchical keys are unique and "fresh" for each run of the protocol which require high computation.

As multiple entities will interact with the data, the authors in [8] explain that access to sensitive resources should be provided only to authorized users and tenants. They adapt Task-Role-Based Access Control, which considers the task in hand and the role of the user. They support both workflow based and non-workflow based tasks and authorize subjects to access necessary objects only during the execution of the task. Classification of tasks and activities has been done on the basis of active and passive access control and inheritable and non-inheritable tasks. Each user is assigned a role, roles are assigned to workflow or non-workflow tasks, and tasks are assigned to permissions. This model only supports the scenarios when the roles are defined within a single healthcare organization. It is designed to support healthcare service provided in a single healthcare organization. So, the access should be restricted and provided only during the execution of a specific task.

In [5] and [9], the authors mainly focuses on access control issues when EHRs are shared with various health care providers in cloud computing environments. In [5], they proposed a unified access control scheme which supports patient-centric selective sharing of virtual composite EHRs using different levels of granularity, accommodating data combination and various privacy defense requirements. However, this approach assumes that all health care providers adopt a unified EHR schema, which is not applicable in cloud environments. In [9], the authors try to overcome this limitation by supporting EHRs aggregation from various health care providers considering different EHR data schemas in cloud environments. They propose a systematic access control mechanism to support selective sharing of composite electronic health records aggregated from various health care providers in the cloud. They present algorithms for EHRs data schema composition and cross-domain EHR aggregation.

In [10], the authors explain that Attribute-Based Encryption ABE (data can only be read by a user with certain attributes [10] suitable for electronic health records system in the cloud, in which many users can retrieve the same EHR while each user can only decrypt the parts that they are allowed to read. The authors here try to handle some problems such as when a user with multiple roles might cause information leakage and computational overhead on EHR owners. Hence, they adopt both ABE and Identity Based Encryption IBE (a type of public-key encryption in which the public key of a user is unique user identity) and integrate them into their hierarchical framework. ABE is used to achieve fine-grained access control while IBE is used to securely transmit ABE keys. EHRs are encrypted on the Trusted Server and then are uploaded to the cloud. Decryption keys are also generated on trusted server and are distributed to domain servers that are then responsible for distributing the decryption keys to authorized entities. This framework addresses only the case of read access. This solution was suitable for an environment which has large number of users (subject) because it depends on their attributes which need not be predefined for each user.

Many research works proposed important and useful concepts of the EHR security [5, 7, 8, 9, and 10]. However, there are several uncertain issues. One of those issues is how to manage information of PHR and bring it near the user to support quick access of these information in timely manner. Therefore, allowing a hospital staff to access patient information (EHRs) in short period is essential. Information stored in the patient’s EHR may help a medical staff to make better decisions. In some emergency healthcare situations, immediate exchange of patient's EHRs is crucial to save lives. In our research, we try to handle the EHR near to the medical staff and provide quick response for patient needs. We will support that by implementing part of EHR in suitable and nearest fog nodes and we propose that Attribute Based Access Control (ABAC) that depends on attributes of subject (who want to access), object (services or information), action attributes (view or delete patient information) and environment conditions (time and location). This approach is flexible and it decreases the administrative overhead [6].

3. FOG COMPUTING APPLICATION IN HEALTHCARE

In this section, we will review some of studies that applied fog computing in health care system. How to develop real-world fog computing-based universal health monitoring system is still an open question.

In [11], pervasive fall detection is employed for stroke mitigation. There were four major contributions in this study: (1) they examined and developed a set of new fall detection algorithms built on acceleration magnitude values and non-linear time series analysis techniques, (2) they designed and employed a real–time fall detection system employing fog computing paradigm, which distributes the analytics through the network by splitting the detection tasks between the edge nodes (e.g., smart phones attached to the user) and the server (e.g., cloud), (3) they examine the special needs and constraints of stroke patients and they proposed patient centered design that is minimal intrusive to patients and (4) their experiments with real-world data displayed that their proposed system achieves the high specificity (low false alarm rate) while it also achieves high sensitivity. Depend on researchers knowledge, their proposed system is the first large scale, real-world pervasive health monitoring system that employs the fog computing paradigm and distributed analytics.

Ultraviolet (UV) radiation has a great effect on human health. Since sensors in mobile phone cameras are very sensitive to UV, mobile phones have the potential to be an ideal equipment to measure UV radiance. The research [12] investigated theoretical foundations that control mobile phone cameras without any add-on to measure solar UV in open environment. Theoretical foundations accomplished to a procedure that can be deployed to any mobile phone with a camera. In addition, by utilizing fog computing, results can be collected and edited locally through fog server to provide accurate UV measurement. Furthermore, an Android app called UV Meter was established based on the procedure that can be implemented in mobile phones. Verification was conducted under unlike weather conditions and their results showed that the procedure is valid and can be implemented onto mobile phones for everyday UV measurement.

In another study [13], efficient IoT-enabled healthcare system architecture which benefits from the concept of fog computing is presented. The effectiveness of fog computing in IoT-based healthcare systems in terms of bandwidth utilization and emergency notification is demonstrated. In addition, they utilized ECG feature extraction at the edge of the network in their implementation as a case study. They proposed that to perform functionalities of gateways, the smart gateway should have the ability to offer a high level of advanced services in the fog computing platform. The smart gateway architecture including physical and operational structures is elaborately designed and described.

4. ABAC IN FOG COMPUTING ENVIRONMENT

Our policy framework adopts attribute-based security framework where in all users are authenticated and identified based on a set of attributes which are associated to each request. In our proposed framework, the ABAC is implemented and enforced at fog node which receives user access request. Each fog node which received requested action will analyze the attributes that is associated with the request. Then, based on these retrieved attributes and policies schema, the permission will be granted to user.

Our proposed solution used the recommended architecture of ABAC [14] as shown in Figure 1. As mentioned above, this architecture implemented in the edge of network (fog nodes). For that, authenticated and authorized access into EHR is applied on request at the nearest fog instead of at the core of the network (cloud).

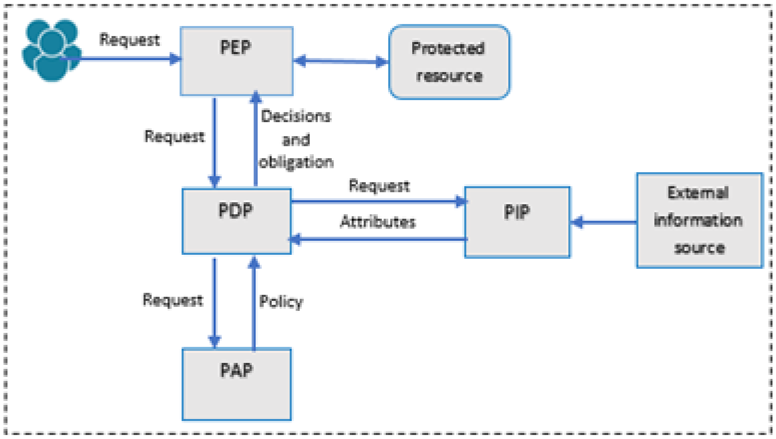


Figure 1. ABAC Architecture

- The PEP or Policy Enforcement Point examines the request and produces an authorization request and sends to the PDP.

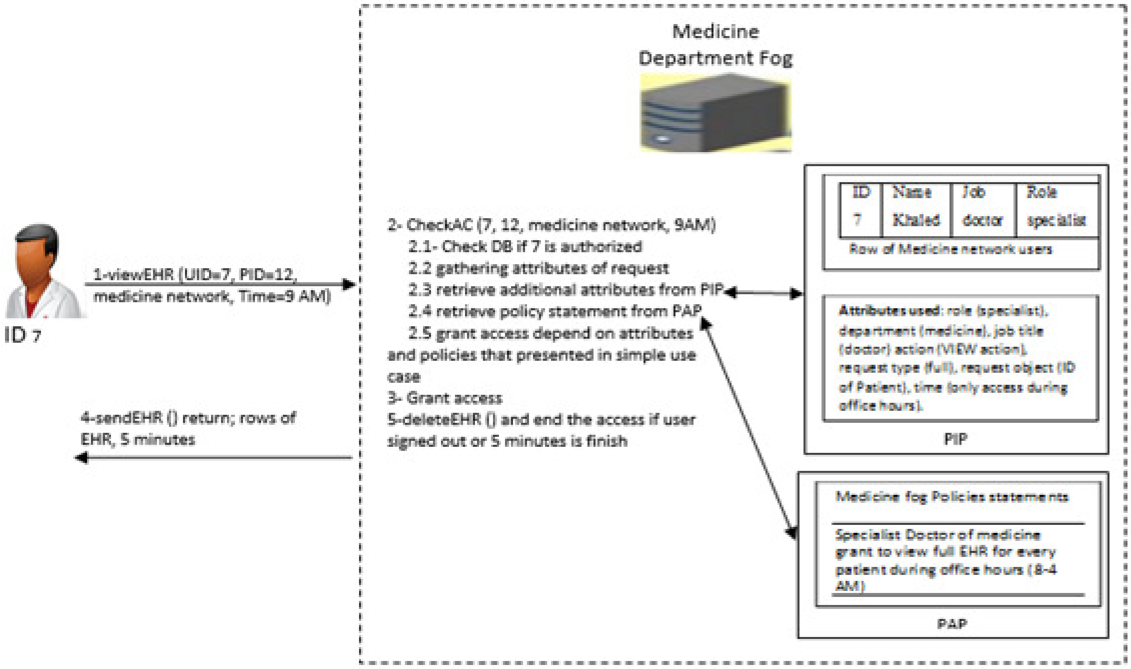
- The PDP or Policy Decision Point evaluates incoming requests against policies that has been constructed. The PDP returns a Permit / Deny decision.

- The PAP or Policy Administration Point maintaines the policies and bridges PDP to policies statements. The administrator of host (fog nodes/cloud) is responsible to defines policises of its host. The multi-tenant nature of the fog computing model raises the requirement for an administrator to define policies that bind a user to healthcare system and implement policy schema. Each fog node has specific polices which are applied only to its users.

- The PIP or Policy Information Point maintains descriptive attributes and bridges the PDP to external sources of attributes e.g. databases. The administrator of host (fog nodes/cloud) is responsible to define PIP of its host. He prepares data schema that specifies a set of defined attributes associated with a physical or virtual component. Each fog has data schema of its users only to avoid unused stored database. The attributes considered in our proposed ABAC are:

- Subject attributes (department, role and job title)
- Action attributes (view or delete patient information)
- Object attributes (object type, sensitivity of data)
- Environment attributes (time and location)

Simple use case of requested action (user 7 view patient 12 record) from Dr Khaled to medicine department fog is presented in Figure 2 below.


Medicine
Department Fog
ID 7
1-viewEHR (UID=7, PID=12, medicine network, Time=9 AM)
2- CheckAC (7, 12, medicine network, 9AM)
2.1- Check DB if 7 is authorized
2.2 gathering attributes of request
2.3 retrieve additional attributes from PIP
2.4 retrieve policy statement from PAP
2.5 grant access depend on attributes and policies that presented in simple use case
3- Grant access
5-deleteEHR () and end the access if user signed out or 5 minutes is finish
4-sendEHR () return; rows of EHR, 5 minutes
Row of Medicine network users
Attributes used: role (specialist), department (medicine), job title (doctor) action (VIEW action), request type (full), request object (ID of Patient), time (only access during office hours).
PIP
Medicine fog Policies statements
Specialist Doctor of medicine grant to view full EHR for every patient during office hours (8-4 AM)
PAP

ID	Name	Job	Role
7	Khaled	doctor	specialist

Figure 2. Simple use case of ABAC

5. Location Based ABAC Fog Computing Architecture

Depending upon location of fog device, the managing and sharing of EHRs between multiple fog nodes and cloud is maintained. The important issue that we considered in our solution is the low capabilities of storage and computing of fog nodes. We focus on reducing the storing and processes in fog nodes to serve the availability of fog, to improve its performance and efficiency. To achieve these goals we proposed that:

- All hospital information and needs are maintained in Cloud (data center)
    - Full version of Electronic Medical System (EMR) which contain EHRs of all patients in hospital are implemented in cloud (data center).
    - It serves all the hospital department's users.

- There is a fog device for every single hospital department
  - Part of EMR is implemented in fog, which provide only services that are needed by department's users to do their job.
  - Contains information and attributes of department network user and predefined access policies.
  - Fog applies ABAC into incoming request for each attempted access.
  - Fog maintains temporary and timely storage of EHR.
- Scheduling of EHR sharing between cloud and specific fog in specific location.
  - Movement timeline of visiting patient in hospital is estimated first. This estimation is assigned once patient visits reception department and reception user tries to access to patient information in cloud.
  - After first access of the patient record in cloud by the receptionist, scheduling of EHR sharing between cloud and specific fog in specific location occurred depends on proposed estimation. For example, patient Khaled will be directed to laboratory department after reception department within 5 minutes. So, depending upon proposed estimation, the cloud will send copy of visiting patient EHR into specific location of fog within 5 minutes.
  - Timing of patient services in specific department is estimated. Once patient arrives and user in this location (department) starts to serve him/her, the timer is started and after the timer ends, the EHR is deleted from the temporary storage of fog.

Simple scenario is presented in Figure 3 to explain simple patient workflow from reception to laboratory and time of EHR sharing between cloud and specific fog in specific location (laboratory). It is estimated that patient after 6 minutes (360 s) will go to laboratory department. Before the patients' arrival to laboratory department, the cloud will send copy of visiting patient EHR to specific location (laboratory department).

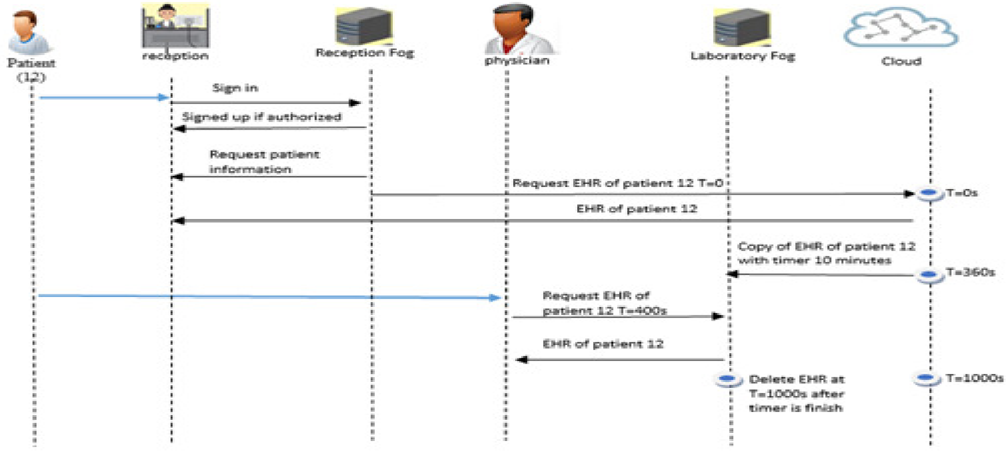

Patient (12)
reception
Reception Fog
physician
Laboratory Fog
Cloud
Sign in
Signed up if authorized
Request patient information
Request EHR of patient 12 T=0
T=0s
EHR of patient 12
Copy of EHR of patient 12 with timer 10 minutes
T=360s
Request EHR of patient 12 T=400s
EHR of patient 12
Delete EHR at T=1000s after timer is finish
T=1000s

Figure 3. EHR sharing between cloud and fog

6. Conclusion and Future Work

We provided ABAC into the EHR in fog to prevent unauthorized access. Also, we considered in our solution the low capabilities of storage and computing of fog nodes by focusing on reducing the storing and processes in fog nodes to serve the availability of fog, to improve its performance and efficiency.

In our future work, we will simulate our solution by using iFogSim tool and we will evaluate the results of our solution.

References

[1] Hong K, Lillethun D, Ramachandran U, Ottenwälder B, Koldehofe B, editors. Opportunistic spatio-temporal event processing for mobile situation awareness. Proceedings of the 7th ACM international conference on Distributed event-based systems; 2013: ACM.

[2] Zhu J, Chan DS, Prabhu MS, Natarajan P, Hu H, Bonomi F, editors. Improving web sites performance using edge servers in fog computing architecture. Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on; 2013: IEEE.

[3] Hong K, Lillethun D, Ramachandran U, Ottenwälder B, Koldehofe B, editors. Mobile fog: A programming model for large-scale applications on the internet of things. Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing; 2013: ACM.

[4] Bonomi F, Milito R, Zhu J, Addepalli S, editors. Fog computing and its role in the internet of things. Proceedings of the first edition of the MCC workshop on Mobile cloud computing; 2012: ACM.

[5] Wu R, Ahn G-J, Hu H, editors. Secure sharing of electronic health records in clouds. Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on; 2012: IEEE.

[6] NIST GS, Goguen A, Fringa A. Risk Management Guide for Information Technology Systems. Recommendations of the National Institute of Standards and Technology. 2002.

[7] Boyd C, Mathuria A. Protocols for authentication and key establishment: Springer Science & Business Media; 2013.

[8] Narayanan HAJ, Güneş MH, editors. Ensuring access control in cloud provisioned healthcare systems. 2011 IEEE Consumer Communications and Networking Conference (CCNC); 2011: IEEE.

[9] Jin J, Ahn G-J, Hu H, Covington MJ, Zhang X. Patient-centric authorization framework for electronic healthcare services. computers & security. 2011;30(2):116-27.

[10] Huang J, Sharaf M, Huang C-T, editors. A hierarchical framework for secure and scalable ehr sharing and access control in multi-cloud. 2012 41st International Conference on Parallel Processing Workshops; 2012: IEEE.

[11] Cao Y, Chen S, Hou P, Brown D, editors. FAST: A fog computing assisted distributed analytics system to monitor fall for stroke mitigation. Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on; 2015: IEEE.

[12] Mei B, Cheng W, Cheng X, editors. Fog Computing Based Ultraviolet Radiation Measurement via Smartphones. Hot Topics in Web Systems and Technologies (HotWeb), 2015 Third IEEE Workshop on; 2015: IEEE.

[13] Gia TN, Jiang M, Rahmani A-M, Westerlund T, Liljeberg P, Tenhunen H, editors. Fog Computing in Healthcare Internet of Things: A Case Study on ECG Feature Extraction. Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on; 2015: IEEE.

[14] Coyne E, Weil TR. ABAC and RBAC: scalable, flexible, and auditable access management. IT Professional. 2013;15(3):0014-16.

INTENTIONAL BLANK

MODELLING AND APPLICATION OF A COMPUTER-CONTROLLED LIQUID LEVEL TANK SYSTEM

Hayati Mamur[1], Ismail Atacak[2], Fatih Korkmaz[3] and M.R.A. Bhuiyan[1]

[1]Department of Electrical and Electronics Engineering, Faculty of Engineering,
Manisa Celal Bayar University, 45100, Manisa, Turkey
[2]Department of Computer Engineering, Faculty of Technology, Gazi University,
06100, Teknikokullar, Ankara, Turkey
[3]Department of Electrical and Electronics Engineering, Faculty of Engineering,
Cankiri Karatekin University, 18100, Cankiri, Turkey

ABSTRACT

Liquid level tanks are employed in many industrial and chemical areas. Their level must be keep a defined point or between maximum-minimum points depending on changing of inlet and outlet liquid quantities. In order to overcome the problem, many level control methods have been developed. In the paper, it was aimed that obtain a mathematical model of an installed liquid level tank system. Then, the mathematical model was derived from the installed system depending on the sizes of the liquid level tank. According to some proportional-integral-derivative (PID) parameters, the model was simulated by using MATLAB/Simulink program. After that, data of the liquid level tank were taken into a computer by employing data acquisition cards (DAQs). Lastly, the computer-controlled liquid level control was successfully practiced through a written computer program embedded into a PID algorithm used the PID parameters obtained from the simulations into Advantech VisiDAQ software.

KEYWORDS

Computer-controlled system, Data acquisition card, Level control, PID, Process control

1. INTRODUCTION

Liquid level tanks are used to keep the liquid level a certain point or between particular values in chemical industry. To accomplish this goal, many automatic control methods have been suggested in literature. When these methods carefully inspected, the objective of them are that the changing of the tank level due to the variations of inlet and outlet liquid quantities brings to as quickly and accurately as the defined point [1–12].

A closed loop automatic control system is given in Fig. 1. In practice, some controllers such as computers connected data acquisition cards (DAQs), compact proportional-integral-derivative (PID) devices, programmable logic controllers (PLCs), microcontrollers (μCs) and digital signal processors (DSPs) are commonly utilized.

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 97– 106, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70210

The controllers connect the physical and non-physical parts of the systems. Firstly, they take the sensor signal *y*(t) measuring process variables (PV), then compare with a set point *r*(t) (SP) and after that find out an error *e*(t). Lastly, they give an output signal (CV - MV) *u*(t) in order to zeroize the error depending on the used control methods and accumulated errors.

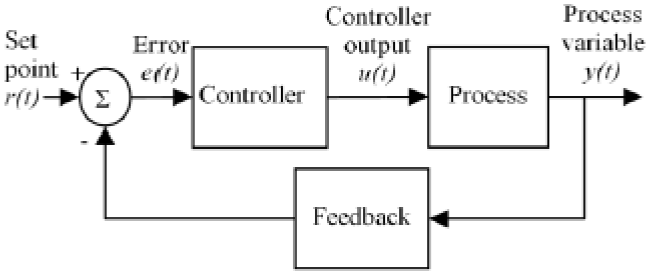


Figure 1. A block diagram of a closed loop automatic control system

The adjusting of the PID parameter values in terms of the used process in controller is essential during the taking into operation. Several methods are used for adjusting of the PID parameter values [1]. There is a specific parameter adjusting for each process. However, the certain adjusting of the PID parameter cannot be achieved to zeroize the error in some cases. The adjusting of parameter values and also the understanding of how each parameter value affects a process are very significant topic for automatic control systems [2].

In this study, in order to turn into application to knowledge, a real computer-controlled liquid level tank control embedded into a conventional PID control algorithm was accomplishedly executed by means of the DAQs. The simulations of the system were performed by MATLAB/Simulink software.

This executed study was presented as follow: In the second section, fundamental control methods were shortly explained. The mathematical model of the system and the simulation practices and results were clarified in the third section. Then, the installation of the developed liquid level control set was displayed in the fourth section. After that, the results and discussion of the experimental applications applied to the mathematical model and the conventional PID algorithm on the system were displayed in the fifth section. Finally, conclusion and future perspectives were given in sixth sections, respectively.

2. FUNDAMENTAL CONTROL METHODS

On-off, proportional (P), proportional-derivative (PD), proportional-integral (PI), and PID controls are fundamental control methods employed to regulate process variable to a specific set point in industrial control systems. These methods were initially realized using mechanical devices and after that, they were designed by pneumatic and analog electronic devices.

In on-off control, a controller opens or closes a final control element in accordance with the case in which the process variable is over or under the set point value. PID control is one of the most basic methods widely used in industrial control systems [3]. The method can be easily applied to the systems with linear and simple structure [4]. In contrast, it is quite difficult to apply nonlinear systems especially when with dead time delays [5]. The general output equations of the PID controller in time and Laplace domains are given as follow:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau + K_d \frac{de}{dt}, \tag{1}$$

$$u(t) = K_p \left[e(t) + \frac{1}{T_i} \int_0^t e(\tau)\, d\tau + T_d \frac{de}{dt} \right], \tag{2}$$

$$G(s) = K_p + \frac{K_i}{s} + K_d\, s = \frac{K_d s^2 + K_p s + K_i}{s}. \tag{3}$$

Where *u(t)* is the controller output signal fed to the process to be controlled, *e(t)* is the error signal: The difference between the set point and the measured process variable ($e(t) = r(t) - b(t)$), K_p, K_i and K_d are the controller gain parameters, T_i is the integral time constant and T_d is the derivative time constant. The gain parameters K_i and K_d in Laplace domain equation are calculated by the equations $K_i = K_p/T_i$ and $K_d = K_p.T_d$, respectively.

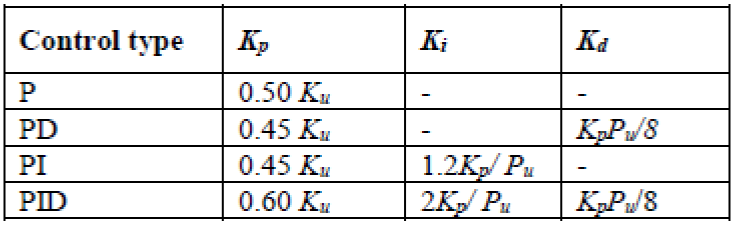
PID control accumulates itself all characteristic features of P control, I control and D control and thus, sometimes called three-term control. This type control is usually used more quickly to stabilize the process controlled by PI control.

2.1. Setting of the PID controller gain parameters

In the setting of PID controller parameters by the Ziegler-Nicholas's (Z-N) sustained oscillation method, the sufficient knowledge and experience about the process makes possible more accurate and faster parameter optimization [6]. This method is experimental and subjects the closed loop control system to an experiment through only proportional gain.

As a result of the applying of the processing procedures to the controller connected to the process, the data required for the calculation of gain parameters are obtained experimentally. Then, these data are placed in the equations in Table 1, recommended by Ziegler and Nichols, and the final values are obtained for the gain parameters K_p, K_i and K_d.

Table 1. According to the Z-N's sustained oscillation method.



Control type	K_p	K_i	K_d
P	0.50 K_u	-	-
PD	0.45 K_u	-	$K_pP_u/8$
PI	0.45 K_u	1.2K_p/ P_u	-
PID	0.60 K_u	2K_p/ P_u	$K_pP_u/8$

3. Mathematical Model and Simulation

Before proceeding to the implementation stage, the knowing of how to exhibit a dynamic behavior of liquid level control systems is vital in terms of the determining of a number of parameters related to the system in the design stage [7]. The dynamic behavior of a system can be observed and analyzed through some simulation studies after the obtaining of the mathematical model to the system [8]. As in many physical systems, liquid level control systems also exhibit a non-linear dynamic behavior due to the inherent characteristics [9].

Nevertheless, the theories and theorems in control systems, the most of which can be only applied to linear systems, necessitates to be modeled nonlinear systems as a linear system [10]. When considering small changes in the level control system, the system can be modeled with linear differential equations. In this section, the linearization of a nonlinear liquid level control system model has been discussed. The schematic diagram of the first order liquid level control system is shown in Fig. 2.

This system consists of a tank with an inlet (control) valve and outlet (load) vane and represents a single input single output (SISO) control system. In the system, while the outflow liquid from the tank is manually controlled through the load vane, the inflow liquid into the tank is adjusted by a proportional valve. Normally, the outflow liquid from the tank is a load which is needed by process and continuously changes due to reasons beyond control. Therefore, the inflow liquid into the tank represents a manipulated variable (MV) depending on the liquid level [11]. The outflow liquid from the tank refers to a load or a disturbance. The top of the liquid tank is open and it has a cylindrical structure. The dimensions and some calculations of the liquid level control tank are given in Table 2. The liquid level control system has been modeled taking into account the change in the liquid level, which results from the difference between the inlet flow rate and outlet flow rate of the liquid in the tank. This system can be considered as a simple circuit including a capacity and a resistance.

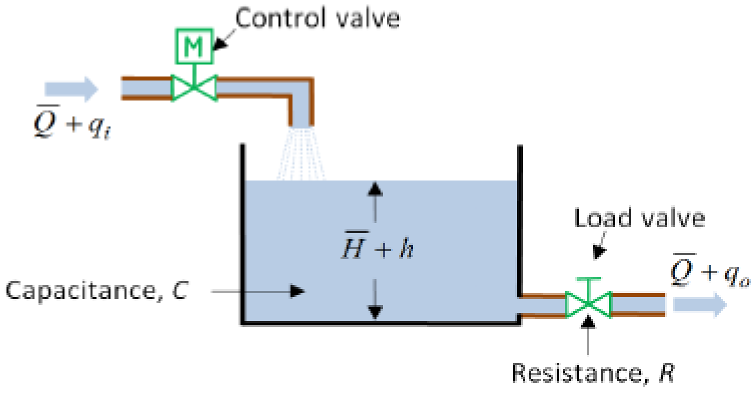


Figure 2. A schematic diagram of the liquid level control system

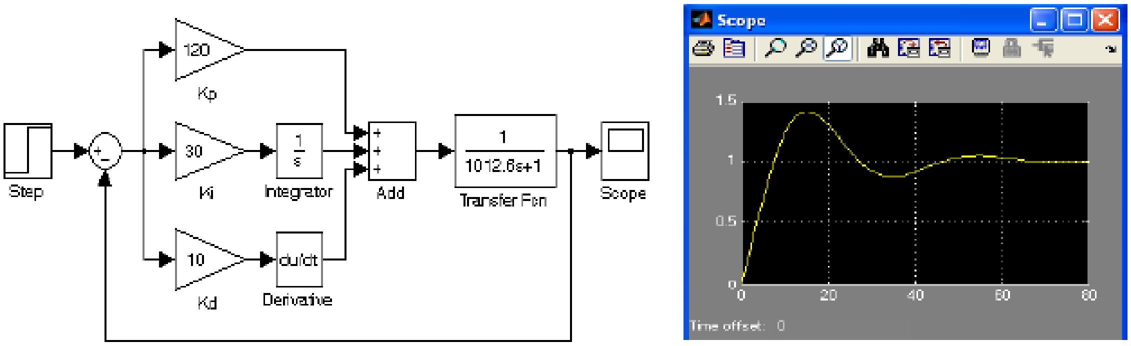
Table 2. The dimensions of the tank

Properties	**Values**
The height of the tank, h (m)	1
The diameter of the tank, d (m)	0.15
The cross sectional area of the tank, A (m^2)	0.5063
The capacitance of the tank, C (m^3/m)	0.5063
The volume of the tank, V (m^3)	0.0176
The maximum liquid flow, Q_o (l/s)	0.5
The maximum liquid flow, Q_o (m^3/s)	0.0005
Resistance, R (s/m^2)	2000

The characteristic equation of the obtained transfer function, simply the denominator of this function, is first order. Therefore, the dynamic behavior of the system is defined in form of time constant [12]. When the calculated the resistance of the liquid level system R and the capacitance C (m^3/m) values in Table 2, the transfer function of the liquid level control system is achieved by:

$$G(s) = \frac{Q_o(s)}{Q_i(s)} = \frac{1}{1012.6s + 1} \tag{4}$$

The Simulink model of the whole system, which consists of the mathematical model derived for the liquid level control system and the PID controller with the tuned gain parameters, is given in Fig 3a. The step response of the system is shown in Fig. 3b.

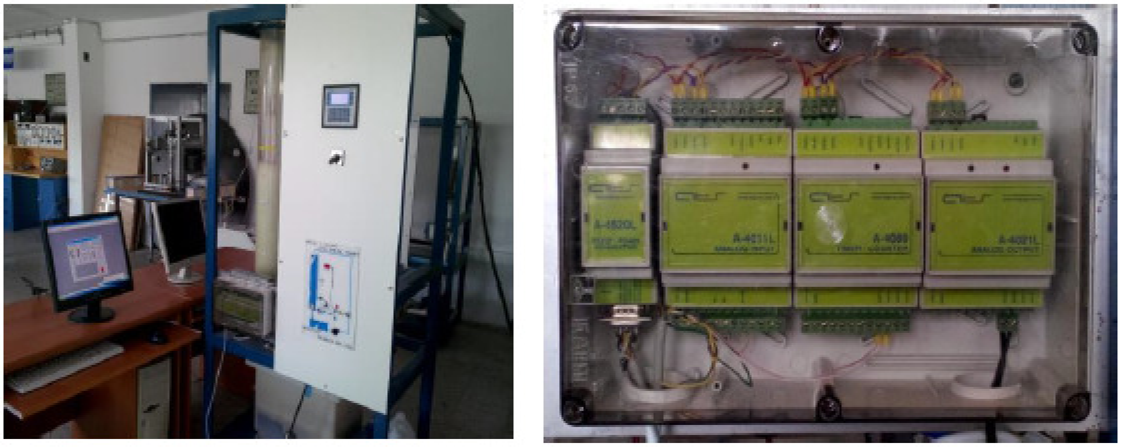


(a) (b)

Figure 3. (a) MATLAB-Simulink model of the liquid level control system with PID controller and (b) Step response of the system

4. APPLICATION

The liquid level set whose experimental setup has been constituted on a prototype and its DAQs are shown in Fig. 4.



(a) (b)

Figure 4. (a) The liquid level control system with computer-controlled and (b) DAQs

The communication between the DAQ cards and the SCADA software on the PC is done by ATS A-4520L DAQ RS232-RS484 converter The RS232 communication protocol is employed in the data communication from the PC to the DAQ cards. Besides, the RS485 communication protocol is used to transfer the data from the DAQ cards to the PC. The DAQ cards are fed by Meanweal power supply of 24 V DC and 100 W. The computer-based PID control algorithm is carried out over the DAQ SCADA program developed by the authors.

The liquid level in the tank is sensed through Foxboro differential pressure transmitter. The transmitter has a measurement range of 0-6.6 kPa. The high pressure input of the transmitter P_H is connected to the bottom side of the cylindrical tank ($P_H = P_{atm} + P_{liquid}$). Also, the low pressure input of the transmitter P_L is left open to the atmospheric pressure ($P_L = P_{atm}$). Since both the top of the tank and the low pressure input of the transmitter are left open to the atmospheric pressure, the transmitter output changes depending on the height of the liquid level in the tank ($P_{difference} = P_H - P_L = P_{liquid}$). Water is used as the process variable to be controlled in the system. The specific gravity of water is about 998.2 kg/m^3 at 1 atm and 20°C. The laboratory temperature, in which there exists the experimental set, is about 20°C.

5. SOFTWARE OF THE DAQ-SCADA SYSTEM

The software of the SCADA system with the DAQ card has been realized by the Advantech VisiDAQ software which is a Windows-based data acquisition, control, analysis and presentation development package. The feedback control algorithm with PID controller is executed on this software developed for the computer-based control. The block diagram of the developed software is given in Fig. 5. In the diagram, while the level information from the A-4011L card is taken by the AI1 block, the output command is sent by the AO1 block to the proportional vane over the A-4021L card The data from the flow transmitter, which is only utilized in monitoring the flow rate of the liquid, is taken by the CTFQ1 block. NTCL1, SPIN1, SPIN2 and SPIN3 tags are used to enter the values of set point, K_p, K_i and K_d, respectively. The PID1 block includes the PID control algorithm which is required to obtain the control signals. The output position of the system is regulated by PRG2 and PRG3 user programs. Preview image of the DAQ-SCADA window which appears on the computer screen when running the software is shown in Fig. 6

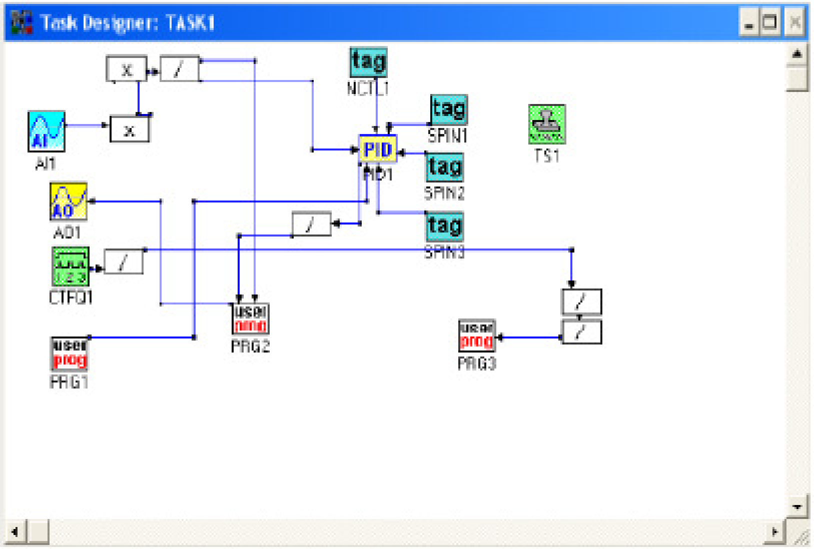
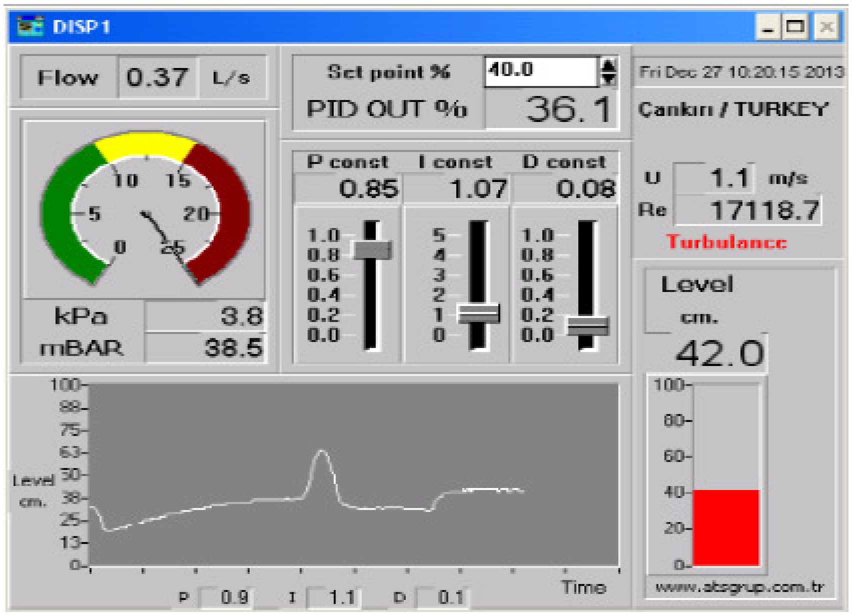


Figure 5. The block diagram of the DAQ-SCADA software



Flow 0.37 L/s

Set point % 40.0

PID OUT % 36.1

Fri Dec 27 10:20:15 2013

Cankırı / TURKEY

P const 0.85 I const 1.07 D const 0.08

kPa 3.8

mBAR 38.5

U 1.1 m/s

Re 17118.7

Turbulance

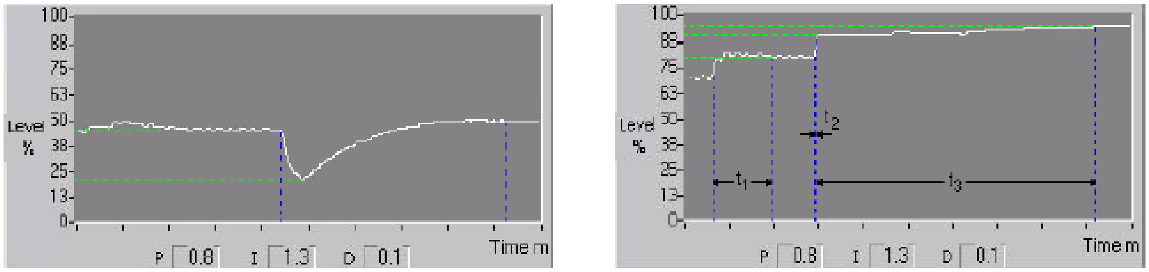
Level cm. 42.0

P 0.9 I 1.1 D 0.1

Figure 6. Preview image of the DAQ-SCADA window

6. Results and Discussion

A SCADA system with DAQ card developed by authors has been used in control of the system, as a computer-based controller. In the system, the PID controller gain parameters have been determined by fine tuning to the gain parameters, which have been previously obtained by the Z-N method, on the SCADA screen given in Fig. 6. According to the mentioned procedure, the tuned optimum controller gains have been obtained as K_p = 0.80, Ki = 1.27 and K_d = 0.11 for the set value of 40% of the process variable. On the SCADA screen, the chart at the bottom of the window shows the change of the process variable versus time and this change is also monitored by a bar graph which is located at the right side of the chart. The vertical and horizontal axes on the chart are scaled as percent (%) and minute, respectively. The maximum deviation of the process variable from the set value, which occurs before the system comes to the steady state, has been measured as 15% in negative direction. The system has reached the steady state at 4 min and the steady state error has been obtained as 0%. The transient response to the parameter changes of the system has been tested over the set point, the output load and the input load. Fig. 7 shows the transient responses to these changes of the liquid level control system which is controlled by the computer-based SCADA system.



(a)

(b)

Figure 7. Transient responses for $K_p = 0.80$, $K_i = 1.27$ and $K_d = 0.11$.

For the case that the output load is abruptly dropped while the system is working with the last steady state conditions in last case, the measured transient response is illustrated in Fig. 7a. In that case, a deviation of 23% has happened in the process variable and this has been corrected at about 5 min. The steady state error has been obtained as 0%.

In Fig. 7b, the transient response is illustrated for the cases that the set value of the system is changed from 70% to 80%, from 80% to 90% and from 90% to 95%, consecutively. In the case of the change from 70% to 80% of the set value, the transient response of the system is obtained as about t_1 = 1.20 min and the steady state error has been 2%. When the set value is changed from 80% to 90%, the process variable has come the steady state in a very short time of about t_2 = 5 sec along with the zero steady state error. The system has reached the steady state in a long time about t_3 = 6.20 min along with the zero steady state error when the set value is changed from 90% to 95%.

7. Conclusion and Future Perspective

In this study, modeling and application of a computer controlled liquid level tank system was executed for practical applications of conventional PID control method. The results of the experimental studies were clearly carried out the fundamental control algorithms on the liquid level process. By means of the DAQ-SCADA software containing a visual and flexible interface, the process could be controlled by the computer-based control structures and analyzed under the different operating conditions in detail.

For next studies, other control algorithms would be applied on the installed liquid level process. Then the results of these algorithms would be compared with each other as simulation results and experimental results.

REFERENCES

[1] Mostafa A. Fellani & Aboubaker M. Gabaj, (2015), PID controller design for two tanks liquid level control system using Matlab, International Journal of Electrical and Computer Engineering (IJECE), vol. 5, ISSN 2088-8708, p. 336-342.

[2] Kunal Chakraborty, Sankha Subhra Ghosh, Rahul Dev Basak &Indranil Roy, (2015), Temperature control of liquid filled tank system using advance state feedback controller, TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 14, ISSN 2302-4046, p. 288-292.

[3] Juri Belikov, Sven Nomm, Eduard Petlenkov & Kristina Vassiljeva, (2013), Application of neural networks based SANARX model for identification and control liquid level tank system, Proceedings of 12th International Conference on Machine Learning and Applications, Miami, FL, 4-7 December 2013, p. 246-251.

[4] Lei Zhao, (2013), Single tank liquid level control based on improved BP neural network PID control algorithm, Proceedings of Third International Conference on Information Science and Technology, Yangzhou, 23-25 March 2013, p. 230-232.

[5] Ravi Kumar Jatoth, Ayush Kumar Jain & T. Phanindra, (2013), Liquid level control of three tank system using hybrid GA-PSO algorithm, Proceedings of Nirma University International Conference on Engineering (NUiCONE), Ahmedabad, 28-30 November 2013, p. 1-7.

[6] Arun Kumar, Munish Vashishth & Lalit Rai, (2013), Liquid level control of coupled tank system using fractional PID controller, International Journal of Emerging Trends in Electrical and Electronics, vol. 3, ISSN 2320-9569, p. 61-64.

[7] Si-Wu He, Chaoying Liu, Zheying Song & Zengfang Wang, (2014), Real-time intelligent control of liquid level system based on MCGS and MATLAB, Proceedings of IEEE International Conference on Machine Learning and Cybernetics (ICMLC), Lanzhou, 13-16 July 2014, p. 131-136.

[8] Kristina Vassiljeva, Juri Belikov & Eduard Petlenkov, (2014), Application of genetic algorithms to neural networks based control of a liquid level tank system, Proceedings of International Joint Conference on Neural Networks (IJCNN), Beijing, 6-11 July 2014, p. 2525-2530.

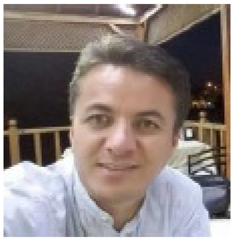
[9] Oscar Castillo & Patricia Melin, (2014), A review on interval type-2 fuzzy logic applications in intelligent control, Information Sciences, vol. 279, ISSN 0020-0255, p. 615-631.

[10] Ming Hao, Fengying Ma & QingYin, (2014), Improved fuzzy PID control algorithm applied in liquid mixture regulating system, Proceedings of International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), Shengyang, 20-22 December 2013, p. 358-361.

[11] Subramanian Saju, R. Revathi & K. Parkavi Suganya, (2014), Modeling and control of liquid level non-linear interacting and non-interacting system, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 3, ISSN 2320-3765, p. 8003-8013.

[12] Guangdong Qiu, Zhijie Luo & Guofu Zhou, (2015), Modeling and control of liquid level non-linear interacting and non-interacting system, International Journal of Smart Home, vol. 9, ISSN 1975-4094, p. 211-224.

AUTHORS



Hayati Mamur Hayati Mamur received the B.S. degree from Department of Electrical Education, Gazi University, in 1996, the M.S. and Ph.D. degrees from Department of Electronics and Computer Education, Gazi University, Ankara, Turkey, in 2005 and 2013, respectively. He is working as an Asst. Prof. at the Department of Electrical and Electronics Engineering, Faculty of Engineering, Manisa Celal Bayar University, Manisa, Turkey. His current research interests include micro wind turbines, permanent magnet generators, convertors and thermoelectric modules.

Ismail Atacak received the B.S., M.S. and Ph.D. degrees from Department of Electronics and Computer Education, Gazi University, in 1994, 1998, and 2005, respectively. From 2007 to 2012, he worked as an Assistance Professor at the Department of Electronics and Computer Education, Faculty of Technical Education, Ankara, Turkey. He is currently working as an Assistance Professor at the Department of Computer Engineering, Faculty of Engineering, Ankara, Turkey. His research interests include power systems, artificial intelligent based algorithms, optimization based algorithms and Engineering Education.

Fatih Korkmaz was born in Kırıkkale, Turkey, in 1977. He received the B.S., M.S., and Ph.D. degrees in electrical education, from University of Gazi, Ankara, Turkey, respectively in 2000, 2004 and 2011. Since 2012, he is working as an Asst. Prof. Dr. at the Department of Electrical and Electronics Engineering, Faculty of Engineering, Cankiri Karatekin University, Cankiri, Turkey. His current research field includes Electric Machines Drives and Control Systems.

Mohammad Ruhul Amin Bhuiyan received the B.Sc., M.Sc. degrees in Applied Physics and Electronic Engineering from Rajshahi University, Bangladesh in 1994 and 1995, respectively. Ph.D. degree in Applied Physics, Electronics and Communication Engineering from Islamic University, Bangladesh in 2008. Now at present he is working as a visiting scientist at the Department of Electrical and Electronics Engineering, Faculty of Engineering, Manisa Celal Bayar University, Manisa, Turkey.

DEVELOPMENT OF A LOCATION-BASED APPROACHING NOTIFICATION SYSTEM USING ANDROID PLATFORM

Hisham AlMajed and Abdelaziz Khamis

Department of Computer,
Arab East Colleges, Riyadh, Saudi Arabia

ABSTRACT

Mobile application uses and development is a rapidly growing sector. Nowadays mobile devices are more powerful and portable with plenty of useful tools for assisting people handle daily life. The main objective of this paper is to develop a mobile application that solves the problems facing bus drivers and parents when parents do not show up, and when kids wait for a long time. The application also produces the current drivers map to be used for bus fleet management purposes. The application makes use of the location service on Android to specify the current location of the driver, and the Google's cloud to device messaging to push approaching notifications to parents. The application is developed using an Extreme Programming (XP) based methodology that performs the analysis, design, implementation, and testing iteratively.

KEYWORDS

Notification; Mobile application development; Location-based services; Android studio; Android SDK; Software engineering; Extreme Programming.

1. INTRODUCTION

Since the beginning of the so-called smart-phones and their associated mobile software applications or apps, users could experience the functionality of personal computers on pocket-sized devices. These apps are becoming increasingly everywhere in our daily life [1]. Currently, categories of mobile applications include: games, banking, travel, social networking, location based apps, fitness, and medical apps.

The target audiences of many applications have shifted from the use of traditional personal computers to using mobile devices such as smart phones for performing the tasks they want or obtaining the information they seek [2]. This shift has motivated the software engineering community to provide guidance for many issues that are related to the development of mobile applications.

Different mobile operating systems are provided by different mobile companies. Mobile application developers have the choice of developing apps for each platform or a platform

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 107– 119, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70211

independent app [3]. The process of choosing between developing platform-dependent and platform-independent applications involves many parameters, such as budget, project timeframe, target audience and app functionality.

Numerous development tools have been created to aid developers in building mobile applications. These tools may be classified into two categories. The first category of tools is for platform-dependent app development, while the second category is for cross-platform app development. In [4], the authors introduced a framework for evaluating cross-platform mobile application development tools.

A mobile application is nothing but a software product with a different level of complexity. A SDLC is required to develop high quality software products that meet or exceed customer expectations, and reach completion within times and cost estimates. The phases of the SDLC include: analysis, design, implementation, and testing. Each phase is itself composed of a series of steps, which rely on techniques that produce deliverables [5].

There are many development methodologies that can be used to implement the SDLC. These methodologies vary in terms of the progression that is followed through the phases of the SDLC. Methodology options include waterfall and agile development [6].

In the Saudi community, bus drivers and kids' parents are facing problems when kids do not show up, and when parents wait for a long time. In this paper, a mobile application has been developed to solve these problems. The state of the art technology in mobile application development has been used. In addition, the location-based serves and the cloud to device messaging have been investigated to make use of their latest technology in the developed application.

This paper is structured as follows. Section 2 provides the objectives of this project, and presents the project methodology used. The literature review is introduced in section 3. The adopted SDLC methodology is given in section 4. The analysis phase is presented in section 5. The design phase is presented in section 6. The implementation and testing phases are presented in sections 7 and 8. Finally, the conclusion is given in section 9.

2. PROJECT METHOD

2.1. Project Questions

The main project questions of this paper are:

- What are the tools that best support Android platform development?
- What is the best fit methodology that can be used for mobile application development?
- What is the latest technology that can be used for cloud to device messaging?
- What are the latest location-based services that can be used in mobile application development?

2.2. Development Approach

In the next section, we start with a literature study to gather information from different resources to answer the above mentioned project questions. To guarantee the integrity of the information gathered, the literature study will depend on information that will be collected from published academic literature and industry whitepapers.

Using the build methodology, it will be possible to explore the application of the state of the art tools and methodology in mobile application development. The following good practices will be considered:

- Reuse components. Component-based development has been successful in many application domains.

- Test, Test, and Retest. Waiting until after an application has been implemented to uncover any deficiencies can be costly, and time-consuming. To minimize these kinds of problems, user interfaces must be continually tested and refined as development proceeds.

3. BACKGROUND

3.1. Mobile Application Development

There are many software engineering issues that are related to the development of mobile applications [7]. In this section, we will focus only on four issues namely, portability, development tools, and development process.

3.1.1. Portability

An important issue to consider by the mobile application developer is portability, since there are a lot of different platforms available nowadays including: Apple iOS, Google Android and Microsoft Windows.

There are various types of mobile applications: native applications, mobile web applications, and hybrid applications [8, 9, and 10]. A native mobile application is built specifically for a particular platform, using tools provided by the operating system vender.

The main benefit of a native application is performance; native applications deal directly with the mobile operating system, and make full use of all the functionality that modern mobile devices have to offer [11]. The most critical limitation of native applications is portability; code written for one mobile platform cannot run on another. That is, native mobile applications are not potable.

A mobile web application is developed using web technologies such as HTML, CSS, and JavaScript, and accessed through the mobile device's web browser. Such a browser is in itself a native app that has direct access to the OS APIs, but these APIs are only partially available to web apps or not available at all [11].

A hybrid mobile application is a blended application that combines native development with web technology. A significant portion of such an application is developed in web technology. The native portion enables the hybrid application to make full use of all the features of mobile devices.

Developer can choose between coding the native part of the hybrid app or use ready-made solutions such as PhoneGap that provide a uniform JavaScript interface to selected device features that is consistent across different platforms [12].

3.1.2. Development Tools

Development tools allow developers to write, test and deploy applications into the target platform environment. Each platform has its own development tools. For example, Android Studio IDE is the official tool for the android platform [13]. Android Studio is a powerful open source integrated development environment that provides great features for Android developers.

Before any work can begin on the development of an Android application, the first step is to configure a computer system to act as the development platform. This involves a number of steps consisting of installing the Java Development Kit (JDK) and the Android Studio Integrated Development Environment (IDE) which also includes the Android Software Development Kit (SDK) [13].

Once an Android application has been developed, an Android emulator environment is required to perform a test run of the application. Android Virtual Devices (AVDs) are essentially emulators that allow Android applications to be tested without the necessity to install the application on a physical Android based device [13].

3.1.3. Development Process

The first and foremost reason for low quality mobile application development is that app developers are not conforming to the development life cycle phases [5]. Other reasons include: lack of experience on the app development SDKs, and not enough testing is done. Therefore, app developers need to use a software development life cycle to develop high quality applications.

Choosing a methodology to use by application developers is not a simple task, because no one methodology is always best. The methodology selection criteria include: clarity of user requirements, familiarity with technology, application complexity, application reliability, short time schedules, and schedule visibility [6].

Various existing SDLC methodologies have been adapted to mobile application development [14]. Suitability and contribution of some effective and commonly used agile methods has been discussed [15]. An appropriate agile method could be selected for a given project and can be tailored to a specific requirement based upon project complexity, and time schedule [16].

3.2. Location-Based Services

Location-Based Services (LBSs) are services that make use of the geographic location of an object. LBSs have two major types: Position Aware Services and Location Tracking Services

[17]. The first type of LBSs aims to provide some useful information within the environment around the location of an object (i.e. Restaurants, Banks, Parks, etc.). It answers questions like: Where am I? And where is the nearest point of interest? [18].

The second type of LBSs aims to provide the user's location to another party to track his/her locations (For example, Parcel shipment). These services rely on others' locations and keep tracking it to provide the location information to the recipient. These services answer questions like: Where is my object of interest? And how do I get there? A survey on location based services, which includes their history and generations, exists in [19].

3.3. Cloud to Device Messaging

Many of the mobile applications rely on remote services on the cloud. Google Cloud Messaging (GCM) is one of such services which allow developers to send push messages to Android devices. Here, GCM acts as proxy server in between the android client and server [20].

Understanding the performance of GCM is essential for time sensitive applications, such as fire alert and instant messaging. An evaluation of GCM shows that the GCM message arrival latency is unpredictable. That is, a reliable connection to the Google's GCM servers does not guarantee a timely message arrival [21].

Recently, Google announced that it has acquired Firebase, a backend service that helps developers build real time mobile applications for iOS, Android and the web [22]. Using firebase cloud messaging (FCM), you can send notification messages to drive user reengagement and retention [23].

4. THE ADOPTED SDLC METHODOLOGY

Agile Development methods are considered to have the potential to help deliver enhanced speed and quality for mobile application development [24]. These methods focus on the iterative development of applications. They break the SDLC into smaller iterations to reduce the risk and allow the development to adapt rapid modification. Examples of Agile Development methodologies include Scrum, and extreme programming (XP).

In this paper, an XP-based methodology will be adopted. The main features of this methodology include the close interactions with end users and continuous testing. After a shallow planning process, the developer performs the analysis, design, implementation, and testing phases iteratively and incrementally, as shown in figure 1. Iterations used in the proposed system are: client-server connectivity, maps drawing, and notification messaging.

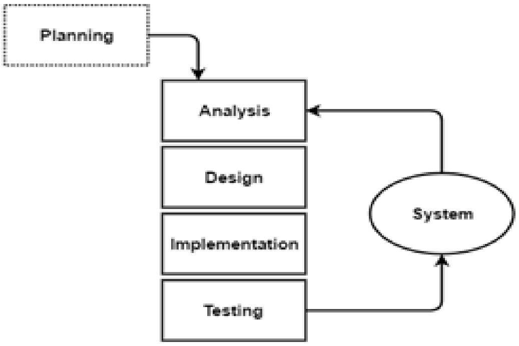
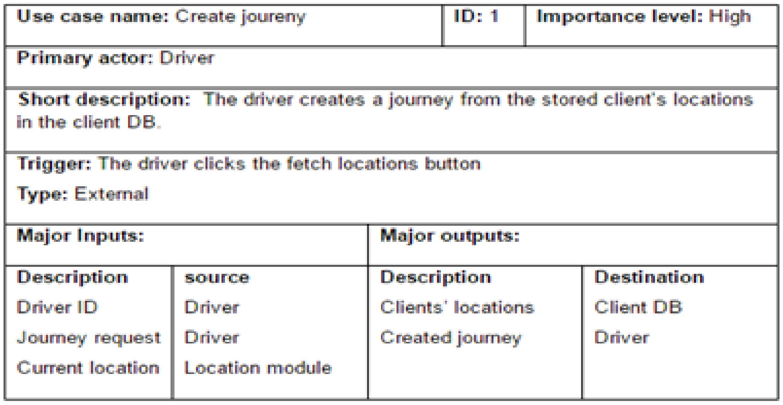


Figure 1. Extreme Programming Methodology

5. ANALYSIS PHASE

The purpose of the analysis phase is to express what the system should do by drawing process models and data models. For interactive applications, the first activity in the analysis phase is developing use cases as a means of expressing user requirements.

In this paper we present one of the use cases required for the "client-server connectivity" iteration that aims to create the application connectivity with the database. The selected use case, as shown in figure 2, creates a journey for the bus's driver to all kids' locations. The primary actor is the driver who launches the application and press the journey creation button.



Use case name: Create joureny		**ID:** 1	**Importance level:** High
Primary actor: Driver			
Short description: The driver creates a journey from the stored client's locations in the client DB.			
Trigger: The driver clicks the fetch locations button **Type:** External			
Major Inputs:		**Major outputs:**	
Description	**source**	**Description**	**Destination**
Driver ID	Driver	Clients' locations	Client DB
Journey request	Driver	Created journey	Driver
Current location	Location module		

Figure 2. Create Journy Use Case

The second activity in the analysis phase is process modelling as a means of describing the business processes. In this paper we present the context diagram that shows the entire system in context with its environment. In addition, we present a DFD fragment corresponding to the use case "Create journey".

The context diagram of the proposed system is shown in figure 3. It shows the entire system as just one process and shows the data flows to and from external entities.

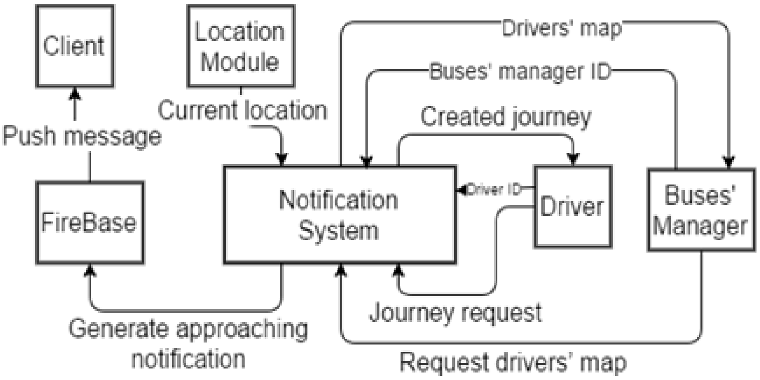


Figure 3. Notification System Context Diagram

The context diagram is decomposed to a more detailed process model called level 0 DFD that contains a fragment for each use case. The DFD fragment corresponding to the “Create journey” use case is given in figure 4.

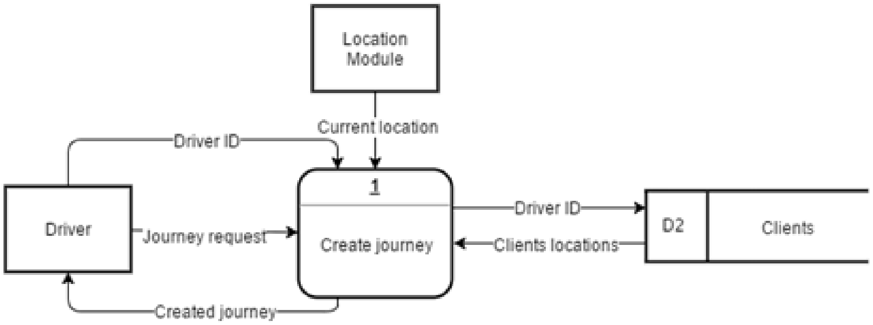


Figure 4. DFD Fragment for the “Create journey” Use Case

The third activity in the analysis phase is data modelling as a means of describing the data that are used and created by a business system. It demonstrates people, places, or things about which information is captured and how they are related to each other. The logical data model of the proposed system is shown in figure 5.

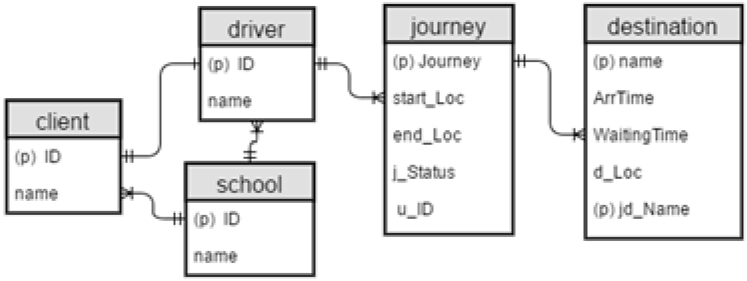


Figure 5. Notification System Logical Data Model

6. DESIGN PHASE

The design phase describes how the system will work. The deliverables of this phase include: architecture design, user interface design, and data storage design. Architecture design describes the system's hardware, software, and network environment.

In the proposed system, the client-server architecture is used, as shown in figure 6. The client is a native Android application that is responsible for the presentation logic. The server is responsible for the data access logic and data storage. PHP is used to connect the client application to a MySQL database via a Web service.

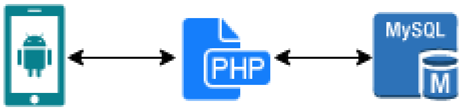


Figure 6. Notification System Architecture Design

In the proposed system, the user interface is designed to be pleasing to the eye and simple to use. It is also designed to minimize the amount of effort needed to accomplish tasks. Figure 7 shows a high fidelity prototype of the main screen of the proposed system. The Android Studio Material Design is used in designing the interface.

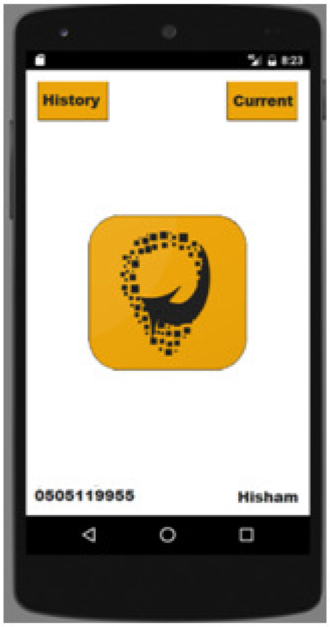


Figure 7. Notification System Main Screen

A driver clicks the button in the middle to request Clients' locations and starts the journey. A small icon will be in the notification bar to indicate that a journey is ongoing and list number of destinations left.

In the second iteration of the system development, Google maps are used to determine the best route to the clients' locations. The driver launches the Google navigator to start the journey. The locations of the clients on Google maps are shown in figure 8.

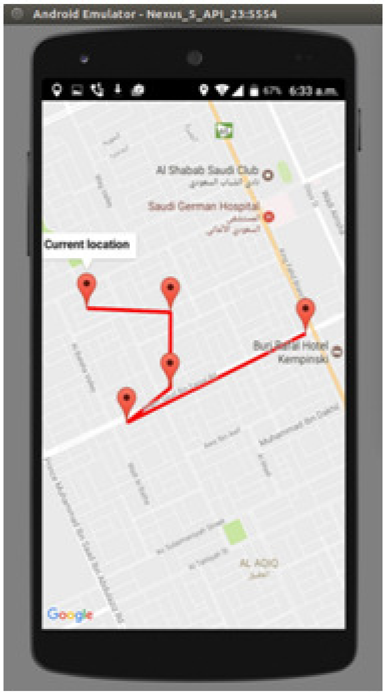


Figure 8 Clients' locations Map

Another activity of the design phase is designing the data storage component of the system. The logical data model will be converted into a physical data model. The physical data model of the proposed system is shown in figure 9.

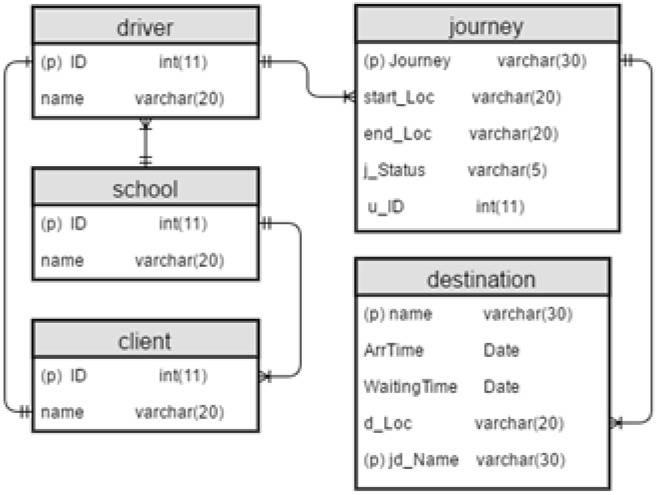


Figure 9. Notification System Physical Data Model

7. IMPLEMENTATION PHASE

The main activity of the implementation phase is writing programs to build the system. The proposed system is built using Android Studio version 2.2. For the Server side, POST and GET provided by PHP version 5.6 are used to send and receive data from the database. PHP code is hosted and accessed using the domain “approaching.halmajed.sa” defined directly in the client side to connect the application to the database. Finally, the database is constructed using MySQL version 5.7, and accessed using phpMyAdmin version 4.6.4.

For the “client-server connectivity” iteration, the Volley package is used to connect the application to the PHP server. Volley package is ready made library make Android networking easy to implement. For the “Notification messaging” iteration, Firebase API is used to notify client of arriving driver. Finally, for the “maps drawing” iteration, the Google Maps API is used to draw drivers’ locations for busses’ management department and to draw the clients’ locations for buses’ drivers.

8. TESTING PHASE

A program is not considered finished until it has passed its testing. For this reason, programming and testing are tightly coupled. Unit testing is done after an iteration programming. Integration testing is done after integrating a new iteration with the system.

Another important type of testing is the so called usability testing. Usability is a qualitative attribute that assesses how easy user interfaces are to use. This type of testing is done early in the development process. A prototype of the user interface is developed and used to get the users’ feedback. Figure 10 shows screens of the initial user interface prototype.

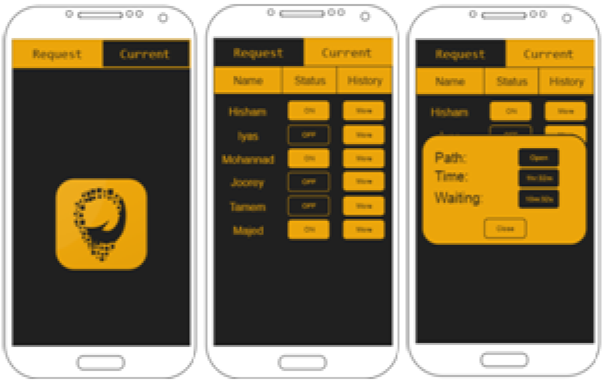


Figure 10. UI Prototype used for Usability Testing

In the usability testing three questions are considered. 1. Does the application logo indicate its purpose? 2. Is the application easy to use? 3. Is the application pleasing to the eye? After collecting answers, the overall user satisfaction is calculated. The percentage of the users who are satisfied with the initial prototype was only 40%. Therefore, the UI prototype is redesign and the usability test is repeated. This process is repeated until we got the improved results shown in figure 11. Screens of the revised user interface are shown in figure 12.

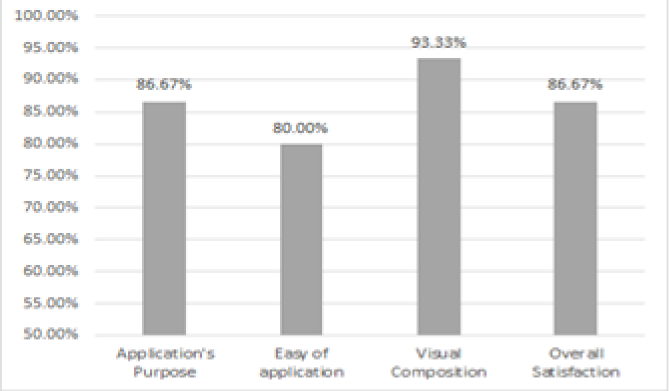


Figure 11. Results of Usability Testing

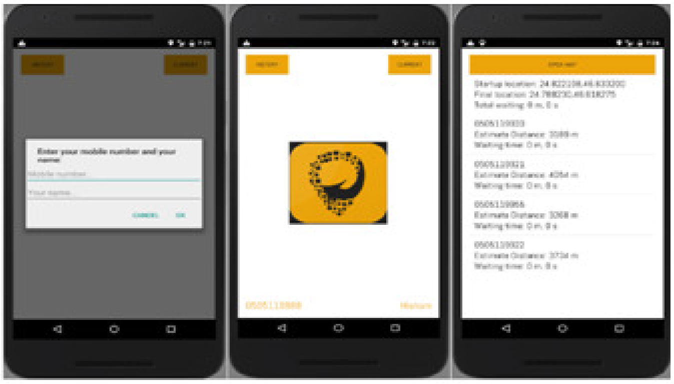


Figure 12. Revised UI Prototype

The final type of tests is done during an actual use of the application. The application is uploaded to Google play store as a beta version. Then, some selected users are invited to try the application in a real situation.

The application is integrated with Firebase to collect application logs. Figure 13 shows a statistical report produced from the Firebase console for the last 30 days. It indicates that there are 5 errors during the specified time frame, and only one user is impacted. It also indicates that these errors fall into 4 clusters (error categories).

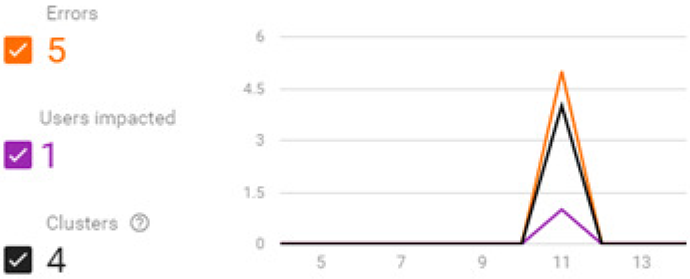
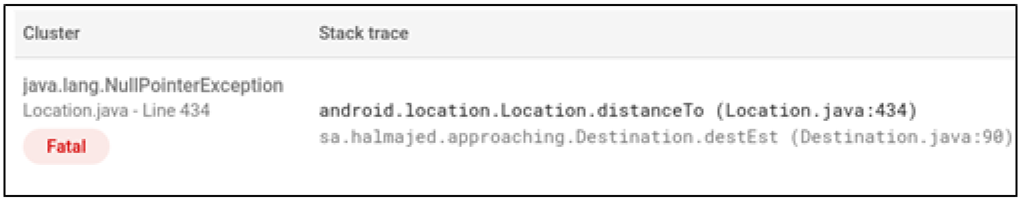


Figure 13. Application Log Report for the Last 30 Days

In addition, Firebase gives details about all error in the log file. An example of the details about an error is shown in figure 14. “NullPointer Exception” indicates that the root cause of the error is a null value is passed to the function. While “destEst (Destination.java:90)” indicates the class, function name and line number that cause the application to crash.



Cluster	Stack trace
java.lang.NullPointerException Location.java - Line 434 Fatal	android.location.Location.distanceTo (Location.java:434) sa.halmajed.approaching.Destination.destEst (Destination.java:90)

Figure 14. Error Details Example

9. CONCLUSIONS

The good practices that have been considered in our development process include reuse components and continuous testing. The reused components include third party packages and created codes. The third party packages that have been used include: Firebase package and Volley library. On the other hand, the created code for reuse include ListView adapter that is used to list the destinations of the current journey. Reuse of components reduce the error and increase the quality of mobile applications.

The second good practice is continuous testing. Impact of ignoring testing until the finishing of the application development can be frustrating, costly, and time-consuming. Therefore, during the design phase, the user interface was continually tested and refined before it is implemented. Unit and integration testing is done for each iteration of the application. Final testing is done during an actual use of the application, and Firebase console logging.

REFERENCES

[1] T. Rakestraw, R. Eunni, and R. Kasuganti, “The mobile apps industry: A Case Study”, Journal of Business Cases and Applications, Volume 9 - September 2013.

[2] L. Williamson, “A Mobile Application Development Primer: A Guide for Enterprise Teams Working on Mobile Application Projects”, IBM Whitepaper, 2012.

[3] M. Curran, N. McKelvey, K. Curran, and S. Nadarajah, “Mobile App Stores”, IGI Global, 2015.

[4] S. Dhillon and Q. Mahmoud, “An Evaluation Framework for Cross-Platform Mobile Application Development Tools”, Softw. Pract. Exper, 2015.

[5] V. Inukollu, D. Keshamoni, T. Kang, and M. Inukollu, “Factors Influencing Quality of Mobile Apps: Role of Mobile App Development Life Cycle”, International Journal of Software Engineering & Applications, Vol. 5, No. 5, September 2014.

[6] A. D. Wixom, B. Wixom, and R. Roth, “Systems Analysis and Design”, John Wiley & sons, 2012.

[7] A. Wasserman. “Software Engineering Issues for Mobile Application Development”. Proceedings of the FSE/SDP workshop on Future of software engineering research. November 7-8, 2010.

[8] A. Charland and B. Leroux. "Mobile Application Development: Web vs. Native", Communications of the ACM 54.5, 2011.

[9] S. Avinash and P. Anandkumar, "Implementation of Cross-Platform Mobile Application using PhoneGap Framework", IJCSE, Voli. 3, Issue 3, May 2014.

[10] IBM Whitepaper, "Establishing an Effective Application Strategy for your Mobile Enterprise", 2012.

[11] IBM Whitepaper, "Native, Web or Hybrid Mobile-App Developments", 2012.

[12] J. Wargo "PhoneGap Essentials Building Cross-Platform Mobile Apps", Pearson Education, 2012.

[13] N. Smyth, "Android Studio Development Essentials", eBookFrenzy, 2015..

[14] A. Kaur, and K. Kaur, "Suitability of Existing Software Development Life Cycle (SDLC) in Context of Mobile Application Development Life Cycle (MADLC)", International Journal of Computer Applications, Vol. 116, No. 19, April 2015.

[15] A. Khalid, S. Zahra, and M. Khan, "Suitability and Contribution of Agile Methods in Mobile Software Development", I. J. Modern Education and Computer Science, Vol. 2, 2014.

[16] H. Flora and S. Chande, "A Review and Analysis on Mobile Application Development Processes using Agile Mythologies", International Journal of Research in Computer Science, Vol. 3, No. 3, 2013.

[17] I. Junglas and R. Watson, "Location-Based Services", Communications of the ACM, Vol. 51, No.3, 2008.

[18] A. Kushwaha and V. Kushwaha. "Location Based Services using Android Mobile Operating System", International Journal of Advances in Engineering & Technology, 2011.

[19] M. Mohammadi, E. Molaei, and A. Naserasadi. "A Survey on Location Based Services and Positioning Techniques". In: International Journal of Computer Applications, Vol. 24, No.5, 2011.

[20] C. Tamilselvi and B. Kumar," Cloud to Device Messaging with Voice Notification Using GCM", Proceedings of the World Congress on Engineering and Computer Science 2015 Vol I, WCECS 2015, October 21-23, 2015.

[21] Y. Selim, B. Aydin, and M. Demirbas, "Google Cloud Messaging (GCM): An Evaluation", Symposium on Selected Areas in Communications: GC14 SAC Internet of Things, Globecom 2014.

[22] F. Lardinois, "Google Acquires Firebase To Help Developers Build Better Real-Time Apps", https://techcrunch.com/2014/10/21/google-acquires-firebase-to-help-developers-build-better-realtime-apps/, Posted Oct 21, 2014

[23] "Firebase Cloud Messaging", https://firebase.google.com/docs/cloud-messaging/

[24] F. Harleen, S. Chande, and X. Wang. "Adopting an agile approach for the development of mobile applications." International Journal of Computer Applications Vol. 94, No.17, 2014.

INTENTIONAL BLANK

RESOLVING CYCLIC AMBIGUITIES AND INCREASING ACCURACY AND RESOLUTION IN DOA ESTIMATION USING ARRAY ROTATION

AbdelhamidDjouadi[1] and NebojsaI. Jaksic[2]

[1]Nokia, 4575 Rings Rd, Dublin, Ohio, USA
[2]Colorado State University - Pueblo, 2200 Bonforte Blvd.,
Pueblo, Colorado, USA

ABSTRACT

A method to resolve cyclic ambiguities and increase the accuracy and the resolution in the direction-of-arrival (DOA) estimation using the Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT)algorithm is proposed. It is based on rotating the array and sampling the received signal at multiple positions. Using this approach, the gain in accuracy and resolution is addressed as function of the mean and variance of the DOA. Simulations results are provided as a means of verifying this analysis.

KEYWORDS

Direction Of Arrival (DOA) estimation, Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT), Total Least Square (TLS), Cyclic Ambiguities

1. INTRODUCTION

Because of the widespread use of sensor arrays, and the continuing development of their capabilities, sensor arrays have experienced an increased range of applications. One of these applications that has been given special attention is the high resolution direction-of-arrival (DOA) estimation. This estimation is mainly based on the processing of the received signal to extract the desired parameters of DOA of plane waves upon which the sensor outputs depend. The estimation of Signal Parameters by Rotational Invariance Techniques (ESPRIT) is one of the many approaches that have been used for implementing the DOA functions[1].In essence, and for the narrowband direction-of-arrival case, ESPRIT algorithm estimates a unitary diagonal matrix Φ with diagonal elements, given by $\varphi_i = e^{-j\mu_i}; i = 1,...d$, where d is the number of sources impinging on the array from distinct locations $\theta_1, \theta_2, ..., \theta_d$, as shown in Figure 1, and where the parameters $\mu_i\ ; i = 1,..., d$, denote the phase shifts to be estimated. For a uniform linear array (ULA), it is well known that the phase shift μ is related to the angle of arrival θ by

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 121– 138, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70212

$$\mu = 2\pi D \sin\theta / \lambda \tag{1}$$

where λ is the wavelength of the narrow-band signal and D is the displacement vector between two sub-arrays.

Various algorithms were proposed successively to improve the ESPRIT performance. However, much of the studies show that the performance of these algorithms depends only on the number of snapshots(n), the number of array sensors(m), and the signal-to-noise (SNR) of the received array signals [2]-[5].

Further in [6]-[8], it was shown that the performance is improved by using multi-scale method, where the short $(D \leq \lambda/2)$ and long baselines $(D > \lambda/2)$ are, respectively, utilized to derive the coarse unambiguous and fine ambiguous estimates for each signal. The final DOA estimate in this case is determined by using the coarse estimate to disambiguate the fine estimate.

This work shows that the performance of the ESPRIT algorithm is also constrained by the angle of arrivals of incidence with which the sources are impinging the array, and that the ESPRIT algorithm produces ambiguities in the estimated DOAs when the antenna element spacing on the linear array has a measurement error [9]-[10], or when the sources are impinging on the array with an angle of incidence along-side the ULA, especially for low SNR and limited number of snapshots.

The purpose of this work is to show that by rotating the array and sampling the received signal at multiple positions, ambiguities in the DOA estimates are resolved and also more accurate DOA estimates are obtained. Using this approach, the gain in accuracy and resolution are addressed as function of the mean and variance of the angle of arrival errors of the sources. An analysis of simulation results is provided to verify the method described.

2. PROBLEM FORMULATION

This section formulates the DOA estimation problem when using the ESPRIT algorithm and shows that its performance depends on the angle-of-arrival (incidence) with which the source is impinging the array. Also it describes the ambiguities produced in the estimated DOA results, when the antenna element spacing on a linear array is more than half a wavelength because of measurement errors, or when the sources are impinging on the array at an angle of incidence along-side the ULA.

2.1. Data Model

Consider an array consisting of two sub-arrays as shown inFigure1. Each sub-array consists of m elements. The two sub-arrays are assumed to be overlapping, and separated by a displacement vector D(where $D$is a fixed distance equal or smaller than half a wavelength). Assume that $d < m$ narrow-band sources impinge on the array from distinct locations $\theta_1, \theta_2, ..., \theta_d$ .For simplicity, it is assumed that the sources are narrow-band, no-coherent, coplanar, and that they are in the far field of the array. This assumption allows modeling the propagation delays between sensor elements as simple phase shifts, and thus the only parameters that characterize the sources locations are their DOAs. In this case, the output at any sensor of the array is the superposition of the individual emitter's signals- $s_1(t), s_2(t),..., s_d(t)$, weighted by the sensor response.

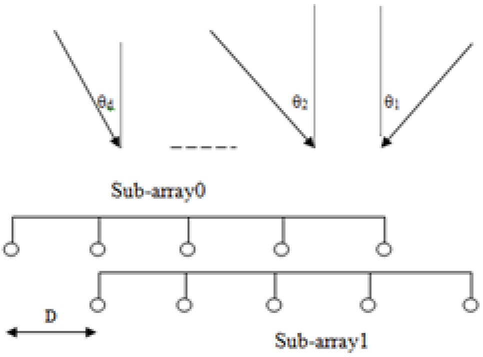


Figure 1. Array Configuration

The received signal at the k^{th} sensor can be written as

$$x_k(t)=\sum_{i=1}^{d} a_k(\theta_i)s_i(t)e^{-j\omega_0\tau_k(\theta_i)}+n_k(t) \tag{2}$$

where $\tau_k(\theta_i)$ is the propagation delay between a reference point and the k^{th} sensor for the i^{th} waveform impinging on the array from direction θ_i, $a_k(\theta_i)$ is the corresponding sensor element complex response (gain and phase) at frequency ω_0, and $n_k(t)$ is the additive noise at the k^{th} sensor. If we let $\underline{a}(\theta_i)=\left(a_1(\theta_i)e^{-j\omega_0\tau_1(\theta_i)}, a_2(\theta_i)e^{-j\omega_0\tau_2(\theta_i)},...,a_m(\theta_i)e^{-j\omega_0\tau_m(\theta_i)}\right)^T$ to be the data model representing the outputs of the m sensors of the first (reference) sub-array, then the data model representing the outputs of the m sensors is given by

$$X_0(t)=\sum_{i=1}^{d}\underline{a}(\theta_i)s_i(t)\,. \tag{3}$$

Now by letting $A(\theta)=\left(\underline{a}(\theta_1),\underline{a}(\theta_2),...,\underline{a}(\theta_d)\right)$, $S(t)=\left(s_1(t),\, s_2(t),....,\, s_d(t)\right)^T$, and

$N_0(t)=\left(n_{01}(t),\, n_{02}(t),....,\, n_{0m}(t)\right)^T$, $X_0(t)$ can be rewritten as:

$$X_0(t)=A(\theta)S(t)+N_0(t) \tag{4}$$

where $A(\theta)$ is called the direction matrix. The columns of $A(\theta)$ are elements of a set, termed the array manifold, composed of all array response vectors obtained as θ ranges over the entire space.

Likewise, the data model representing the outputs of the m sensors of the second sub-array is given by

$$X_1(t) = A(\theta)\Phi S(t) + N_1(t) \tag{5}$$

where $N_1(t) = (n_{11}(t), n_{12}(t), ..., n_{1m}(t))^T$, and where Φ is the previously defined unitary diagonal matrices with diagonal elements φ_i given by $\varphi_i = \exp\{-j2\pi D \sin\theta_i / \lambda\}$, and where λ is the wavelength of the narrow-band signal. $N_0(t)$ and $N_1(t)$ represent the uncorrelated noise present at each antenna element of the first and second sub-arrays, respectively.

In order to estimate the DOA, ESPRIT exploits the shift structure inherent in the relevant signal subspace that contains the output $Z(t)$ given by:

$$Z(t) = \begin{pmatrix} X_0(t) \\ X_1(t) \end{pmatrix} = \overline{A} S(t) + \begin{pmatrix} N_0(t) \\ N_1(t) \end{pmatrix} \tag{6}$$

where $\overline{A} = \begin{pmatrix} A(\theta) \\ A(\theta)\Phi \end{pmatrix}$.

The DOA's estimation is achieved by separating the $2m$-dimensional complex vector space C^{2m} of output snapshots into orthogonal subspaces, namely the signal subspace and the noise subspace. This is achieved by performing the eigen decomposition of the covariance matrix

$$R_{ZZ} = E(Z^*(t)Z(t)) = \overline{A} R_{ss} \overline{A}^* + \sigma^2 I \tag{7}$$

where $E(.)$ denotes expectation, R_{zz} is the covariance matrix of measurements, R_{ss} is the positive definite of the stationary (zero-mean) of the signals, and $\sigma^2 I$ is the spatial correlation matrix of the uncorrelated noise vector $N(t) = (N_0(t), N_1(t))^T$. In practice, the covariance matrix is obtained by first collecting n snapshots, $Z(t_1), Z(t_2), ..., Z(t_N)$ of the output, and then computing the sample covariance matrix as:

$$\hat{R}_{ZZ} = \frac{1}{n} \sum_{k=1}^{n} Z(t_k) Z^*(t_k) \tag{8}$$

The eigen decomposition of the positive definite and hermetian sample covariance matrix $\hat{R}_{zz}$ is given by

$$\hat{R}_{ZZ} = \sum_{i=1}^{2m} \lambda_i e_i e_i^* = E_s \Lambda_s E_s^* + \sigma^2 E_n E_n^* \tag{9}$$

where Λ_s is a diagonal matrix with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d > \lambda_{d+1} = ... = \lambda_{2m}$

$E_s = [e_1, e_2, ..., e_d]$ is the matrix composed of the eigenvectors corresponding to the d largest eigen values that define the signal subspace, and $E_n = [e_{d+1}, e_{d+2}, ..., e_{2m}]$ is the matrix of the eigenvectors that span the complement orthogonal noisesubspace. Since the matrices $E_s = \begin{pmatrix} E_0 \\ E_1 \end{pmatrix}$ and $\overline{A} = \begin{pmatrix} A \\ A\Phi \end{pmatrix}$ have the same range space, then the intent of the ESPRIT algorithm is to find a nonsingular matrix T of rank d such that

$$\begin{pmatrix} E_0 \\ E_1 \end{pmatrix} = \begin{pmatrix} A \\ A\Phi \end{pmatrix} T \tag{10}$$

By eliminating A in (10), the following expression is obtained:

$$E_1 = E_0 T^{-1} \Phi T = E_0 \Psi \tag{11}$$

where $\Psi = T^{-1}\Phi T$. In general, there is no matrix Ψ that satisfies (11) exactly because E_0 and E_1 are equally noisy and their estimates do not span the same column subspace.

The conventional ESPRIT estimates Ψ using the Least Square (LS) criterion, and thus mayyield in overall inferior results as the LS solution is known to be biased [11]. To circumvent the conventional ESPRIT algorithm drawback to some extent, the Total Least Square (TLS)-ESPRIT [11] is used instead to provide asymptotically unbiased and efficient estimates of the DOAs. The TLS estimates have been shown to be strongly consistent (convergeswith *probability one* to the true values). Following [11], a total least square (TLS) estimates of Ψ, given E_0 and E_1,is provided by

$$\Psi_{TLS} = -E_{12} E_{22}^{-1} \tag{12}$$

respectively, where E_{12} and E_{22} are implicitly defined by the eigen decomposition of

$$E_{xy}^* E_{xy} \overset{def}{=} \begin{bmatrix} E_0^* \\ E_1^* \end{bmatrix} [E_0 \mid E_1] = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} L \begin{bmatrix} E_{11}^* & E_{12}^* \\ E_{21}^* & E_{22}^* \end{bmatrix} \tag{13}$$

where $L = diag\,(l_1, l_2, ..., l_d)$, $l_1 > l_2 > ... > l_d$.

In practical situations where only a finite number of noisy measurements are available, we have only an estimates of Ψ_{TLS} of Ψ. In this case, the eigenvalues of Ψ_{TLS}, denoted by $\hat{\phi}_i$,$i=1,...,d$, give estimates of the DOAs as:

$$\hat{\theta}_i = \sin^{-1}(\lambda\,\hat{\mu}_i/2\pi D),\ i=1,...,d. \tag{14}$$

where $\hat{\mu}_i = \arg(\hat{\phi}_i)$.

2.2. ESPRIT Performance

As mentioned earlier, the TLS-ESPRIT algorithm[11] leads to unbiased estimates phase shifts under the assumption that a very large number of data snapshots are available, a condition difficult to satisfy in practical situations. When the collected data snapshots are limited, the estimates do not always approach the asymptotic ones, and the realizable results vary as function of the estimation error variance of the phase shifts.

In general, if we consider that the estimated phase shifts $\hat{\mu}$ of μ are determined with an estimation error $\Delta\mu$, and with the assumption that $\Delta\mu$ is a Gaussian independent process with zero mean and with variance $E(\Delta\mu^2)=\sigma_\mu^2$, then from (1), the estimates of the DOAs $\hat{\theta}$ of θ are estimated with an error of

$$\Delta\theta = \frac{\lambda(\Delta\mu)}{2\pi D\cos\theta} \tag{15}$$

Thus, $\Delta\theta$ is also Gaussian with zero mean and variance

$$E(\Delta\theta^2) = \frac{\lambda^2\sigma_\mu^2}{4\pi^2 D^2\cos^2\theta} \tag{16}$$

One can clearly see that, as the absolute value of incidence angle θ increases, the variance of the error $\Delta\theta$ will also increase. For instance, consider the case of only one source impinging on the array with θ_1 corresponding to a phase shift μ_1. If this same source is impinging on the array with θ_2, it will correspond to a phase shift μ_2. With the assumption that $|\theta_2|>|\theta_1|$, and since $\sigma_{u_1}=\sigma_{\mu_2}=\sigma_u$, the forms of the Gaussian probability density functions of the estimated phase shifts with means μ_1 and μ_2 are shown in Figure 2. Likewise the Gaussian probability density functions of the corresponding estimated DOAs with mean θ_1, and θ_2, are shown in Figure 3. If we choose the intervals $\mu_1 \pm z_{\beta/2}\sigma_\mu$ and $\mu_2 \pm z_{\beta/2}\sigma_\mu$,where $z_{\beta/2}$ is the z value that locates an area of $(1-\beta/2)$ in the upper tail of a normal distribution, then we are certain that these intervals will contain estimates of μ_1 and μ_2 with a probability equal to $(1-\beta)$ or a confidence level $\Im=(1-\beta)100\%$. However, we point out that while maintaining a constant confidence interval for the estimates of the phase shifts, the confidence interval of the estimates of the DOAs will increase with the angle of incidence. This is depicted in Figure 3, where the confidence interval for the estimates of θ_2 is clearly wider than the confidence of the estimates of θ_1 since the error of the estimates of θ_2 has larger sample variance than the error of the estimates of θ_1 .

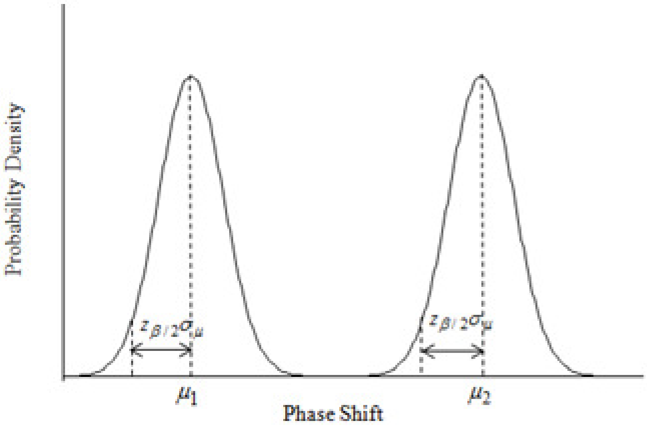


Figure2. Probability Density Functions of the Phase Shifts

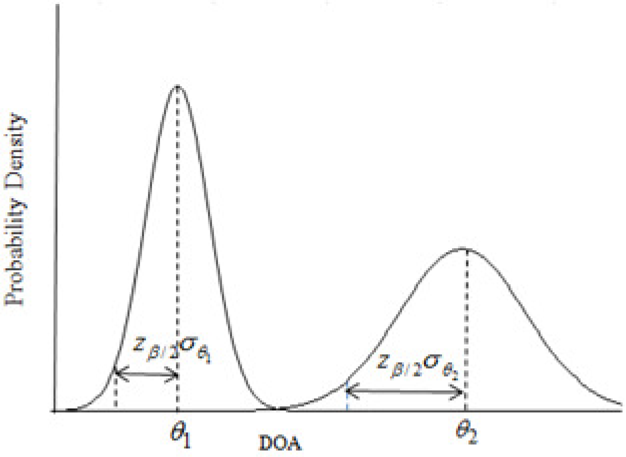


Figure 3. Probability Density Functions for the DOAs

It is obvious that (16) cannot be made arbitrary small, especially for angles of incidence alongside the ULA, without putting a constraint on σ_μ . Maintaining a good accuracy of the estimated DOA simplies having a very small σ_μ . This means a higher *SNR* and/or an increase of the number of snapshots *n.* For instance, this is clearly shown for the one source case [3], where the phase shift estimation error, the *SNR*, and the number *n*of snapshots are related by

$$\sigma_\mu^2 = \frac{1}{SNR}\left(\frac{1}{m^2 n}\right). \tag{17}$$

For more than one source, the relationship between σ_{μ} and the *SNR* and number of snapshots *n* is more complicated [7]. Nevertheless due to the non-linear relationship between the phase shift and the DOA given by (1), the basic trend is that higher *SNR* and/or larger *n* are required to obtain more accurate DOA estimates or to distinguish between two or more close sources as the absolute of their angles of incidence increases. That is, ESPRIT is limited by its ability to resolve two closely spaced sources when their angles of incidence are along-side the ULA. In fact, for low *SNR* and limited number of snapshots, ESPRIT algorithm will fail in this case to determine that there are actually two sources present. It will indicate that there is only one source present.

2.3. Cyclic Ambiguities

Consider the relation, given by (1), where we have tacitly assumed that $D = f(\lambda)$ holds perfectly. In practice however, due to measurement errors, this holds only approximately. Let ΔD, represent the error on D, this will introduce an error on the phase shift as

$$\Delta\mu = \frac{2\pi\Delta D \sin\theta}{\lambda} \tag{18}$$

Because the effective angular range of the ULA has a value $-90^{o} \leq \theta \leq 90^{o}$, the phase shift has a value in the range $-\pi \leq \mu < \pi$. Since the inverse of the mapping $\mu \rightarrow e^{-ju}$ is aliased outside this range, then any error on the phase shift, given by (18), may introduce cyclic ambiguities in estimating the DOAs. For illustration purposes, suppose that $D = 0.56\lambda$ $(> \lambda/2)$ and a source impinging on the array with $\theta = 85^{o}$, that correspond to a phase shift of $\mu = 3.505$. This phase shift is cyclically equivalent to a phase shift of $\mu = -2.777$. ESPRIT will misinterpret this value and compute the estimated DOA, assuming $D = \lambda/2$,to be $\theta = -62.16^{o}$ which is very different from the true DOA at 85^{o}.

In short, ambiguous errors due to misinterpretation of cyclically ambiguous phase shift occur in ESPRIT if

$$\Delta\mu > \pi - |\mu|. \tag{19}$$

Similarly, for low *SNR* and small number of snapshots, cyclic ambiguities may occur in estimating the DOAs for sources impinging on the array at an angle of incidence along-side the ULA. For $D = \lambda/2$, consider the phase shift $\mu = \pi \sin\theta$, resulting for a θ close to 90^{o}, as shown in Figure 4 .

It is noticeable that a small perturbation on the phase shift may lead to a phase shift $\hat{\mu}_1 > \pi$. This phase shift is cyclically equivalent to the phase shift $\hat{\mu}$. Again, ESPRIT algorithm will misinterpret the estimated phase and computes the estimated DOA at a value that is a very different from the true DOA.

3. Proposed Method

In this work, a method is proposed to address the ambiguity, resolution and accuracy problems of the DOA estimation.

3.1. DOA Estimation

From Figure 5, showing the phase Shift as function of the Angle of Arrival for $D = \lambda / 2$, one can notice that if the DOA is kept within the range $-30^o \le \theta \le 30^o$, then the phase shift μ can be approximated as

$$\mu = \frac{2\pi D \sin\theta}{\lambda} \approx \frac{2\pi^2 D\theta}{180\lambda} \tag{20}$$

where θ is expressed in degrees. This implies that $\Delta\mu \approx \frac{2\pi^2 D}{180\lambda}\Delta\theta$, and thus approximately the same *SNR* and/or the same number of snapshots are needed to maintain the same accuracy of the DOAs.

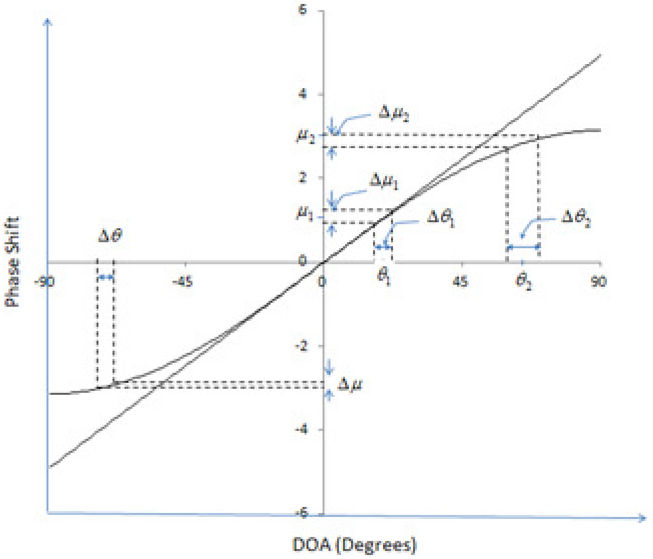


Figure 5.Phase Shift as function of the Angle of Arrival for *D*=λ/2

Thus, the main idea behind the new method is to map any estimate $\breve{\theta}$ of the DOA that iswithin the range of $-90^o \le \breve{\theta} \le -30^o$ or $30^o \le \breve{\theta} \le 90^o$ into the range of $-30^o \le \breve{\theta} \le 30^o$. This calls for the exploration of rotating the array in the elevation plane, as shown in Figure 6, and sampling the received signals at three different positions.

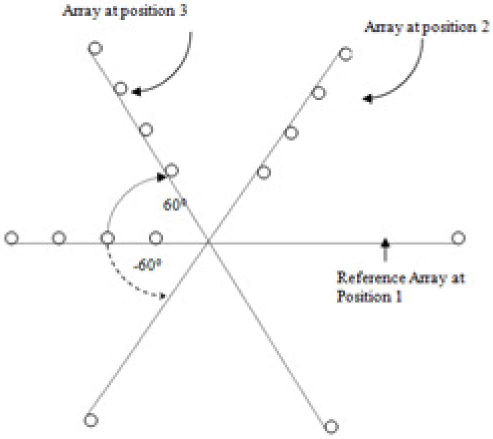


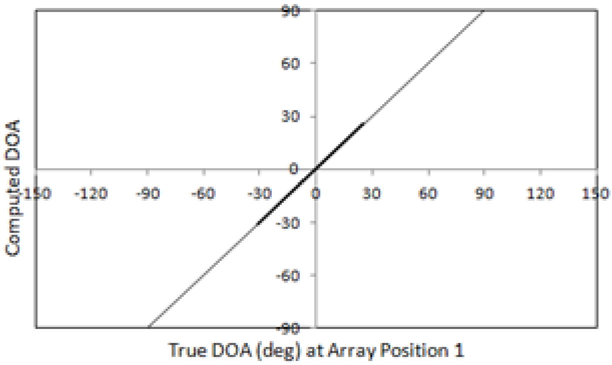
Figure 6. Array Rotation Positions

In this case, one expects to get three estimates of the DOAs at the different array positions in the range of $-90^o \le \hat{\theta}^{(1)} \le 90^o$, $-30^o \le \hat{\theta}^{(2)} \le 150^o$, and $-150^o \le \hat{\theta}^{(3)} \le 30^o$, respectively. However because the estimated phase shifts outside the range of $-\pi \le \hat{\mu} < \pi$ are aliased, any angle $\hat{\theta}^{(2)}$ in the range $90^o \le \hat{\theta}^{(2)} \le 150^o$ will appear in the range $30^o \le \hat{\theta}^{(2)} \le 90^o$, and any angle $\hat{\theta}^{(3)}$ in the range $-150^o \le \hat{\theta}^{(3)} \le -30^o$ will appear in the range $-90^o \le \hat{\theta}^{(3)} \le -30^o$, as depicted in Figure 7.Thus, by considering only the estimates of DOAs in the range of $-30^o \le \hat{\theta} \le 30^o$ at each array position, we have the following cases:

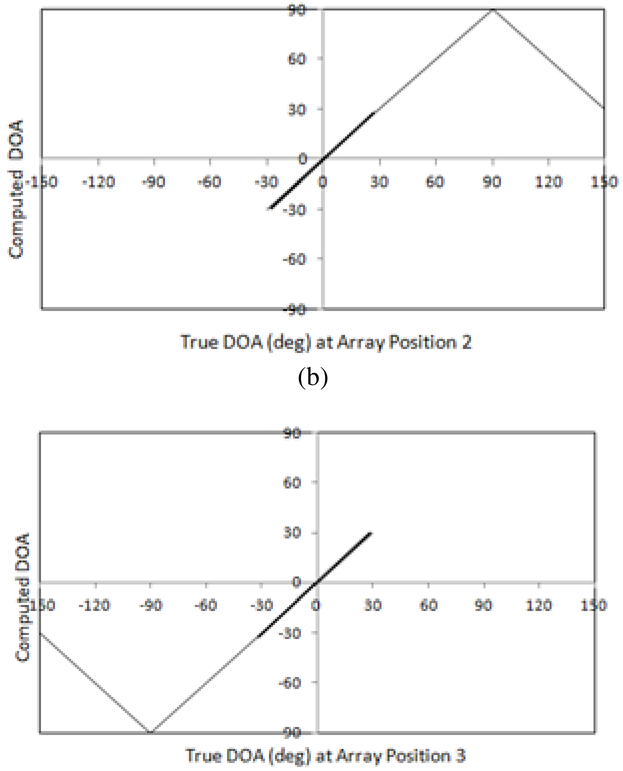
a) If $\hat{\theta}_i^{(1)}$ is within the range $-30^o \le \hat{\theta}^{(1)} \le 30^o$, the actual DOA estimate is $\hat{\theta}_i = \hat{\theta}_i^{(1)}$.

b) If $\hat{\theta}_i^{(2)}$ is within $-30^o \le \hat{\theta}^{(2)} \le 30^o$, the actual DOA estimate is $\hat{\theta}_i = \hat{\theta}_i^{(2)} - 60^o$.

c) If $\hat{\theta}_i^{(3)}$ is within $-30^o \le \hat{\theta}^{(3)} \le 30^o$, the actual DOA estimate is $\hat{\theta}_i = \hat{\theta}_i^{(3)} + 60^o$.



(a)



(b)

(c)

Figure 7. Computed DOAs as Function of the True DOAs at the Different Array Positions

To be more specific the estimation of the DOAs is performed as follows:

Step 1: Using (8), get estimates of the covariance matrices $\hat{R}_{zz_1}$, $\hat{R}_{zz_2}$, and $\hat{R}_{zz_3}$, obtained from the data collected at array positions1, 2and3, respectively.

Step 2: Compute the generalized eigen decompositions of $\{\hat{R}_{ZZ_1}, I_n\}$, $\{\hat{R}_{ZZ_2}, I_n\}$, and $\{\hat{R}_{ZZ_3}, I_n\}$, where I_n is the identity matrix.

Step 3: Using Akaike's information criterion (AIC) or the minimum description length (MDL)[12], estimate the number of sources d_1, d_2 and d_3 at each array position. Note that ESPRIT may be limited by its ability to resolve two closely spaced sources when their angles of incidence are along-side the ULA, the estimation of the number of sources may vary at the different array positions.

Step 4: Use(9)-(11) to obtain the signal spaces estimates, composed of the eigenvectors corresponding to the d_1, d_2, and d_3 largest eigenvalues of $\hat{R}_{ZZ_1}$, $\hat{R}_{zz_2}$, and $\hat{R}_{zz_3}$, respectively.

Step 5: Using (12), obtain the corresponding three mapping matrices Ψ_{1TLS}, Ψ_{2TLS} and Ψ_{3TLS} .

Step 6: Estimate the DOA's $\hat{\theta}_i^{(j)} = \sin^{-1}(\hat{\mu}_i^{(j)}\lambda / 2\pi D)$, where $\hat{\mu}_i^{(j)}; i = 1,..d_j; j = 1,2,3.$ are the phases shifts obtained from the eigen decomposition of Ψ_{1TLS}, Ψ_{2TLS} and Ψ_{3TLS}, respectively.

Step 7: By considering only the estimates of DOAs in the range of $-30^o \le \hat{\theta} \le 30^o$, classify them according the three cases specified above.

The above algorithm also resolves ambiguities in estimating the DOAs.

For $D = \lambda / 2$, if we denote by $\theta^{(1)}$, $\theta^{(2)}$, and $\theta^{(3)}$the angle of incidence at the different positions of a source impinging on the reference array with θ, and by $\mu^{(1)}$, $\mu^{(2)}$, and $\mu^{(3)}$their corresponding phases shifts, then we will have the following three cases detailed in Table.1

Table 1. DOAs and Corresponding Phase Shift at the Different Array Positions

	DOA	Phase Shift
Case 1: $-90^o \le \theta \le -30^o$	$-30^o \le \theta^{(2)} \le 30^o$ $-90^o \le \theta^{(1)} \le -30^o$ $-90^o \le \theta^{(3)} \le -30^o$	$-\pi/2 \le \mu^{(2)} \le \pi/2$ $-\pi \le \mu^{(1)} \le -\pi/2$ $-\pi \le \mu^{(3)} \le -\pi/2$
Case 2: $-30 \le \theta \le 30$	$30^o \le \theta^{(2)} \le 90^o$ $-30^o \le \theta^{(1)} \le 30^o$ $-90^o \le \theta^{(3)} \le -30^o$	$\pi/2 \le \mu^{(2)} \le \pi$ $-\pi/2 \le \mu^{(1)} \le \pi/2$ $-\pi \le \mu^{(3)} \le -\pi/2$
Case 3: $30 \le \theta \le 90$	$30^o \le \theta^{(2)} \le 90^o$ $30^o \le \theta^{(1)} \le 90^o$ $-30^o \le \theta^{(3)} \le 30^o$	$\pi/2 \le \mu^{(2)} \le \pi$ $\pi/2 \le \mu^{(1)} \le \pi$ $-\pi/2 \le \mu^{(3)} \le \pi/2$

By considering the phase shifts in the complex plane as shown in Figure 8 for the different cases, only the phase shifts that are in *quadrant 1* or *quadrant 4* should be considered to estimate the DOAs. All others should be rejected. However, one can notice that ambiguities only occur if the estimated phase shifts that are intended to be in *quadrant 2* or *quadrant 3*are misclassified to be in *quadrant 1* or *quadrant 4*. However, for this to happen, it will require a perturbation on the phase shift, given by:

$$\Delta\mu > |\mu| - \pi/2 \tag{22}$$

For illustration purposes, suppose that $\theta = 85^o$. Using the conventional case, this corresponds to a phase of $\mu = 3.1296$. From (19), any estimate of the phase shift that has a deviation of $\Delta\mu > 0.012$ from the true (noiseless) phase shift will introduce ambiguities.

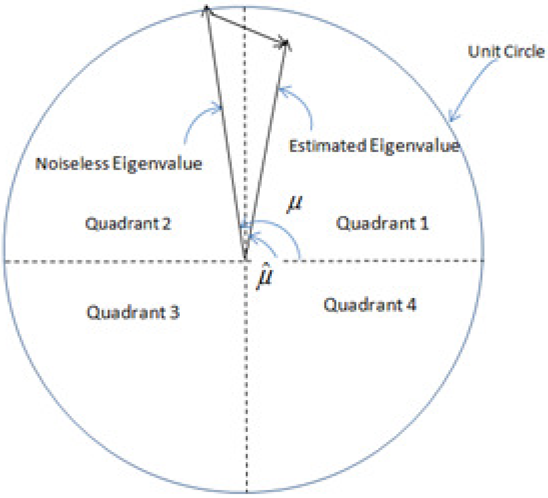


Figure 8. Error Effect on the Eigen value Estimation when the three arrays are used.

Now, using the new algorithm, a $\theta = 85^o$ will correspond to $\theta^{(1)} = 85^o$ ($\mu^{(1)} = 3.1296$), $\theta^{(2)} = 145^o$ ($\mu^{(2)} = 1.8019$), and $\theta^{(3)} = 25^o$ ($\mu^{(3)} = 1.3277$). Note that $\mu^{(2)}$ also corresponds to the aliased value $\theta'^{(2)} (= 35^o)$ of $\theta^{(2)}$.

It is clear that estimates of $\theta^{(3)}$ that are in *quadrant 1,* offset by 60^o, will be considered to be the estimates of the actual DOAs. However, since $\mu^{(2)}$ is close to $\pi/2$, ambiguities may still occurif any of its estimates that are intended to be in *quadrant 2*are misclassified to be in *quadrant 1*. In this case, the new ESPRIT algorithm will compute the estimated DOA with an offset of -60^o instead of an offset of 60^o, which leads to a value that is a very different from the true DOA.But for this to occur, and using (22), it will require a deviation of $\Delta\mu > 0.231$ from the noiseless phase shift, as compared to $\Delta\mu > 0.012$ using the conventional algorithm**.** To completethis discussion, the performances of the two cases are compared. The criterion used for this comparison is the probability *Pa* that ambiguities will occur. For the conventional case,and using (19), the probability of ambiguities can be easily shown to be given by the complementary error function

$$P_a^{(1)} = \frac{1}{2} erfc\left(\frac{\pi - |\mu|}{2\sigma_u}\right) \tag{23}$$

Now considering the new method, the probability of ambiguities is given by:

$$P_a^{(2)} = \frac{1}{2} erfc\left(\frac{|\mu^{(1)}|-\pi/2}{2\sigma_u}\right)+\frac{1}{2} erfc\left(\frac{|\mu^{(2)}|-\pi/2}{2\sigma_u}\right) \qquad (24)$$

Using the same example for one source impinging on the array with $\theta = 85^o$ and $D = \lambda/2$, $P_a^{(1)}$ and $P_a^{(2)}$ are plotted in Figure 9 for various values of the *SNR*, where σ_u is computed from (17)with *m=8* and *n=64*. The plot indicates that using the new algorithm eliminates ambiguities in this case even for low *SNR*.

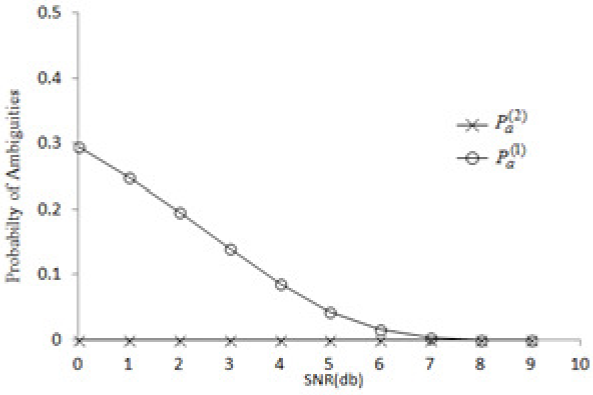


Figure 9. Probability of ambiguities

4. Simulation Results

In order to demonstrate the performance of the proposed method, the following simulations were performed. First to show the ambiguities in estimating the DOAs, *n*(=64) snapshots were generated at the different *2m* sensors with a *SNR=0dB* for one source impinging on the reference array with DOA of 85^0. The result is brought out by presenting in Figure 10, estimates of the phase shifts resulting from 10 experiments. One observation may be made at this point.When using the conventional method, out of the 10 estimates of the phase shifts, 3 of the estimates are in *quadrant 3*, resulting in an ambiguous estimation of the DOAs.

However, when considering the estimates derived from the data collected at the different positions of the array, one can notice that all the estimates of $\mu^{(3)}$ are included in *quadrant 1*.

These estimates will be used to estimate the actual DOAs. One also can notice that no ambiguities are occurring as all the estimates of phase shifts $\mu^{(1)}$ and $\mu^{(2)}$ are either within *quadrant 2*, or*quadrant3*. This is an encouraging result in when compared to the result wherethe conventional algorithm is used.

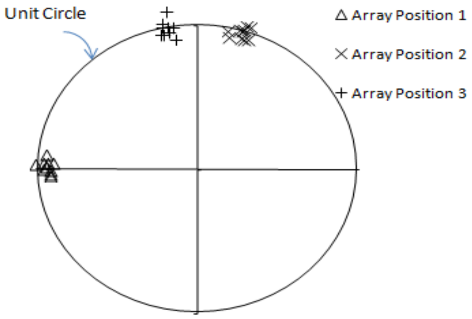
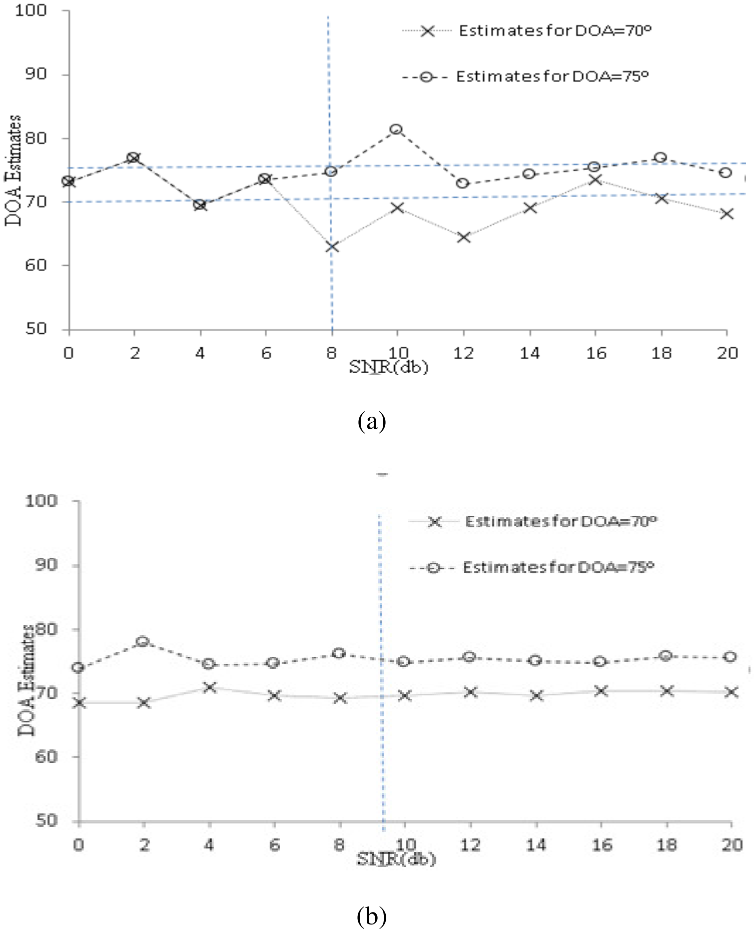


Figure 10. Estimated phase shifts obtained from data collected at the three different arrays positions

To show that for a low *SNR* and limited number of snapshots, ESPRIT algorithm fails to distinguish between two or more close sources when the angles of their incidence is along-side the ULA, two sources with DOAs of 70^o and 75^o were considered to be impinging on the array. We notice that when the conventional ESPRIT algorithm is used, it fails to distinguish between the two sources for low *SNR,* as shown in Figure 11(a). However, the new algorithm always distinguishes between the two sources even for low *SNR* as shown in Figure 11(b).



(a)

(b)

Figure 11. DOA Estimates Using the Conventional and the New Algorithms

Now, in order to demonstrate further advantages of the performance of the proposed method, the same two sources with DOAs of 70^0 and 75^0 were considered to be impinging on the array. n(=1000) snapshots were collected at each array position. The performance of new approach is brought out by examining the results illustrated in Figure 12, and Figure 13, where the bias and standard deviation are plotted respectively. The estimated DOAs were averaged over 100 experiments. To make the figure less crowded, only the results of DOA of $70^0$are shown. By rejecting any estimate outside the range of $-30^o \le \breve{\theta} \le 30^o$, only the estimates of $\theta^{(2)}$ were considered, and these estimates were offset by 60^o to get the actual estimates. To show the gain in accuracy, the estimates were compared with the ones obtained in the conventional way which corresponds in this case to the estimates of $\theta^{(1)}$. As expected, the new algorithm performed far better, especially for low *SNR*. It should be noted that below a $SNR = 8db$, the conventional ESPRIT failed to distinguish between the two sources or introduced ambiguities, and thus its estimates were excluded in this range

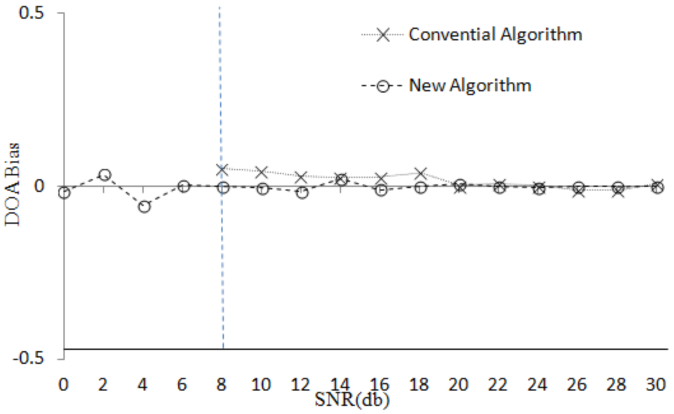


Figure 12. DOA Bias for the Conventional and the New Algorithms

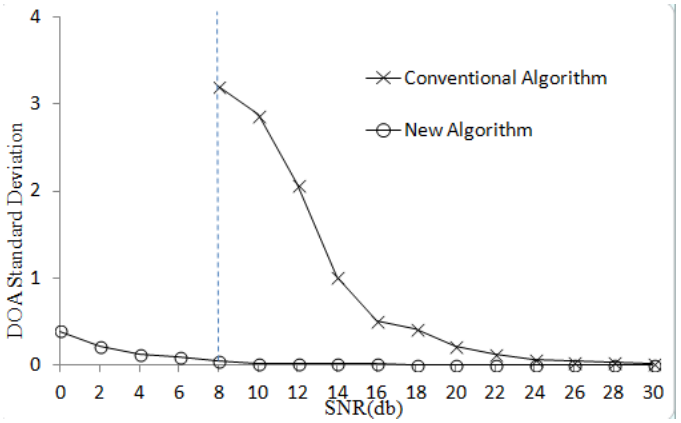


Figure 13. DOA Standard Deviation for the Conventional and the New Algorithms

5. CONCLUSION

The primary goal of this study has been to develop a procedure to resolve the cyclic ambiguities of the DOA estimates, and increase their resolution and accuracy. Although the new approach calls for the exploration of rotating the array and sampling the received data of signals at multiple positions, this disadvantage is overcome by the gain in accuracy and resolution in the DOA estimation, and the considerable improvement in reducing the ambiguities that are more likely to occur when the angle of incidence is along-side the ULA, especially for low *SNR* and small number of snapshots. Note that one can first only use the reference array (conventional method) in estimating the DOAs. If any estimates are found to be in the range $-90^o \le \hat{\theta} \le -30^o$ or $30^o \le \hat{\theta} \le 90^o$, then these estimates are considered to be coarse estimates, and the array will be rotated by 60^o and -60^o to obtain estimates at a finer scale.

REFERENCES

[1] E. Tuncer and B. Friedlander, (2009) Classical and Modern Direction-of-Arrival Estimation, Ed. Elsevier, USA.

[2] F. F. Gao and A. B. Gershman, (2005) "A generalized ESPRIT approach to direction-of-arrival estimation," IEEE Signal Processing Letters, vol. 12, no. 3, pp. 254-257.

[3] N. P. Waweru, D. B. O. Konditi, P. K. Langat, (2014) "Performance Analysis of MUSIC, Root-MUSIC and ESPRIT DOA Estimation Algorithm," International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, Vol. 8, No:1, pp. 209-216.

[4] M. H. Bhede, D. G. Ganage, S. A. Wagh, (2015) "Performance Analysis of MUSIC and Smooth MUSIC Algorithm for DOA Estimation," International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, Issue. 7, pp. 4397-4402.

[5] C. L. Srinidh, S. A. Hariprasad, (2012) "Comparative Study on Performance Analysis ofHigh Resolution Direction of Arrival Estimation Algorithms," International Journal of Advanced Research in Computer Engineering & Technology, Volume 1, Issue 4, June 2012, pp. 67-79.

[6] K.T. Wong, M. D. Zoltowski,(1998) "Direction finding with sparse rectangular dual-size spatial invariance array", IEEE Trans. Aerosp. Electron. Syst., 34, (4), pp. 1320– 1327.

[7] A. N. Lemma, A. J. van der Veen, and E. F. Deprettere, (1999) " Multiresolution ESPRIT Algorithm," IEEE Transactions on Signal Processing, Vol. 47, No. 6, June 1999, pp. 1722-1726.

[8] V. I. Vasylyshyn,(2005) "Unitary ESPRIT-based DOA estimation using sparse linear dual size spatial invariance array". Proc. European Radar Conf., Paris, France, pp. 157– 160.

[9] Tan, C. M., et al, (2002) "Ambiguity in MUSIC and ESPRIT for direction of arrival estimation," Electronics Letters, vol.38, no. 24, pp. 1598- 1600.

[10] K. Yang, Z. Zhao, X. Zhu, and Q. H. Liu, (2013) "Resolving ambiguities in DOA estimation by optimizing the element orientations," in Proc. IEEE Antennas Propag. Soc. Int. Symp. (APSURSI), Jul. 2013, pp. 1326–1327.

[11] R. Roy and T. Kailath, (1989) "ESPRIT-Estimation of Signal Parameters Via Rotational Invariance Techniques," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37. No. 7, July 1989, pp. 984-995.

[12] D. B. Williams, (1999) "Detection: Determining the number of sources," in Digital Signal Processing Handbook, V. K. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC.

AUTHORS

Abdelhamid Djouadi received the B.S. degree in electrical engineering from the University of Sciences and Technologies, Oran, Algeria, in 1980, and the M.S. and Ph.D. degree in electrical engineering from The Ohio State University, Columbus, OH, USA, in 1983 and 1987, respectively. After a successful academic career, since 1997, he has been with Nokia (former Alcatel-Lucent Technologies) as a technical lead driving improvement in 4G and 3G products. His research interests include wireless telecommunication, pattern recognition, sensor array, and signal processing.

Nebojsa I. Jaksic received the Dipl. Ing. degree in electrical engineering from Belgrade University in 1984, the M.S. in electrical engineering in 1988, the M.S. in industrial engineering in 1992, and the Ph.D. in industrial engineering in 2000 from The Ohio State University, Columbus, OH, USA. He is currently Professor at Colorado State University-Pueblo. Dr. Jaksic has published over 60 papers and holds two patents. He is a registered Professional Engineer with the State of Colorado. Dr. Jaksic is a senior member of IEEE and SME, and a member of ASEE. His research interests include robotics, automation, and nanotechnology.

SINGLE IMAGE SUPER RESOLUTION: A COMPARATIVE STUDY

Aliaa Youssef[1], Sameh Zarif[2] and Amr Ghoneim[1]

[1]Department of Computer Science, Helwan University, Helwan, Egypt
[2]Department of Information Technology, Menofia University, Menofia, Egypt

ABSTRACT

The majority of applications requiring high resolution images to derive and analyze data accurately and easily. Image super resolution is playing an effective role in those applications. Image super resolution is the process of producing high resolution image from low resolution image. In this paper, we study various image super resolution techniques with respect to the quality of results and processing time. This comparative study introduces a comparison between four algorithms of single image super-resolution. For fair comparison, the compared algorithms are tested on the same dataset and same platform to show the major advantages of one over the others.

KEYWORDS

Super resolution, Interpolation, Neighbour filling, Resizing, Low resolution

1. INTRODUCTION

Supper resolution (SR) is the process of enhancement the resolution of images. Resolution is a measure of frequency content in an image. There are always requests for good quality images from low one, although the cameras for high resolution (HR) images are expensive. Also image capturing setting is not ideal so the resulting images are blurred and noisy. Regarding that using of supper resolution techniques to enhance resolution of images and maintain the details of them is preferable [1-4].

HR images are frequently used in large applications such as satellite imaging, sports images, medical imaging, computer vision, remote sensing, surveillance systems, object detection and recognition. The need of zooming of images to analyze visual information also increases the request for super-resolution [5-7].

In general, super resolution techniques are divided into two categories, which are multi image super-resolution and single image super-resolution [2]. Multi image super-resolution is the process of generation high resolution image from multiple low resolution images. Single image super-resolution is the process of generating high-resolution image from its low resolution image [8]. This study focuses on single image super-resolution techniques.

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 139– 147, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70213

Many researchers have developed algorithms for solving super resolution issues. These algorithms could be classified into three types: interpolation-based, learning-based and reconstruction-based [9-10]. Interpolation based algorithms such as nearest neighbour interpolation, bilinear interpolation, bicubic interpolation, and lanczos interpolation are simple but the resulted image is blurred [8]. The learning-based algorithms main idea is that the lost details in (LR) images could be retrieved from a dictionary or a data set retrieved from fixed (HR) images set or website [11-13]. The reconstruction based algorithms enforce a constraint that the version of the estimated (HR) image should be consistent with its (LR) image according to predefined values [1].

The sections of paper are organized as follow. Section II presents image super resolution related techniques. Section III introduces a brief description about four compared techniques that are used in the comparison. Section IV describes the experimental results that show the advantages and disadvantages of each technique. Conclusion is presented in section V.

2. RELATED WORK

The authors in [14] proposed fast and robust multi frame super-resolution. This method based on normalization and Gaussian model. In normalization stage, the output images are generated with sharp edges. The sharp images are followed by Gaussian model to remove noise.

J. Sun et al [1] presented an image super-resolution using gradient profile prior. The gradient profile prior is learned from a huge number of natural images. It provides a constraint on image gradients when it estimates a high-resolution image from a low-resolution one. This gradient constraint helps to sharpen the details and putdown ringing along edges.

M. Bevilacqua et al [15] presented super-resolution through neighbour embedding. In this method the generation of a high-resolution image patch does not depend on only one of the nearest neighbours in the training set. Instead, it depends simultaneously on multiple nearest neighbours in a way similar to LLE for manifold learning.

The SR algorithms which depend on reconstruction-based require image patches from one or several images. This is achieved by registration and alignment of multiple LR image patches of the same scene with sub-pixel level accuracy [5], [7]. If the images haven"t insufficient patch self-similarity, these methods are not able to produce satisfying results [9]. A recent methods proposed in [17] moderate this limitation by learning image prior models via kernel principal component analysis from multiple image frames.

Another type of super resolution is learning based methods. The information is learned/observed from the training image data. Chang et al. [18] introduced the method of locally linear embedding (LLE) for super resolution dedications. Support vector regression (SVR) is proposed by Ni et al. [2] to fit the patches of low resolution image and the corresponding pixel value of the high resolution image in DCT and spatial domains. In order to achieve better SR results, one needs to carefully/manually select the training data. In such cases, the computation complexity of training and difficulty of training data selection take into account. Sparse Representation, it was first applied to SR by Yang et al. [8], [9]. They considered the image patch from HR images as a sparse representation with respect to an over-complete dictionary composed of signal-atoms. Kim and Kwon [19] proposed an example-based single image SR for learning the mapping function between the low and high resolution images by using sparse regression and natural image priors.

S. Derin et al [20] presented a novel Bayesian formulation for joint image registration and super resolution. The unknown high resolution image, motion parameters and algorithm parameters, including the noise variances, are modelled within a hierarchical Bayesian framework. The proposed framework can be extended to more general super resolution applications with more complex motion models. Y. Zhu et al [21] introduced a single image super resolution method using deformable patches. By considering each patch as a deformable field rather than a fixed vector, the patch dictionary is more expressive. This algorithm doesn"t have the ability for various of texture e.g. logo, animal, flowers.

Recently many researchers have been used sparse super-resolution algorithm for image interpolation and inpainting. Sparse super-resolution Estimators algorithm introduced a group of inverse problem estimators computed by mixing adaptively a group of linear estimators. Sparse mixing weights are calculated over a blocks of coefficients in a frame providing a sparse signal representation [22]. Computing adaptive directional image interpolations over a wavelet frame provides effective nonparametric of inverse problems. Curvelet frames and contour let frames build sparse image approximations by taking advantage of the image directional regularity. The instability of these algorithms come from their flexibility .Sparse super-resolution algorithm in [22] can be improved by Computing an adaptive signal representation in blocks. In which they are obtained as an adaptive mixing of linear Tikhonov estimators, over blocks of vectors in a frame. A fast orthogonal block matching pursuit algorithm is introduced to reduce the number of process by Applying of mixing directional interpolators over oriented blocks in a wavelet frame.

In spatial domain, multi-images SR algorithms are mostly working on aliasing artifacts that are present in LR images. The representative methods in this category include iterative back projection (IBP), Projection onto convex sets (POCS), Maximum Likelihood (ML). Iterative back projection (IBP) based method in [16] is initially a guess the HR targeted image. It is needed and then it is refined. A guess can be obtained by registering the LR images over an HR grid and then averaging them. This initial guess can be refined by using the simulated imaging model with a set of available LR observations. Then the error between the simulated LR images and the observed ones is obtained and back-projected to the coordinates of the HR image to improve the initial guess. In this method the back-projected error is the mean of the errors that each LR image causes.

3. COMPARED METHODS

This section describes four image super resolution methods. The compared methods that used in our comparative study are selected to be two states of art super resolution methods, and the other two are recently developed methods. The compared four methods are:

- Image Super-Resolution via Sparse Representation [2010] [9]
- Generative Bayesian Image Super-Resolution with Natural Image Prior [2012] [23]
- Super-resolution from Transformed Self-Exemplars [24]
- Deep Networks for Image Super-Resolution with sparse Prior [2015] [17]

3.1. Image Super-Resolution via Sparse Representation

Super resolution could be discussing from other point of view. J.Yang et al [9] introduced image super-resolution via sparse representation to generate SR based upon sparse representations by make training of coupled dictionaries from high and low resolution patch pairs. Let X is the high resolution image, Y the low resolution image, x the high resolution image patch, y the low resolution image patch, Dh dictionaries for high resolution and Dl dictionaries for low resolution. In order to recover the high resolution image (X), the method is based on two constrains. First constrain is the reconstruction constrain Y=SHX where S is down-sampling filter and H is blurring filter. Second constrain is Sparse prior, which patches (x) of the high resolution image (X) and patches (y) of low resolution image can represented like a sparse linear grouped in a dictionaries (Dh) and (Dl). The high resolution feature can be reconstructed as a combination of the sparse linear of the learned Dh dictionaries for high resolution atoms by assuming that low resolution and high resolution features share the same sparse recovered coefficients. The main disadvantage of this algorithm is the highly exhaustive computation caused by the used optimization function.

3.2. Generative Bayesian Image Super-Resolution with Natural Image Prior

In this method, initially a guess the high resolution image is needed and then it is refined. This guess can be obtained with the posterior mean, rather than the posterior mode [23] .To estimate the high resolution image, we take the advantage of sampling of high dimensional data for Gaussian Model and develop for MMSE for the HR image. In this method, Results compared with SR algorithms verify its effectiveness. This method has flexibility in using natural images priors in Bayesian model, and it use MCMC sampling based generative approach. The main drawback of this method is that it is not fast as MAP Solution.

3.3. Super-resolution from Transformed Self-Exemplars

The authors in [24] proposed an image super resolution method based on transformed self-exemplars. The algorithm produces the high resolution image by using the following steps:

1) Compute a transformation matrix T (homography) that warps target patch P to its best matching patch Q (source patch) in the down sampled image ID for each patch P in the low resolution image I. To obtain the parameters of such a transformation, we estimate a nearest neighbor field between image I and down sampled image ID using a modified Patch Match algorithm.

2) Extract the high resolution patch QH from the image I, which is the high resolution version of the source patch Q.

3) To obtain the self-exemplar PH, the inverse of the computed transformation matrix T to „unwarp‟ the high resolution patch QH, which is estimated HR version of the target patch P. After that, the algorithm paste PH in the HR image at the location corresponding to the LR patch P.

4) For all target patches, we repeat the above steps to obtain an estimation of the HR image.

5) Run the iterative back projection method to ensure that the estimated high resolution image satisfies the rebuilding constraint with the given low resolution.

Finally this method is difficulty conducting with fine details when the planes are not well detected. In addition to that, it is computationally complex due to the training procedure.

3.4. Deep Networks for Image Super-Resolution with Sparse Prior

Deep networks learning algorithms have been successfully applied in many areas of computer vision and image processing, including low-level image restoration issues. Several models depend on deep neural networks have been recently shown up for image super-resolution and gained better performance that overcome all previous invented models. We argue that domain expertise offered by the conventional sparse coding technique is still valuable, and how it can be concatenated with the key ingredients of deep networks learning to achieve further enhanced results. The method in [17] assumed that a sparse coding technique designed particularly for super-resolution could be shaped as a neural network, and trained in an ordered structure from end to end. The performance of the deep network based on sparse coding technique leads to much more efficient and effective training, as well as a reduced model size.

4. EXPERIMENTAL RESULTS

We conducted a subjective evaluation of the super resolution results for several state of art methods, for comprehensive and fair comparison between the compared methods, they have been tested on a laptop with core i5 CPU and six GB of ram. To demonstrate the strengths and drawbacks of each of them, the comparison is done on the same dataset. According to the changes that happened to the super resolution image after resizing, it is very difficult to evaluate the quality by traditional objective evaluation such as pixel to noise ratio (PSNR). So, the quality depends on the human visual perception system rather than mathematical measures. Figure 1 shows an example of resizing low resolution face and natural images by using the four state-of-art methods. First row in figure 1 shows the original images. Second row shows the super resolution output from the method in [9]. The super resolution output from method in [23] is shown in third row. Fourth row introduces the super resolution output from method in [24]. Finally, Fifth row represents the super resolution output from method in [17].

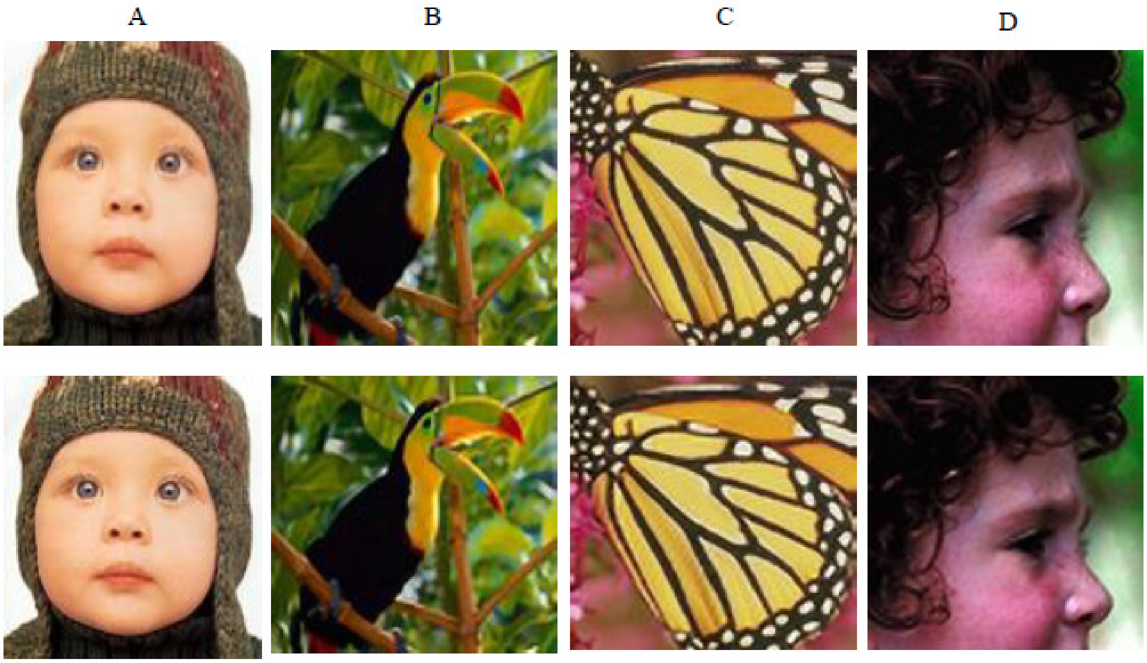
In this comparative study, we conducted a mathematical quality measure by using PSNR. Therefore, we created artificial low resolution images. Then, we applied the four state-of-art methods to reconstruct the high resolution images from the artificial low resolution images. After that, we calculated the PSNR between the original high resolution images and the reconstructed high resolution images. Table I shows PSNR comparison for the images in figure 1. Table II presents processing time comparisons between the four methods for the images in figure 1.

Table I. PSNR comparison between the four state-of-art methods.

IMAGE	SIZE	METHOD IN [9]	METHOD IN [23]	METHOD IN [24]	METHOD IN [17]
A	128X128	25.8	26.32	**28.12**	27.85
B	250X300	19.45	20.21	**23.31**	22.51
C	250X250	22.84	23.59	**25.53**	24.87
D	128X128	9.52	12.59	**13.21**	11.54

Table II. Processing time comparison between the four compared methods (in seconds).

IMAGE	SIZE	METHOD IN [9]	METHOD IN [23]	METHOD IN [24]	METHOD IN [17]
A	128X18	540	12	7.32	6.33
B	250X300	286	8.84	163.23	4.22
C	250X250	170	5.5	152.45	3.84
D	128X128	125	3.2	3.52	1.95



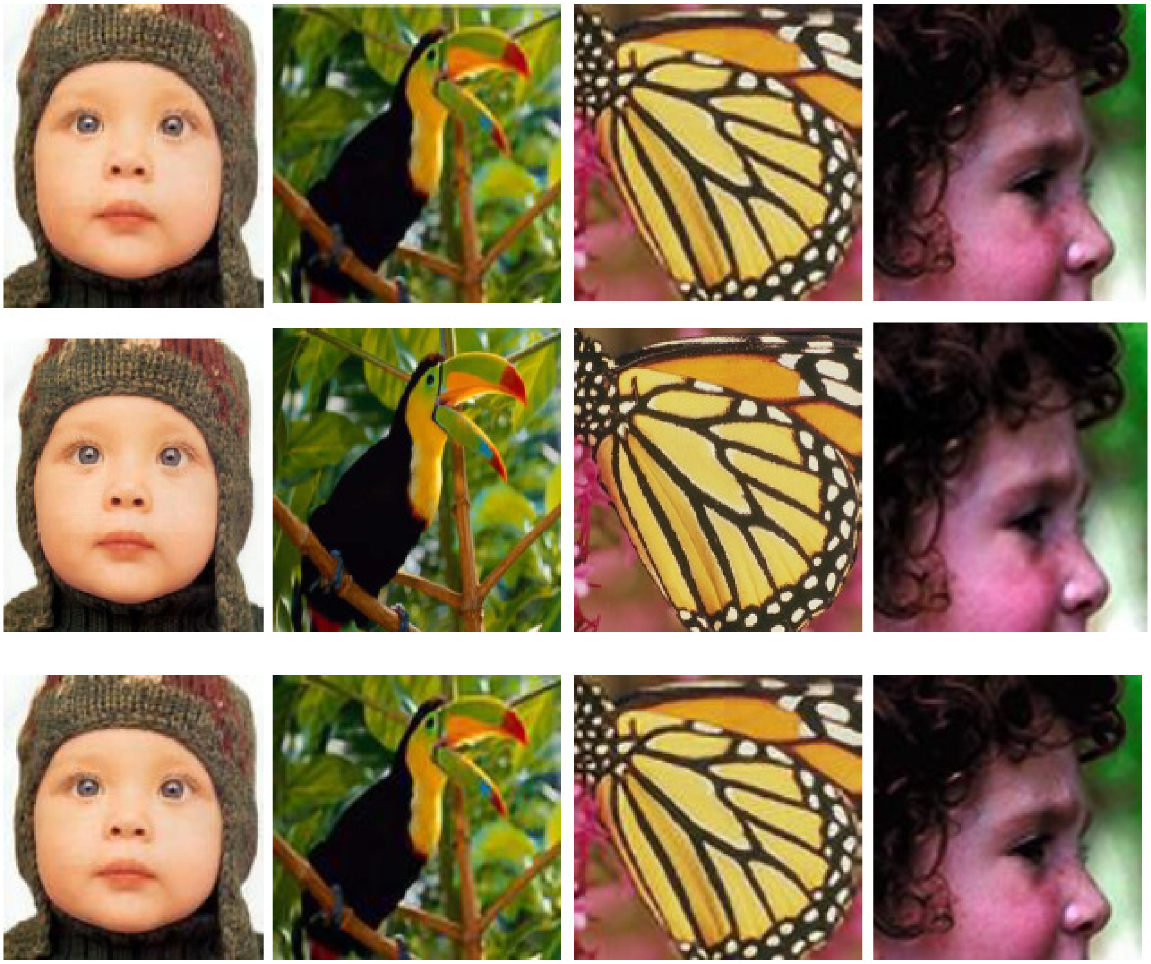


Figure. 1 Examples of recovering super resolution images. Row 1 represents the original images, row 2 represents the results of method [9], row 3 represents the results of method [23], row 4 represents the results of method [24], and row 5 represents the results of method [17].

5. CONCLUSIONS

In this research paper, we introduce image super resolution comparative study between four recent methods. We conducted to be fair as much as possible by comparing the four image super resolution methods in the same images as well as device hardware. The experimental results illustrate the strengths and drawbacks of each method. According to the PSNR table and Time table we can conclude that, Method [24] produces the higher PSNR value over the other methods. In the second table Method [17] produces the minimum Time. Therefore, we recommend them for all researchers to be applied in their super resolution applications future work.

ACKNOWLEDGEMENTS

The authors gratefully would like to thank the Pre-Master students of Faculty of Computers and Information, Menofia University, year 2016 for their efforts.

REFERENCES

[1] J. Sun, Z. Xu, and H. Shum, “Image super-resolution using gradient profile prior,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8, 2008.

[2] K. S. Ni and T. Q. Nguyen, “Image super-resolution using support vector regression,” IEEE Trans. Image Process., vol. 16, no. 6, pp. 1596–1610, 2007.

[3] J. Sun, J. Zhu, and M. F. Tappen, “Context-constrained hallucination for image super-resolution,” in Proc. IEEE Conf. Comput. Vision and Pattern Recognition, pp. 1-8, 2010.

[4] Zeyde, R., Elad, M., Protter, M “single image scale-up using Sparse-representations” Curves and Surfaces, pp. 711–730, 2012.

[5] A. Chakrabarti, A. N. Rajagopalan, and R. Chellappa, “Super-resolution of face images using kernel PCA-based prior,” IEEE Trans. Multimedia, vol. 9, no. 4, pp. 888–892, 2007.

[6] M. Protter and M. Elad, “Image sequence denoising via sparse and redundant representations,” IEEE Trans. Image Process., vol. 18, no.1, pp. 27–35, 2009.

[7] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Non-local sparse models for image restoration,” in Proc. ICCV, pp. 2272–2279, 2009.

[8] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution as sparse representation of raw image patches,” in IEEE Conf. Comput. Vision and Pattern Recognition, pp. 1-8, 2008.

[9] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861–2873, 2010.

[10] K. I. Kim and Y. Kwon, “Example-Based Learning for Single-Image Super-Resolution and jpeg Artifact Removal” Max-Planck-Institute for Biological Cybernetics, pp. 1-28, 2008.

[11] Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T. “Coupled dictionary training for image super-resolution” IEEE Transactions on Image Processing, vol. 21(8), 3467–3478, 2012.

[12] Dong, C., Loy, C.C., He, K., Tang, X. “Learning a deep convolutional network for image super-resolution” uropean Conference on Computer Vision, pp. 184–199, 2014.

[13] Yang, J., Lin, Z., Cohen, S.: Fast image super-resolution based on in-place example regression. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1059–1066 (2013).

[14] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, “Fast and Robust Multi-frame Super-resolution”, IEEE Transactions on Image Processing, vol. 13, no. 10, pp. 1327-1344,

[15] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Morel. “Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding” Proceedings of the British Machine Vision Conference (BMVC). , pp. 135.1–135.10, 2012

[16] Patel Shreyas A,”Novel Iterative Back Projection Approach ”, Journal of Computer Engineering , Volume 11, Issue 1, PP 65-69, 2013.

[17] Z. Wang, D. Liu, J. Yang, W. Han and T. Huang, "Deep Networks for Image Super-Resolution with Sparse Prior," IEEE International Conference on Computer Vision (ICCV), pp. 370-378, 2015

[18] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in Proc. IEEE Conf. Comput. Vision and Pattern Recognition, pp. 1-8, 2004.

[19] K. I. Kim and Y. Kwon, “Single-image super-resolution using sparse regression and natural image prior,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[20] S. Derin, R. Molina and A. K. Katsaggelos “Variational Bayesian Super Resolution” IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 20, NO. 4, APRIL 2011.

[21] Y. Zhu, Y. Zhang, A. L. Yuille “Single Image Super-resolution using Deformable Patches” in Proc. IEEE Conf. Comput. Vision and Pattern Recognition, pp. 1-8, 2014.

[22] S. Mallat, G Yu “Super-resolution with sparse mixing estimators”, IEEE Transactions on Image Processing, VOL. 19, NO. 11, pp. 2889 – 2900, 2010.

[23] H Zhang, Y Zhang, H Li, “Generative Bayesian image super resolution with natural image prior” IEEE Transactions on Image processing, vol. 21, no. 9, pp. 4054-4067, 2012.

[24] J. Huang, A. Singh, N. Ahuja “Single Image Super-resolution from Transformed Self-Exemplars” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197 - 5206, 2015.

AUTHORS

Sameh Zarif received his BSc and MSc degrees in information technology from Menofia University, Egypt, in 2005 and 2009 respectively. He completed his Doctor of Philosophy from centre of intelligent signal & imaging research (CISIR), Universiti Teknologi PETRONAS (UTP), Malaysia 2015. Currently he is an assistant Professor in Department of Information Technology at Menofia University Egypt. In addition to his current research into image super resolution, texture synthesis, and image completion, his interests lie in image processing, computer vision, and pattern recognition.

INTENTIONAL BLANK

CHECKING BEHAVIOURAL COMPATIBILITY IN SERVICE COMPOSITION WITH GRAPH TRANSFORMATION

Redouane Nouara and Allaoua Chaoui

Laboratory MISC University Abdel Hamid Mehri Constantine 2,
Constantine Algeria

ABSTRACT

The success of Service Oriented Architecture (SOA) largely depended on the success of automatic service composition. Dynamic service selection process should ensure full compatibility between the services involved in the composition. This compatibility must be both on static proprieties, called interface compatibility which can be easily proved and especially on behavioural compatibility that needs composability checking of basic services. In this paper, we propose (1) a formalism for modelling composite services using an extension of the Business Process (BP) modelling approach proposed by Benatallah et al. and (2) a formal verification approach of service composition. This approach uses the Graph Transformation (GT) methodology as a formal verification tool. It allows behavioural compatibility verification of two given services modelled by their BPs, used as the source graph in the GT operation. The idea consists of (1) trying to dynamically generate a graph grammar R (a set of transformation rules) whose application generates the composite service, if it exists, in this case (2) the next step consist in checking the deadlock free in the resulting composite service. To this end we propose an algorithm that we have implemented using the AGG, an algebraic graph transformation API environment under eclipse IDE.

KEYWORDS

Dynamic Service Composition, Graph Transformation, Service Composition Checking, Modeling Composite Service

1. INTRODUCTION

Service Oriented Architecture (SOA) is an ideal solution to the problems of distributed applications development, characterized by system heterogeneity and low coupling of components, since systems may not be developed by the same teams. Despite the great step made in this field by standardizing protocols of description (WSDL), discovery (UDDI), binding (SOAP) and a series of languages for manipulating services called (WS-*), all researchers and manufacturers are convinced that the success of the SOA approach is inevitably conditioned by a successful automation of dynamic service composition, in which a new service is dynamically created by assembling features of elementary services. In this case, the selection of the composed services is made on the fly. Although software vendors can guarantee the safety of their web services, the development, testing and verification of these web services are independently from other vendors' peers[1]. This raises the problem of composability of services offered by different providers.

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 149– 162, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70214

For this end, several approaches have been proposed in the literature; generally based on planning tools, semantic extensions of service protocols or formal approaches. All these approaches incorporate the behavioural aspect of the service as part of their specification, in which the service's behaviour is associated with its static interface description (specified as a WSDL document). The specification of external and observable behaviour of services is required to achieve the composition operation because having only a syntactic compatibility level in the interaction interfaces cannot by itself guarantee the success of the interaction between two services[2][3]. The crucial problem that has been raised is whether a given service, selected based on some criteria, which can be functional or non-functional, may be successfully composed with the desired service in terms of interaction interfaces; even if they are not compatible in behavioural aspects.

Checking the composability of services plays an important role in the operation of automatic composition. If the non-formal approaches of composition, based on AI planning tools, have shown their limits at the expense of purely formal approaches, characterized by their mathematical basis [4]. These approaches are therefore ideal candidates that can contribute to solve the problem of checking composability.

Among these formalisms, the GT constitute an adequate tool for solving this kind of problems, due to (1) its pure formal basis (algebraic approach) and (2) it handles graphs which are the formalism generally used for modeling service behavior. However, major approaches proposed for service composition conceal an important aspect which is the modelling of composite services. In these approaches, a global view of services is used, which don't specify the real granularity of services, because services are generally modelled as black boxes or as atomic actions, which do not describe exactly the reality of things. This results in a coarse description of the composite service and an inaccurate specification of the interaction between services. To overcome these problems we propose, in this paper, a modelling formalism (an extension of the BP model) for describing the external and observable behavior of composite services that reflect also the interaction between elementary services, and an approach for checking the behavioural composability between two services using the GT formalism in the form of an algorithm for generating the composite service if it exists.

The article is structured as follows. In section 2 we introduce the concepts and definitions of the graph transformation formalism based on the algebraic approach. A state of the art on the use of graph transformation as a tool in service composition literature is presented in Section 3. Section 4 introduces the Business Process (BP) used as formalism for modelling behavioural services in our approach. In section 5 we detail the proposed service composition checking approach. In section 6 we present the rule generation process, our algorithm and its implementation. Finally we conclude this paper and give some future works directions in Section 8.

2. GRAPH TRANSFORMATION

Graphs offer a very rich mathematical formalism for modeling because they are a natural means for expressing complex system situations on an intuitive level. They are used to model all kinds of system states and specially the behavioural aspect with this mathematical basis. Graphs may be subject to compute operations that check some behavioural properties on system models. What justifies their wide uses in the specification data, diagrams, flow control, for the entities and relationships for UML diagrams[5]. One of these tools is Graph Transformation. Its basic idea is the change of a source graph, into another, result graph by applying some transformation rule(s), similar to Chomsky grammars in formal language theory. GT is used in several areas of computing for model transformation such as modeling and specification of visual processing models according to the MDA (Model Driven Architecture) approach or describing the concurrency and distribution of systems [4]. In what follows, we present some basic definitions of the algebraic graph transformation used in our work.

2.1. Graphs and Graphs Morphisms

A labelled graph $G = (V, E, s, t, l_V, l_E)$ is a sextuplet with V a finite set of nodes (also called vertices), E a finite set of edges and two functions s and t defined by $s, t: E \to V$, which define the sources and targets of edges respectively. $l_V: V \to L_V$ and $l_E: E \to L_E$ are labelling functions that attribute a node's label from the set LV (respectively edge's label from LE with $L_V \cup L_E = L$ the labelling set. Let be two graphs G1 and G2 defined by $G_i = (V_i, E_i, s_i, t_i, l_{V_i}, l_{E_i})$ with $i \in \{1, 2\}$. A graph morphism f between G1 and G2 is $f: G_1 \to G_2$ with $f = (f_E , f_V)$ consists in two functions $f_V: V_1 \to V_2$ and $f_E: E_1 \to E_2$, that preserves the source and target functions defined by $f_V \circ s_1 = s_2 \circ f_E$ and $f_V \circ t_1 = t_2 \circ f_E$ in [4].

2.2. Algebraic Graph Transformation

The algebraic graph transformation approach is based on pushout constructions used to model the gluing of graphs. In this approach there are two main variants the Single Pushout (SPO) and Double Pushout (DPO). In the latter two gluing constructions are used, where in the first only one construction is used as depicted in Figure 1. The interested reader can find more details in[4].

The operation of transforming a given source graph G to a resulting graph H is done by applying a production rule p, defined in SPO by: $p = L \xrightarrow{r} R$ where L and R are two graphs called the left hand (LHS) and right hand (RHS) of the rule respectively, r is a morphism between L and R as illustrated in Figure 1.

The rule p is applicable if and only if there is a morphism m between the graphs L and G ($m: L \to G$) which takes the form of an image of L in G. The target graph H is constructed by adding the graph R to the graph G, and from the resulting graph the graph L is removed [4]. To prohibit the execution of a rule, some conditions can be added, called Negative Application Condition (NAC), this forbids some graph structure X to be present in the source graph G before or after applying a rule. Formulated by: A NAC(n) on L is a graph $n: L \to X$, a graph morphism $m: L \to G$ satisfies NAC(n) on L iff: $\nexists\, q: X \to G$ such that $q \circ n = m$ [6].

$$\begin{tikzcd}
L \arrow[r, "r"] \arrow[d, "m"'] & R \arrow[d] \\
G \arrow[r] & H
\end{tikzcd}$$

Figure 1. Graph Transformation Principle.

3. RELATED WORK

The use of Model Transformation (MT) in general and especially the GT in the formalization and checking of distributed architectures and service composition as a special case has got very little attention in the literature. Major works proposed in checking service composition uses other formalisms.

In [7] a structural approach is proposed, where composite service is modelled as a kind of Petri Net called Open Net. The service composition checking used done by using results of structure theory of Petri net, in which the necessary and/or sufficient structural conditions are identified for ensuring a behavioural compatibility between two services. Bentahar [1] used a model-checking based approach in order to verify if composite service design meets some desirable behavioural properties. Composite service is modelled based on a separation between two aspects, an operational behaviour illustrates the business logic that a composite service implements and a control behaviour illustrates and states the constraints which the operational behaviour should satisfy. These two behaviours are formally defined using automata-based techniques. Foster[8] propose a checking service composition approach based on verification of properties. Created from design specifications and implementation models; to confirm expected results from the viewpoints of both the designer, modelled in UML, and implementer. The result compiled into the Finite State Process notation (FSP) in order to reason about the concurrent programs. Bultan [9] propose WSAT4; a framework for analysing the interactions among composed Web services modelled as conversations (a sequence of exchanged messages). The composite web service is modelled as a set of peers (elementary services) which communicate with each other using asynchronous messages via a FIFO queue, where each peer is modelled as a state machine. In [10], the Classical Linear Logic (CLL) is used to verify the correctness of web service composition. The process consists of finding a proof for a requested service with available services stated as assumptions. If the proof is found this means a valid composition exists, and then a process calculus realisation of the composite service can be automatically extracted. Hamadi [11] propose an approach that uses Petri nets for modelling composed services, the service composition is done by a merging process of elementary services to a Petri net that models the control flow of the composite service. This approach uses an algebra that specifies different concurrent execution forms between composed services. The most similar work is that of DING [12], where an approach is proposed for the identification of structural conflicts (behavioural incompatibility) in inter-enterprise business process models. This approach is based on an algorithm that employs condition reachable matrix.

4. SERVICE MODELLING FORMALISM

The choice of the formalism used to model the service behaviour constitutes the key element in any approach for service composition. The model should describe as accurately as possible the behaviour of the service and its interaction with its environment. In what follows we present the modelling formalism used in our approach to describe service behaviours whether, elementary or composite.

4.1 Single Service Modeling

We adopt, in this work, the service model proposed by [13] [3] [14] called Roman model. This model captures conversations, the external and observable behavior that a service supports; it is defined as the ordered set of messages exchanged between the service and its client during their interaction. The Roman model uses deterministic finite state automata (DFA), in which the states represent the different phases through which the service passes during its life cycle, and transitions model the events and/or internal actions that occur during service interaction. These transitions are triggered by messages exchanged between the service and its client, which corresponds to (1) an invocation of a service method or a response to the latter, or (2) the advent of an internal event to the service as an expiration of a waiting period. The model has a single initial state and several final states, the transitions labeled by messages are associated with polarities defined by symbols +,- in [13] or ? and ! in [14] that specifies the origin of messages. Polarity + (respectively -) indicates that the message is received (respectively sent) by the service. Each BP is associated with a current state that describes the current state of the BP, initially following the invocation of the service by the client, it starts at the initial state and at each transition it changes the current state until reaching a final state which indicates the end of the interaction.

To use the graph transformation approach, we formalize the external behavior of services in this article with a graph language notation as mentioned in [4], instead of the automaton notation used in [13] [3]. In order to integrate the BP specificities in a graph model notation, we extended the graph definition by initial and final states. Let A be a BP of a service, formally we use the following definition of $A = (V, E, s, t, l_V, l_E, v_0, F)$ with V and E describe the sets of states and edges respectively, s and t are the start and target functions of edges. The sets lV and lE represent the state and edge labels respectively (with their respective polarities). v_0 the initial state with $v_0 \in V$, F the set of final states (with $F \subset V$).

As an example, Figure 2 describes the modelling of an e-commerce service that manages the order of some goods, with start as initial state and the set *{Cancel, delivery}* as final states. Labels *{login (+), confirm_order (+), payment (+), delivery (-)}* are a sequence of exchanged messages between the service and the client; while the messages depend on the polarity sign. This sequence constitutes a valid conversation, *cancel(-)* is an internal event, automatically generated by the service and sent to the client after the timeout of payment by the customer.

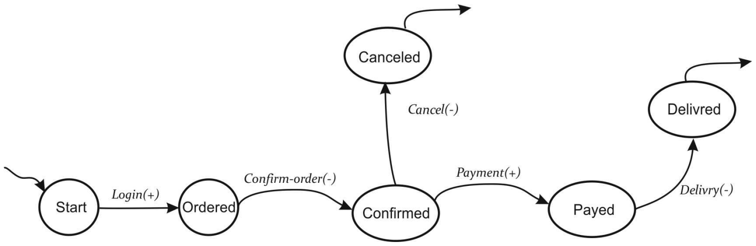


Figure 2. Sample Business Process.

4.2 Composite Service Modeling

The Roman model used to describe service external behaviour is well suited for describing single service behavior. However, it suffers from the lack of concurrence modelling between elementary services in the case of composite service, because DFA formalism has only a single current state describing the entity running at a given time; while in the case of composite services there is a set of services that run in parallel and a fully distributed manner. This drawback inherent from DFA constitutes a big handicap for modelling composite services and specifying the multiple forms of concurrence existing between elementary services. To overcome this obstacle, we propose an extension of the Roman model in order to support the specification of concurrence in a composite service modelling. We use a Multi Current State DFA, for specifying the concurrent execution between elementary services. In this model we formulate a composite service Cs as:

$$C_S = (A, S_0, I)$$

With: *A* a set of service BPs modelling elementary services, S_0 is the set of current active states and I the set of invocation edges. It specifies the execution of composed services and their life cycle progress. *A* is equal to the number of elementary services involved in the composition. When a service calls another one, initially only the caller service has its current state active (in S_0). Each time an elementary service is invoked, its current state (generally the initial state) becomes active and added to S_0. The current state dynamically changes every time the service exchanges messages with accordance to its BP until the end of service execution (reaching a final state). In this case, the current active state is disabled and removed from S_0. The set I describes the interactions between the elementary services. They model either (1) a service invocation or (2) a response from the service following an invocation by another one. Initially, the set I is empty and these invocation edges are dynamically created at each service invocation (added to I) and deleted at the end of executions service. The proposed model is a Multi Current State Automata (as many current states as elementary BPs). The composition is carried out following service invocations. Each time a service S_A invokes, from state s_a, another service S_B to state s_b with a message M. This invocation results in the creation of an invocation edge starting from the state s_a to the state s_b and labelled with message M as depicted in Figure 3.

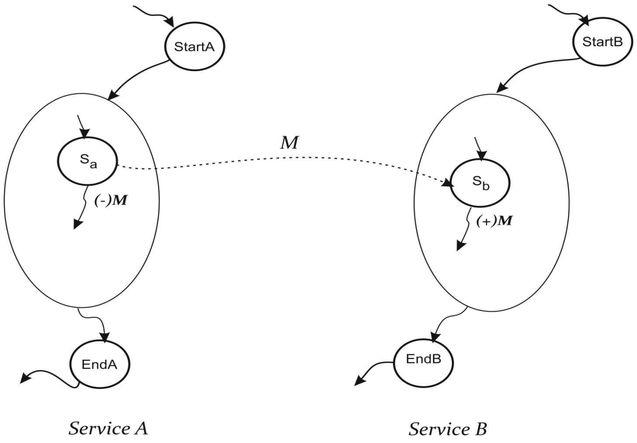


Figure 3. Composite Service Model's.

As an example, let be a service S_A described with its BP shown in Figure 4(a) that interacts with a service S_B (shown in Figure 4(b)) to create a composite service S_c. Initially:

$$C_S = \{(A = \{BP_{SA}\}, S_0 = \{StartA\}, I = \emptyset)$$

Service S_A invokes, from the state A_4 , Service S_B to state $Start_B$, the composition operation creates an invocation edge libelled with the exchanged message m_7 and added to the set I. The created edge connects the state A_4 to $start_B$ as depicted with dotted line in Figure 5. The composite service becomes:

$$C_S = \{(A = \{BP_{SA}, BP_{SB}\}, S_0 = \{A_4, StartB\}, I = \{m_7\})$$

Service S_B responds to Service S_A by sending one of two messages:

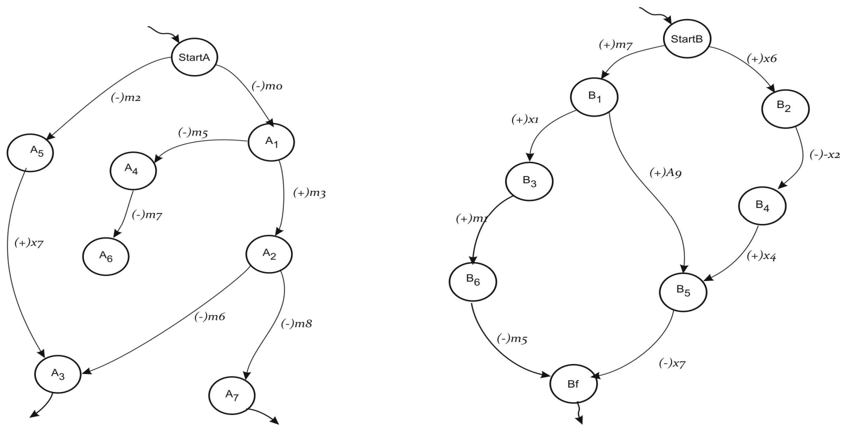
m_5 sent from B_6 to A_l which creates the edge labeled by m_5 between B_6 to A_l and the composite service becomes :

$$C_S = \{(A = \{BP_{SA}, BP_{SB}\}, S_0 = \{A_1, B_6\}, I = \{m_7, m_5\})$$

x_7 sent from B_5 to A_5 which result in :

$$C_S = \{(A = \{BP_A, BP_B\}, S_0 = \{A_5, B_5\}, I = \{m_7, x_7\})$$

After this, Service S_B goes to the final state B_f which will complete the operation of composition between the two services.



(a) BP A. (b) BP B.

Figure 4. Example of elementary services composition.

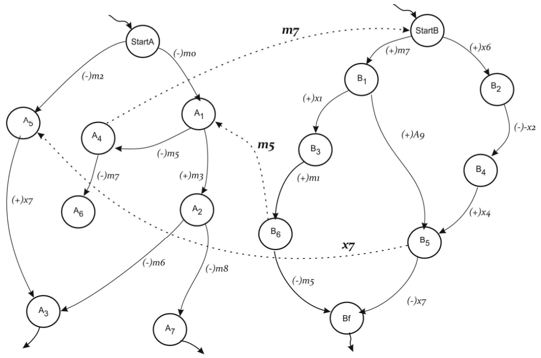


Figure 5. BP of Composite Service.

5. Services Composition Verification Approach

The proposed approach, for checking service composition, uses GT as verification formalism. Since the latter has the major advantage of having a formal process for handling graphs (either simple or typed attributed graphs) [4]. This allows formalizing the necessary conditions that must be met to conclude the success or failure of service composition. The purpose of this approach is to check whether two elementary services S_1 and S_2 modelled by their respective BPs (1) can be syntactically composed by generating a valid composite service S_c i.e. the set of invocation edges I is not null and (2) check behavioural compatibility specially the deadlock free in the composite service. The GT is used as a formal tool to merge the two graphs for giving the composite service Sc (if it exists) by automatically generating a Graph Grammar G = (P, G_0) where P is a set of transformation rules called GTS (Graph Transformation System) and G0 the start graph.

P = {p_i, $1 \leq i \leq n$} where: pi is a rule that represents the interaction between the two services that can be either:

A service invocation: In which one of the two services invoke a method of the second service or inversely a response to a previous invocation of another service.

An internal event generated by a service like timeout expiration.

The P rules have an identical structure which consists of creating an invocation edge between two states, one belonging to each service. The start graph G_0 is represented by the two graphs (S_1 and S_2).

The two services are syntactically composable if the graph grammar G exists i.e. the set P is not empty, in this case the execution of its rules on G_0 generate the composite services Sc.

The existence of Sc does not imply that the two services can be composed because some conditions must be checked before concluding the composability of the two services. In what follows, we define and formalize these necessary conditions.

5.1 Conditions of Services Composability

As mentioned in the beginning of this paper, managing services in a fully open and totally dynamic environment requires, before a service be involved in a composition process, to operate some checks that confirm a priori the success of their composability. These verifications must be done at the same time at syntactic and behavioural levels as detailed in [15] and [16]. Namely, the syntactic consist of checking the mismatches occurring in service interfaces and behavioural aspects (called mismatch in service Business Protocol). The first aspect was already discussed in literature and is not considered in this paper. The second one is to check the behavioral compatibility which can be either (1) a deadlock free of the conversations between the two services or (2) an unspecified reception of a message from the other service. In this paper, only the conversation deadlock free is covered because the unspecified reception of messages cannot be checked (1) before the runtime of service composition and (2) the BPs alone cannot guarantee that the message may be intended for another service.

5.1.1 Existence of Invocation's Message(s)

Checking the existence of exchanged messages between the two services is to verify the composability in a purely syntactic point of view, i.e. that there is at least one message issued by one service (with polarity (-) in its BP) and at the same time expected by the other service (with polarity (+)) as described in Figure 4. In this article, we do not take into account the compatibility of exchanged messages in terms of structure, i.e. the number and parameter types, nor the semantic, i.e. the interpretation of a data element's meaning or an operation's function that can be easily checked. The existence of exchanged messages can be formalized as follows, let be S1 resp S2 two services defined by $\mathrm{S_i} = (\mathrm{V_i}, \mathrm{E_i}, \mathrm{s_i}, \mathrm{t_i}, \mathrm{l_{V_i}}, \mathrm{l_{E_i}}, \mathrm{v_{0_i}}, \mathrm{F_i})$ with $\mathrm{i} \in \{1, 2\}$ and BPS1 (resp BPS2) the Business Process of S1 (resp S2). There is an exchanged message(s) between S1 and S2 if and only if:

$$\exists\ e\ \in (E_1 \cap E_2), \exists\ s_i\ \in V_1, s_k \in V_2\ where\ \begin{pmatrix} s_i.(-)e \in BP_{S_1} and\ s_k.(+)e\ \in BP_{S_2} \\ or \\ s_i.(+)e \in BP_{S_1} and\ s_k.(-)e\ \in BP_{S_2} \end{pmatrix}$$

5.1.2 Deadlock-Free Conversations

As mentioned at the beginning of this paper, syntactic compatibility does not conclude the composability of two services, a second condition must also be checked, it relates to the behavioural compatibility between the two services. This compatibility consists of verifying the deadlock-free between the two services conversations. This situation is characterized by the case where each service is in a waiting situation for reception of a message sent by the other service.

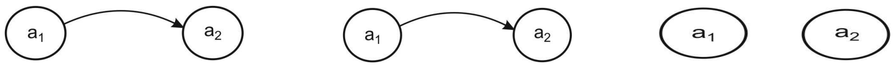
As shown in figure 5, service S1 (stay in state A1) is awaiting receipt of message m5 from service S2 and at the same time S2 (In state SartB) expects the message m7 coming from S1.

To formalize the deadlock-free we define the function *Poid(x)*, with x a BP state, it returns the set of states reachable from x. The set I of invocation messages between S1 and S2 defined by I = $\{m_i(a,b), 1 \leq i \leq n\}$, with a ∈V1 and b ∈ V2 or inversely where $m_i(a, b)$ is a message exchanged between two services (sent from State a to State b). We conclude to deadlock free between $S_1$and S_2 if and only if:

$$\exists\, s_i, s_j \in V_1, \exists\, x_k, x_p \in V_2 \; where \begin{pmatrix} s_j \in Poid(s_i) \wedge x_p \in Poid(x_k) \\ and \\ \exists m_i(x_p, s_i) \in I \wedge \exists m_j(s_j, x_k) \in I \end{pmatrix}$$

5.2 Rules generation

The generated transformation rule set p_i, if it exists, has the same structure, as depicted in Figure 6, and characterized by the facts (1) the left side (LHS) is constituted by two states a_1 and a_2 one belonging to each service (see Figure 6(b)), (2) the right side (RHS) is constituted by the states (a_1 and a_2) connected by an edge (Figure 6(c)) and libelled with the exchanged invocation message. The application of p_i results in creating the edge between a_1 and a_2. In order to avoid an indefinite execution of the grammar rules, and impose a single execution, we add for each rule a NAC which is the RHS of the rule as depicted in Figure 6(a). (3) Grammar rules are not subject to any execution order and therefore can be executed in a random order. In the next section, we present our proposed algorithm for the automatic generation of the grammar whose application creates the composite service if it exists.



(a) NAC (b) LHS (c) RHS

Figure 6. Structure of Generated rules.

6. ALGORITHM FOR CHECKING APPROACH

The BP model used for formalizing the external service behaviour as automata-based graph presents an interesting feature that be an oriented and rooted graph i.e. it has (1) a special single state called root (in our case the BP initial state), from which all other graph states are reachable, and (2) the output edges of each state are bounded by a constant number because BPs are deterministic finite automata. This feature has a major advantage that allows the development of algorithms for processing BPs whose complexity is not exponential i.e. the execution time is limited as proved by [17] and [18] [19]. Based on this, in our approach we propose an algorithm for automatically generating the composite service grammar. The algorithm calls a weight function *Poid* that returns, for a given node, the set of nodes reachable from this node (see Algorithm 1). Essentially based on recursive functions, the algorithm operating principle consist, in the first step of browsing the states of the first graph, starting from the start state, and at each time it searches the existence of an invocation message between this state and another one belonging to the second service, which satisfies the existence of invocation message condition (cited above in Section 5.1). If an invocation message exists, the two identified states are converted to a transformation rule as explained in the previous section and added to Set I . In the case where the invocation message list is empty we conclude to a syntactic incompatibility between services. Otherwise in the second step it executes the generated rules to create invocation edges between services. Finally, it checks the deadlock-free between the two conversations by checking Equation 2, if this condition is meet; it concludes to a behavioural incompatibility and therefore the two services are composable otherwise the two services can be composed.

***Algorithm** 1:Function Poid: compute the set of nodes reachable from a given node.*
***Require**: v a graph node.*
***Ensure**: the list of reachable nodes from v*

```
if Terminal(v) then
    P oid ← {v}
else
    k ← nodecount(v)
    for j = 1 → k do
        P oid ← P oid ∪P oid(nextnode(v; j ))
    end for
end if
return P oid
END.
```

Algorithm 2: Check the composability of two services based on their BPs.
Require: $BP_1 = (V_s, E_s, s_s, t_s)$
Require: $BP2 = (V_t, E_t, s_t, t_t)$ Services.
Ensure: *R* Set of rules.

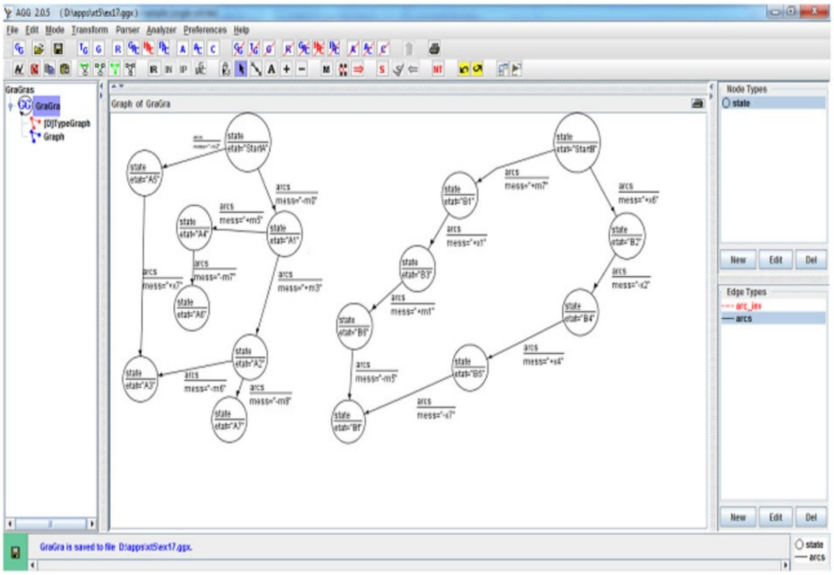
```
$V_s = \{v_s^0, .., v_s^n\}$ and $E_s = \{e_s^0, .., e_s^m\}$
$V_t = \{v_t^0, .., v_t^k\}$ and $E_t = \{e_t^0, .., e_t^l\}$
sta ← start_state(BP 1)
stb ← start_state(BP 2)
compute();
search_event(sta; stb);
run_rules();
if Nbrule = 0 then
    print("SY NTACIC INCOMPAT IBILIY BETWEEN SERVICES")
else
    DEADLOCK ← false
    list_inv_arcs ← get_liste_invocation;
    for k = 1 to size(liste_inv_arcs) do
        arc1 ← liste_inv_arcs(k)
        source_arc1 ← get_source(arc1);
        dest_arc1 ← get_destination(arc1);
        for j = k + 1 to size(liste_inv_arcs) do
            arc2 ← liste_inv_arcs(j );
            source_arc2 ← get_source(arc2);
            dest_arc2 ← get_destination(arc2);
            if source_arc1 ∈ Poids(dest_arc2) and
                source_arc2 ∈ Poids(dest_arc1) then
                DEADLOCK ← true;
            end if
        end for
    end for
    if DEADLOCK then
        PRINT("deadlock between the two bps —> behavioral incompatibility")
    else
        PRINT("behavioral compatibility —> the two services are compatible")
    end if
end if
END.
```

6.1 Algorithm Complexity

The algorithm has as input the two BPs A and B that are rooted graph, suppose that A have n nodes, and the number of edges connected to each node is bounded by an integer k. The graph B has m nodes each one bounded by e edges. The algorithm browse the first BP from the initial sate, and for each edge it check the existence of an invocation edge with a node belonging to the second BP. This operation is done in m^e instructions. With n node in the first graph, the algorithm need: $n^{k(m^e)}$ instructions, in worst case, to achieve the execution.

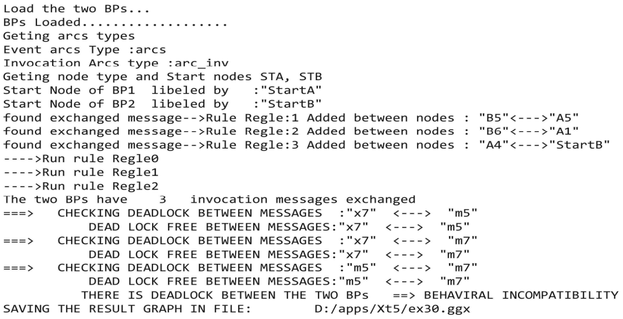
7. IMPLEMENTATION

The above-mentioned algorithm has been implemented using the Eclispe java IDE and the API AGG (see Homepage http://user.cs.tu-berlin.de/˜gragra/agg/) for graph transformation. This choice is guided by the AGG features that provide a set of necessary functions for dynamically manipulating components of GT. AGG is a general development environment for algebraic graph transformation systems which follows the interpretative approach. It allows the dynamic creation of all GT components (typed graphs, graphs, rules, NAC, nodes, edges), and to dynamically manipulate them by adding or removing operations. Its special power comes from a very flexible attribution concept and graphs are allowed to be attributed by any kind of Java objects [20]. These features of dynamically managing the GT and automatic execution of grammar have guided our choice to using the AGG API. As an example, the two services shown in Figure 4 whose corresponding graph represented with the AGG framework (a GUI environment) shown by the screenshot in Figure 7. This graph introduced to our application as input generates an output on the console (see Figure 8) that describes the different steps done during the execution. In which three invocation messages are founded as detailed in Section 4.2 and depicted in dotted line in the resulting generated composite service graph shown in Figure 9. After processing, two invocation edges have a deadlock situation which proves behavioural incompatibility as a result.



GraGra is saved to file D:\apps\xt5\ex17.ggx.

Figure 7. Screen Capture of Input BPs.



```
Load the two BPs...
BPs Loaded..................
Geting arcs types
Event arcs Type :arcs
Invocation Arcs type :arc_inv
Geting node type and Start nodes STA, STB
Start Node of BP1  libeled by   :"StartA"
Start Node of BP2  libeled by   :"StartB"
found exchanged message-->Rule Regle:1 Added between nodes : "B5"<--->"A5"
found exchanged message-->Rule Regle:2 Added between nodes : "B6"<--->"A1"
found exchanged message-->Rule Regle:3 Added between nodes : "A4"<--->"StartB"
---->Run rule Regle0
---->Run rule Regle1
---->Run rule Regle2
The two BPs have    3   invocation messages exchanged
===>   CHECKING DEADLOCK BETWEEN MESSAGES  :"x7"  <--->  "m5"
           DEAD LOCK FREE BETWEEN MESSAGES:"x7"  <--->  "m5"
===>   CHECKING DEADLOCK BETWEEN MESSAGES  :"x7"  <--->  "m7"
           DEAD LOCK FREE BETWEEN MESSAGES:"x7"  <--->  "m7"
===>   CHECKING DEADLOCK BETWEEN MESSAGES  :"m5"  <--->  "m7"
           DEAD LOCK FREE BETWEEN MESSAGES:"m5"  <--->  "m7"
          THERE IS DEADLOCK BETWEEN THE TWO BPs   ==> BEHAVIRAL INCOMPATIBILITY
SAVING THE RESULT GRAPH IN FILE:        D:/apps/Xt5/ex30.ggx
```

Figure 8. Output Execution.

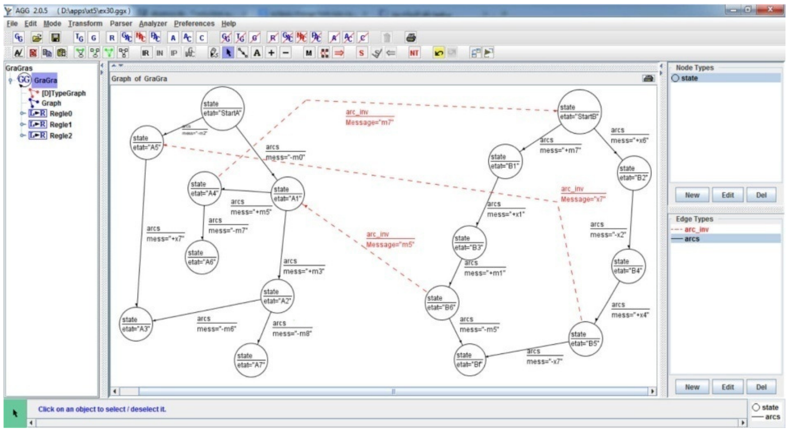


Figure 9. Screen Capture of Resulting Composite Service BP.

8. Conclusion and Future Work Directions

Dynamic services composition is a big challenge facing the success of SOA approach for which several tools have been proposed in literature. Among these tools, we find that formal based methods are the most promising. The choice of formal methods for specifying and dynamically checking service composition is justified by the need to have mathematical based tools, which guarantees the success of these operations. In this context, this paper, explored the possibility of using graph transformation as a tool for service composition checking. Services are modelled by their BPs; a formalism that specifies the external and observable behaviour of services, which is vital in the process of composition. The approach realises the checking composition by an automatic generation of production rules that controls the generation of composite service BP. We have proposed (1) an extension of BP for modelling composite service behaviour (2) a formalisation of necessary and sufficient conditions to check the composability of services (3) and an algorithm for checking services composition that we have implemented with the AGG API. As future work we expect (1) experiment the algorithm on real cases to optimise its complexity (2) extend the BP model to support the specification of service interfaces in order to describe service composition in a more realistic way (3) the use of model transformation tools to translate service BP model to a textual formalism specification such as Lotos.

REFERENCES

[1] Bentahar, J.; Yahyaoui, H.; Kova, M. ; Maamar, Z. ’Symbolic model checking composite Web services using operational and control behaviors’ Expert Systems with Applications. Vol. 40 (2013) pp 508-522.

[2] Papazoglou, M.P., and Georgakopoulos, D. ’Service-Oriented Computing. Communications of the ACM’, 46(10):25-28

[3] Benatallah, B., Casati, F., and Toumani, F. Web ’Service Conversation Modeling ACornerstone for E-Business Automation’ IEEE Internet Computing 2004.

[4] Ehrig, H., Ehrig, K., Prange, U., and Taentzer, G. ’Fundamentals of Algebraic Graph Transformation’ (Monographs in Theoretical Computer Science. An EATCS Series) Springer Verlag.

[5] Elboussasidi, G. ’Développement logiciel par transformation de modèles’, Thèse de doctorat Université de Montréal 2009.

[6] Habel, A., Heckel, R., and Taentzer, G. ’Graph grammars with negative application conditions’, Fundamenta Informaticae vol 26 (1995) pp 287-313. 1996.

[7] Barkaoui, K., Eslamichalandar, M., Kaabachi, M. ’A Structural Verification of Web Services Composition Compatibility.’ In J. Barjis, M.M. Narasipuram, G. Rabadi, J. Ralyté, and P. Plebani (Eds.) CAiSE 2010 Workshop EOMAS’10, Hammamet, Tunisia, pp. 30-41.

[8] Foster, H., Uchitel, S., Magee, J. and Kramer, J. ’Model-based Verification of Web Service Compositions’ Proceedings 18th IEEE International Conference on Automated Software Engineering.

[9] Bultan, F. and Su, T. ’Analysis of interacting BPEL web services’ the 13th international conference onWorld Wide Web. New York, NY,USA:ACM Press 2004.

[10] Papapanagiotou, P., Fleuriot, J. ’Formal verification of web services composition using linear logic and pi-calculus’. In ninth IEEE European Conference on Web services (ECOWS), pp 31-38. IEEE (Septembre 2011).

[11] Hamadi, R. and Benatallah, B ’A Petri Net-Based Model for Web Service Composition’ fourfteenth Australian Database Conference.

[12] Ding, W., Tian, Z., Wang, J., Zhu, J., Liang, H., Zhang, L., ’Conflicts Analysis for InterEnterprise Business Process Model’ Systemics, Cybernetics and informatics Volume 1, Number 3.

[13] Benatallah, B., Casati, F., Toumani, F., and Hamadi, R. ’Conceptual Modelling of Web Service Conversation’ 15th International Conference Advanced Information Systems Engineering (Caise’03) Klagenfurt Austria.

[14] Berardi, B., Calvanese, D., De Giacomo, C., Lenzerini, M., and Mecella ,M. ’Automatic composition of e-services that export their behavior’. In International Conference ServiceOriented Computing 2003.

[15] Benatallah, B., Casati, F., Grigori, G., Nezhad, H. R. M., and Toumani, F. ’Developing adapters for web services integration’ in International Conference Advanced Information Systems Engineering (CAiSE05), 2005, pp. 415-429. 2005.

[16] Nezhad, H. R. M., Benatallah, B., Casati, F., and Toumani, F. ’Web services interoperability specifications, IEEE Internet Computing, vol. 39, no. 5, pp. 24-32, 2006.

[17] Dörr, H.: ’Efficient Graph Rewriting and its Implementation’, volume 922 of Lecture Notes in Computer Science Springer-Verlag, 1995.

[18] Dodds, M., Plump, D.: Extending C for checking shape safety. In Proceedings Graph Transformation for Verification and Concurrency, Electronic Notes in Theoretical Computer Science Elsevier, 2005.

[19] Dodds, M., and Plump, D. (2006) 'Graph Transformation in Constant Time' Third International Conference Graph Tranformation Natal, Rio Grande do Norte, Brazil.

[20] Taentzer, G. 'AGG: A Graph Transformation Environment for Modeling and Validation of Software.' In J. Pfaltz, M. Nagl, and B. Boehlen, editors, Application of Graph Transformations with Industrial Relevance (AGTIVE'03), volume 3062 of LNCS, pages 446 - 456. Springer, 2004.

AUTHORS

Redouane Nouara
Bachelor of Science (B.Sc.), Computer Science, University of Constantine, Algeria, 1993.
Magister, Computer Science,University of Tebessae, Algeria, 2008.
Associated professor 2010.
Research:
- Formal Tool.
- Graph Transformation
- MDA Approach
- Model Transformation.
- System Checking.

Allaoua Chaoui
Bachelor of Science (B.Sc.), Computer Science, University of Constantine, Algeria, June 1986
Master of Science (M.Sc.), Computer Science,University of Constantine, Algeria, 1992.
Doctor of Philosophy (Ph.D.), Computer
Science, University of Constantine, Algeria, 1998. Research:
- Distributed Computing,
- Software Engineering,
- Theory of Computation
- Formal Tool.
- MDA Approach
- Model Transformation.

AUTHOR INDEX

Abdelaziz Khamis 107

Abdelhamid Djouadi 121

Ahmed Ahmed 01

Ahmed Barnawi 87

Aisha Mohmmed Alshiky 87

Aliaa Youssef 139

Alireza Ahmadyfard 77

Allaoua Chaoui 149

Amr Ghoneim 139

Bhuiyan M.R.A. 97

Chandrasekaran Nammalwar 49

Dimitar Radev 23, 37

Elshaimaa Nada 01

Fatih Korkmaz 97

Hayati Mamur 97

Hisham AlMajed 107

Ismail Atacak 97

Kim Myoung Sook 59

Ko Hye Young 59

NebojsaI. Jaksic 121

Omar F.Almallah 11

Park Young Eun 59

Rajeswari Chandrasekaran 49

Redouane Nouara 149

Sameh Zarif 139

Sanaz Hadipour Abkenar 77

Seyed M. Buhari 87

Songül Albayrak 11

Soumaya Chaffar 59

Strahil Sokolov 23, 37

Svetla Radeva 37

Syed Quddus 69

Wafaa Al-Mutiri 01

Yuliyan Velchev 37

Miglena Temelkova 23