International Journal of Computer Applications Technology and Research
Volume 9–Issue 08, 225-228, 2020, ISSN:-2319–8656

The Effect of Queuing System Capacity on the Blocking Probability Value Using M/G/1/C and G/G/m/K Model in Cloud Data Center

Yuniarmelinda Ikha Meru B.
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Sholeh Hadi Pramono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Erni Yudaningtyas
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Abstract: Information technology is growing fast. The growth of information technology is caused by the increasing of human need for technology. The example of it is cloud computing. The more user who accesses the cloud service can cause the possibility of users being rejected or blocking. The more tasks that are not served so that the performance of cloud data centers becomes less effective. One of method to minimize the blocking probability by controlling the queue system capacity [3]. The system model used in this research is a queuing theory model, load balancing and several physical machines. The queuing model used to analyze is M/G/1/C on load balancing and G/G/m/K on physical machine. General distribution queuing model compatible with the dynamic characteristics of cloud computing. The purpose of this research is to see the effect of queuing system capacity on blocking probability and see the effect of load balancing in this system model. The results of the simulation are the capacity of the queue system which is getting bigger, reducing the possibility of blocking. From this simulation the effective blocking probability value with a queuing system capacity of 3000. The existence of load balancing in this model makes physical machine performance effective. Because load balancing divides the task load equally into each physical machine.

Keywords: Data center, Cloud computing, Blocking probability; Response system; Queuing theory

1. INTRODUCTION

Information technology is growing fast. The growth of information technology is caused by the increasing of human need for technology. The example of it is cloud computing. In recent years cloud computing has attracted attention in the industrial world. Definition of Cloud Computing is a computing model that allows ubiquitous (wherever and whenever), convenient, on-demand network access to computing resources that can be quickly released or added. The advantage of cloud computing is flexible and efficient access [1]. In general, cloud computing is divided into three types namely Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS) [1].

Cloud computing technology consists of several components, one of which is the cloud data center. Cloud data centers are like a home for all data that is connected through a server. The Cloud Data Center receives every day from users who want to access services randomly and in large numbers. this makes users have to wait to get service from CDC. To analyze it, you can use queuing theory. Queue theory can be used to examine the activities of service facilities in a series of random conditions from a queue that happen [2].

The growth of cloud computing has resulted in a high increase in users to access cloud services. This can cause blocking probability. Blocking Probability is the possibility of blocking / rejection by the system of user requests for services to the Cloud Data Center. The higher the blocking probability, the more tasks that are not served so that the performance of the cloud data center becomes less effective. To minimize the occurrence of blocking probability, by controlling the capacity of the queuing system [3].

This research use queuing theory with the M/G/1/C queuing model on the load balancing unit and G/G/m/K on the physical machine. The function of load balancing in this research is to balance the workload on each physical machine. So the optimal results can be seen from the parameters of the average response time and the average queue length. Service rate on load balancing and physical machines uses general distribution, because the dynamic characteristics of cloud data centers produce high arrivals. Therefore, this research will increase the capacity of the queuing system to minimize the blocking probability value. This research also uses load balancing to support performance parameters of cloud data center, so it can obtain an effective value used queuing theory. The results of the implementation of this system were simulated using Java Modeling Tools V.1.0.4.

2. RELATED WORK

Although cloud computing has attracted research attention, but only a few of research that discusses this. Research [3], use queuing theory in cloud data centers. In this research, the queuing model on arrival and service time used general distribution with G/G/c model. In this research [2] used a general distribution because the distribution is more relevant to the characteristics of cloud data center. From the results of this study, the value of blocking probability is decreases equivalent with the increasing of system capacity.

Other research, queuing theory is used to analyze the performance of cloud data centers by Said, et al. On this research [5] used load balancing which is modeled by M/M/1/C and M/M/m/K (K>m) queuing models on physical machines. This study analyzes the performance of cloud data center performance with load balancing and makes a numerical model to find out the number of virtual. Therefore, in this research used a queuing model with general distribution service time and adding load balancing to the system model.

3. METHOD

3.1 Queuing Theory

Queuing is a situation that we see in our daily lives. Such as waiting in line in shopping malls or in buildings, vehicles waiting for traffic light, customers waiting in cashier at supermarkets and so on. According to Taha (2007), queuing theory is a theory that discusses mathematical learning from queues or waiting lines.

In general, customers come into a system with a random time, can't be arranged and can't be served immediately so they have to wait. Therefore, queuing theory is used to optimize services without having to wait too long. In addition, queuing theory can also be used to examine the activities of service facilities in a series of random conditions of a queuing system that happen [3].

The queuing factors are the distribution of arrivals, service distribution, service facilities, service discipline and queue length. The user arrival is usually calculated through time between arrivals. It is time between arrival of two consecutive customers in a service. Then, the service is determined by the service time. It is the time needed to serve users in a service. The next component is the system capacity. System capacity is the maximum number of customers, including those being served and those in the queue, which can be accommodated by service facilities at the same time. The last component is queuing discipline.

3.2 M/G/1/C

The queuing system model in this research used M/G/1/C queuing model on the load balancing unit. User arrival process used Markovian, arrival time used exponential distribution, service time used gamma distribution with mean 1/µ . The explanation of the method is:

M/G/1/C

with:

M : arrival rate used exponential distribution
G : service rate used gamma distribution
1 : number of unit is 1
C : queue system capacity is C

The average value of response time (R) and the average value of the number of tasks in the system (q) can be described as follows [4]:

$$R = \frac{\rho\ (1 + \mu^2 \sigma_s^2)}{2\mu\ (1 - \rho)}$$

$$\bar{n} = \bar{q} + \rho$$

$$\bar{q} = \rho^2\ \frac{(1 + \mu^2 \sigma_s^2)}{2(1 - \rho)}$$

with:

R	: Average response time
$\bar{n}$	: Average number task in systems
$\bar{q}$	: queue length
ρ	: average arrival rate divided average service rate
µ	: Service rate
σ_s	: Coefficient of Variation

3.3 G/G/m/K

The queuing system model on the physical machine used the G/G/m/K queuing model. User arrival process used general distribution and arrival time used gamma distribution, service time used gamma distribution with mean 1/µ. The explanation of the method is:

G/G/m/K

with:

G : arrival rate used general distribution
G : service rate used general distribution
m : number of unit is 1
K : queue system capacity is K

This physical machine considers the average queue length, the average number of tasks in the system, the throughput and the average system response time. To calculate the average number of tasks in the system (L) can be described as follow [3]

$$\mathrm{L} = \sum_{n=1}^{N} n\ x\ P_n$$

$$\theta = \sum_{n=1}^{N} U_n\ x\ P_n$$

$$\mathrm{R} = \frac{\mathrm{L}}{\theta}$$

with:

L	: Average number task in the system
N	: System Capacity
n	: Number task in the system
P_n	: Probability of n-task in the system
θ	: Throughput
U_n	: Total number of servers
R	: Average response time

3.4 Blocking Probability

Blocking Probability is the possibility of blocking / rejection by the system of user requests for services to the cloud data centre. The blocking probability value can be determined by queuing theory. The equation to find the value of blocking probability as follows [7]:

$$P_C = \frac{\rho\left(\sqrt{\rho}s^2 - \sqrt{\rho} + 2C\right)/\left(2 + \sqrt{\rho}s^2 - \sqrt{\rho}\right)(\rho - 1)}{(\rho^2\left(\left(1 + \sqrt{\rho}s^2 - \sqrt{\rho} + C\right)/\left(2 + \sqrt{\rho}s^2 - \sqrt{\rho}\right)\right) - 1}$$

$$\rho = \frac{\lambda}{\mu}$$

with:

P_C : Probability C- task in the system
λ : Arrival rate
μ : Service rate
s : Coefficient of Variation
C : System Capacity

3.5 Proposed Model

This research used a system model by IaaS (Infrastructure as a service). IaaS service providers provide IT infrastructure such as server, memory storage, virtual machines and operating system. Data center is a place that provides cloud computing services related to data and information communication [6]. Cloud data center also can be used as a group of servers used by IaaS service providers to meet user needs. Figure 1 below is a picture of a system model in the cloud data center.

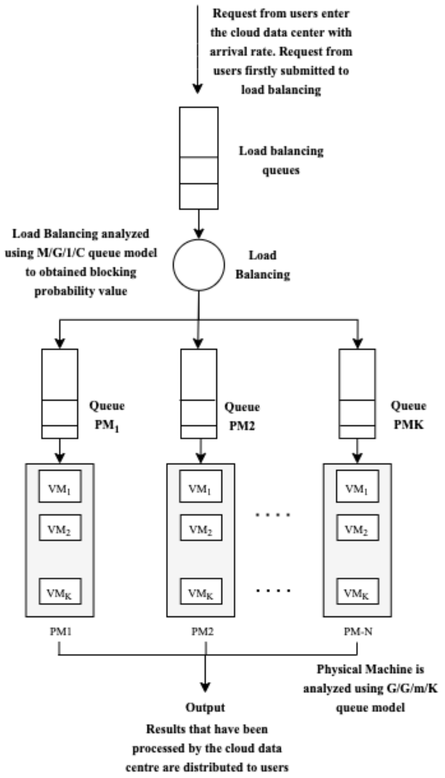

Request from users enter the cloud data center with arrival rate. Request from users firstly submitted to load balancing
Load balancing queues
Load Balancing analyzed using M/G/1/C queue model to obtained blocking probability value
Load Balancing
Queue PM$_1$
Queue PM2
Queue PMK
VM$_1$
VM$_2$
VM$_K$
. . . .
PM1
PM2
PM-N
Physical Machine is analyzed using G/G/m/K queue model
Output
Results that have been processed by the cloud data centre are distributed to users

Figure 1. System Model Cloud Data Center

Cloud data center system model in figure 1 was analyzed using queuing theory, which was tested using Java Modeling Tools simulator. Requests from users sent to the cloud data center, for example requests come to the cloud data center to access websites hosted on a physical machine with arrival rate λ. Requests from the user will enter Load Balancing before being directed to the Physical Machine. The function of load balancing is to balance the load on each physical machine. The N symbol in Figure 1 is the number of Physical Machines in the cloud data center and K is the number of Virtual Machines in each Physical Machine. To analyze load balancing used M/G/1/C queuing model. The symbol C in the queue model is the total number of tasks that can be accommodated in the load balancing queue. To analyze number of Physical Machines used G/G/m/K queuing model, where K is the total system capacity on each physical machine.

4. RESULT AND DISCUSSION

4.1 Blocking Probability Result

User requests that came to access cloud services are received and serve by load balancing with service time of 0.001. Queue system capacity value is changed from 400, 800, 1000, 2000, 3000, 3500. With the changing capacity of the queuing system, a blocking probability value is obtained. the blocking probability value is found in load balancing. The results are as in figure 2.

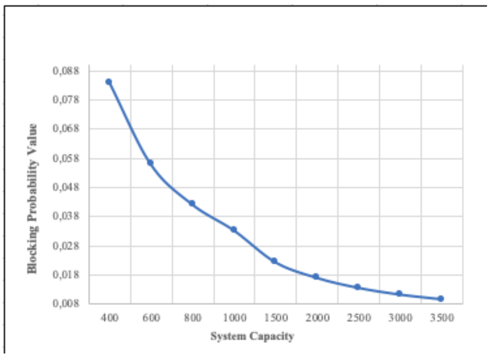


Figure 2. System Capacity vs Blocking Probability

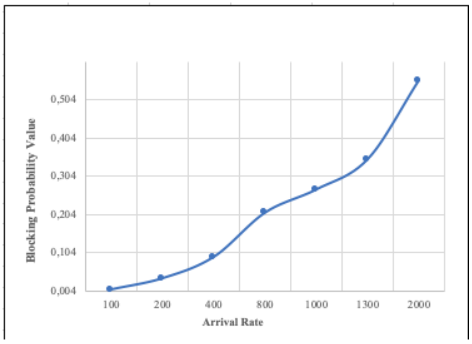


Figure 3. Arrival Rate vs Blocking Probability

From these results it can be analyzed that the blocking probability value decreases with the change in the queuing system capacity. The optimal blocking probability value when the system capacity is 3500. However, the higher arrival rate so the blocking probability value also higher.

4.2 Effect Load Balancing in Physical Machine

The service rate on load balancing used as arrival rate on physical machine. However, to be arrival rate, the amount of service rate is divided by the number of physical machines. Service rate on physical machines is 0.015 with a coefficient of variance 1.4. The number of virtual machines is changed from 22-30 units. The following results are obtained in Figure 4.

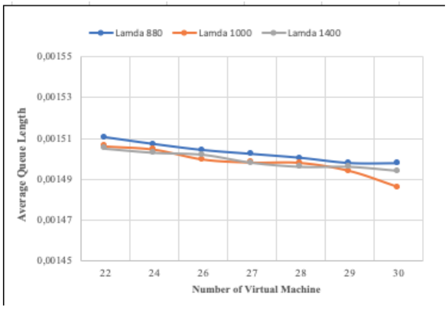


Figure 4. Average Queue Length in Physical Machine

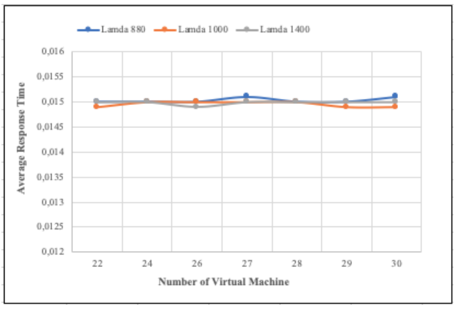


Figure 5. Average Response Time in Physical Machine

Using load balancing in the cloud data center system model, made physical machine work optimally. It can be seen as average queue length and average response time on each physical machine is constant. It is because load balancing divided the same load to each physical machine. The average queue length and response time decreases if the number of virtual machines being used is increasing.

5. CONCLUSION

This research can analyze the effect of system capacity on blocking probability values and load balancing on cloud data center performance. The method used is the M/G/1/C and G/G/m/K queuing model. Service time distribution on load balancing and physical machines used general distribution. The results of this research are to minimize the occurrence of blocking probability by increasing the capacity of the queuing system and with the load balancing unit, the average value of response time and the average length of the queue on the physical machine are more constant, because load balancing can balance the tasks on a physical machine.

6. REFERENCES

[1] Mell, Peter & Grance, Timothy. 2011. The NIST Definition of Cloud Computing. National Institute of Standards and Technology

[2] Kakiay, Thomas J. 2004. Dasar Teori Antrian untuk Kehidupan Nyata. Yogyakarta: Penerbit Andi

[3] Atmaca, T., Begin, T., Brandwajn, A., & Castel Taleb, H. 2016. Performance Evalution Centers with General Arrival and Service. IEEE

[4] Murdoch, J. 1978. Queuing Theory Worked Examples and Problems. The Macmillan Press. Ltd.

[5] El Kafhali, Said & Salah, Khaled. 2017. Stochastic Modelling and Analysis of Cloud Computing Data Center. IEEE

[6] Geng, Hwaiyu. 2015. Data Center Handbook. John Wiley & Sons, Inc., Hoboken, New Jersey

[7] MacGregor Smith, J. 2004. Optimal Design and Performance Modelling of M/G/1/K Queueing Systems. Elsevier

[8] Jagerman, D. L., Balcioglu, B., Altiok, T. & Melamed, B. 2004. Mean Waiting Approximations in the G/G/1 Queue. Kluwer Academic Publisher

International Journal of Computer Applications Technology and Research
Volume 9–Issue 08, 229-233, 2020, ISSN:-2319–8656

Food Ordering Application with Scheduling Feature Based on Mobile Web (Pasti Makan)

I Gusti Bagus JB Surya Bhuana
Departement of Information Technology
Faculty of Engineering, Udayana University
Badung, Bali, Indonesia

I Nyoman Piarsa
Departement of Information Technology
Faculty of Engineering, Udayana University
Badung, Bali, Indonesia

I Made Sukarsa
Departement of Information Technology
Faculty of Engineering, Udayana University
Badung, Bali, Indonesia

Abstract: Over time, the development of information technology has produced a variety of applications that provide convenience for humans, such as the application of food delivery. This application was designed from a background of productive individual conditions who have less time to buy food. Gofood is one of the features in the Gojek application that realizes the concept of food delivery. However, this feature does not give consumers the freedom to determine the food arrival time. Consumers did self estimation about the time when placing an order to obtain the arrival time as desired. Based on these weaknesses, Pasti Makan application was designed. This application provides scheduling features so that food ordering can be done at various times and the order will still arrive in accordance with the time specified by the consumer. This research focuses on solving the problem of the arrival time in order delivery. The purpose of this study is that buyers get their orders within the allotted time. This system made by using the PHP programming language and Javascript.

Keywords: e-commerce; food delivery; ordering; scheduling; website

1. INTRODUCTION

In the era of globalization, technological progress has developed very rapidly [1]. The development of technology, causes information can be received easily and quickly [2]. It cannot be denied that information technology has become one of human's main needs [3]. One of the developments in information technology is the internet [4].

The use of internet in business has developing, from electronic information exchange to business strategy applications, such as marketing, sales, and customer service [5]. The collaboration of internet and digital media, is able to create applications that make it easy for the community in their activities, such as the application in ordering food. Some companies have realized that the application of food ordering is a potential idea that can bring profits.

One example of food ordering application is Gojek. Gojek is one of the providers of online transportation services through applications available on the Android and iOS operating systems [6]. The application offers ordering and delivering food directly to the buyer. However, Gojek drivers do not always deliver their orders quickly and become a problem for Gojek users who have a tight time. The problem of delivering food orders which is fulfill of consumers wishes, is a problem for consumers who have a limit time.

Based on these problems, researchers designed a food ordering application in the form of a website application that provide guaranteed delivery of food orders according to the desired time. This solution is expected to solve the problem of time that is often experienced by consumers who have a limit time. In addition to solving the problem, this application is expected to open new jobs for the community, especially for mothers who want to sell food without having to open a shop.

2. METHODOLOGY

The research method used in compiling a service system called "Food Order Application System with Mobile Website Based Scheduling Features (Pasti Makan)" is waterfall model system design methodology.

2.1 Application Overview

General image of the system from making the Pasti Makan Website Application. The system of mobile website application Pasti Makan uses HTML and PHP languages as the basis for making the system.

Pasti Makan Application Website is an ordering system that can be done online by anyone and at any time using a scheduling system. Figure 1 is an illustration about Application Website Pasti Makan system.

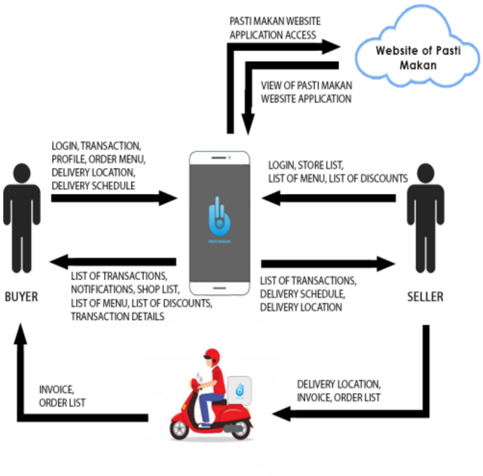

PASTI MAKAN WEBSITE APPLICATION ACCESS
Website of Pasti Makan
VIEW OF PASTI MAKAN WEBSITE APPLICATION
LOGIN, TRANSACTION, PROFILE, ORDER MENU, DELIVERY LOCATION, DELIVERY SCHEDULE
LOGIN, STORE LIST, LIST OF MENU, LIST OF DISCOUNTS
PASTI MAKAN
LIST OF TRANSACTIONS, NOTIFICATIONS, SHOP LIST, LIST OF MENU, LIST OF DISCOUNTS, TRANSACTION DETAILS
LIST OF TRANSACTIONS, DELIVERY SCHEDULE, DELIVERY LOCATION
BUYER
SELLER
INVOICE, ORDER LIST
DELIVERY LOCATION, INVOICE, ORDER LIST

Figure 1. Application Overview

Figure 1 describing that the buyer makes transactions on the system with the desired order menu. After that, the buyer provides information about buyer data, location and schedule to the system. The system begin provides information to store users to process orders that have been ordered. Orders are then delivered to the buyer by the store according to the buyer's specified schedule.

2.2 Data Flow Diagram

Data Flow Diagrams of Pasti Makan Application Website is describe as Figure 2.

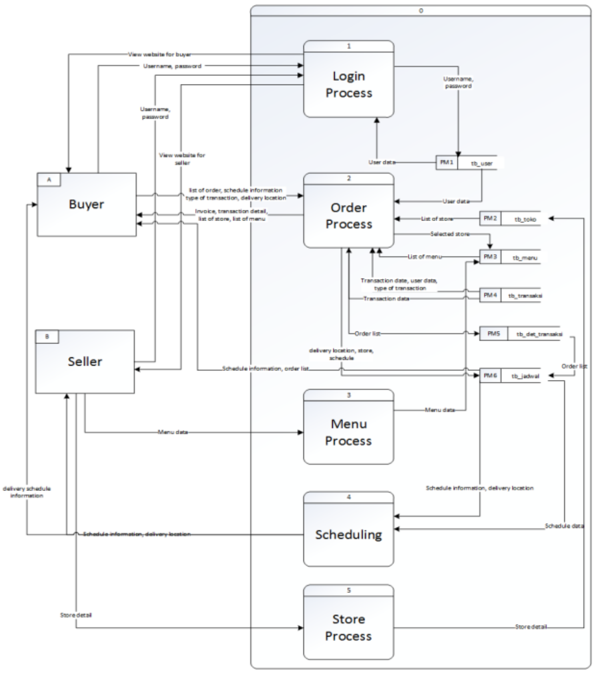


Figure 2. Data Flow Diagram

From Figure 2, there are 6 main processes, namely the login process, the ordering process, the menu process, the store process and the scheduling process. The login process is the process when the buyer or seller user enters the system by entering a username and password. The ordering process is the process when the buyer makes an order through the Pasti Makan system. When ordering process, buyer needs to choose the store and menu to be ordered while entering the delivery location, delivery date and delivery time. The menu process is the seller's process for managing menu data from owned stores, such as adding new menus, changing menus and removing menus. The store process is the seller's process for managing the store, such as deleting a store, changing store information and adding a new store. The scheduling process is the process of the system to display the list of interagency schedules to sellers through transaction data sorted by the closest interagency schedule.

2.3 Entity Relationship Diagram

Entity Relationship Diagram of Pasti Makan can be seen from Figure 3.

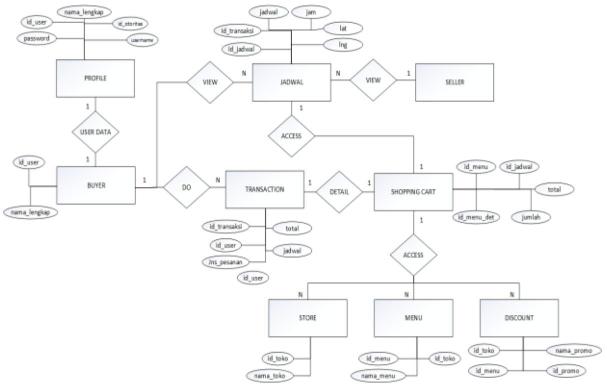


Figure 3. Entity Relationship Diagram

Figure 3 explains how the relationship between entities, processes and databases. This application has three entities namely buyer, seller and admin. Buyers have a relationship between admin and seller. The buyer places an order with the seller. While the seller delivers the buyer's order according to the specified schedule.

3. LITERATURE REVIEW

Literature review is material that is used as a reference in making research. The following references are to an explanation of Cloud Computing, Google Maps API, taking order, PHP, Javascript, and scheduling.

3.1 Cloud Computing

Cloud Computing is a combination of using computer technology and development based on Internet [7]. According to Satya Saputra (2017), cloud computing is a service model for sharing configurable computing resources (for example, networks, servers, storage, applications and services) that can be quickly run over the internet. One of the advantages of cloud technology is allows users to store data centrally on one server based on services provided by cloud computing service providers [8]. Cloud computing services have 3 service models that can be used as needed, 3 services include IaaS, PaaS and SaaS [9].

3.2 Google Maps API

Google Maps is software on the Internet that contains maps of an area or location [10]. Google Maps can be displayed on the web or external applications by using the Google Maps API [11]. The Google Maps application can be displayed on a particular web or application that requires an API key as a unique code generated by Google for a particular website or application so that the Google Maps server can recognize developers who use the Google Maps API service [11].

3.3 Taking Order

Taking Order in a restaurant is the activity of receiving and recording guest orders. In this case, food and drinks will be forwarded to the relevant section, including the kitchen, bar, and cashier [12].

Taking Order includes several activities, such as:

1. Displays accurate information about all available foods and drinks in the menu list.
2. Note the menu ordered, the number ordered, the name of the customer and others.
3. Confirm the order to the customer.
4. Forward the order to the related section.

3.4 Scheduling

Scheduling is a planning activity to determine when and where each operation as part of the overall work must be done on limited resources, as well as allocating resources at a certain time by taking into account the capacity of existing resources.

The main function of production scheduling is to make the production process run smoothly according to the planned time, so that it works at full capacity with minimal costs and the desired quantity of products can be produced on time [13].

3.5 PHP (Hypertext Prepocessor)

PHP is a scripting language that can be embedded or inserted into HTML. PHP is widely used for dynamic website programming [14]. PHP is referred to as HTML embedded server side scripting, because all scripts in PHP are run on the server side. PHP scripts integrated with HTML [15]. PHP code can be built on a web page system by building it with pure PHP language, combined with HTML code, or combined with various template engines and web frameworks. [16].

PHP is used and run on a web page to process the contents of the website seen by visitors of the website [17].

3.6 Javascript

Javascript is a scripting language, which is a set of instructions commands used to control some parts of the operating system [18]. Javascript is developed to be able to run on a web browser or client side [18].

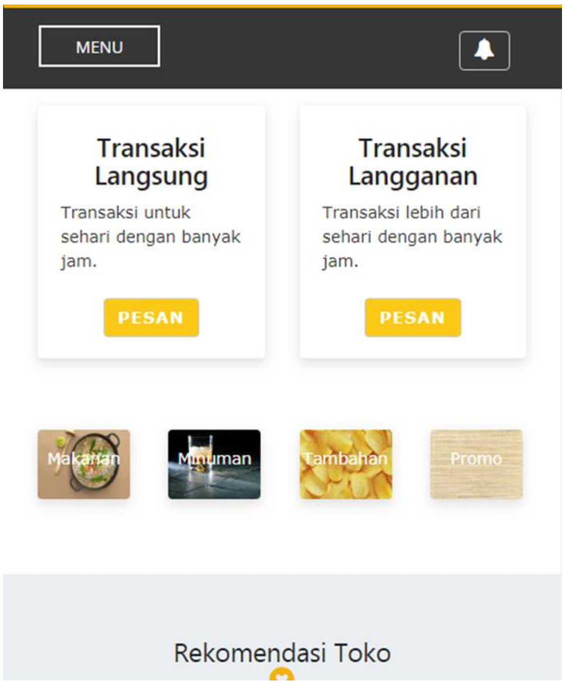
Javascript makes the server load lighter and web pages will respond much faster, even on the lowest internet connections [19]. JavaScript is a category of language that is case sensitive which means that it can distinguish between variables and functions in the use of upper and lower case letters [19].

4. RESULTS AND DISCUSSION

Testing the creation of a Website Application Pasti Makan on the user's authority as a buyer and seller. Buyer transactions in this application can be divided into two, namely direct transactions and customer transactions. In direct transactions, buyers can only schedule on the same day, whereas in subscription transactions, buyers are not limited when scheduling.

4.1 Preview of Main Menu

The buyer's main page in Figure 4 is the first main page the buyer encounters shortly after logging in.



Transaksi Langsung

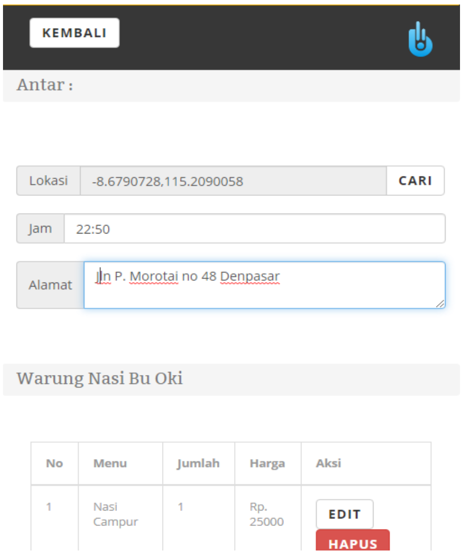
Transaksi untuk sehari dengan banyak jam.

Transaksi Langganan

Transaksi lebih dari sehari dengan banyak jam.

Rekomendasi Toko

Figure 4. Preview of Main Menu



Antar :

Lokasi: -8.6790728,115.2090058 CARI

Jam: 22:50

Alamat: Jln P. Morotai no 48 Denpasar

Warung Nasi Bu Oki

No	Menu	Jumlah	Harga	Aksi
1	Nasi Campur	1	Rp. 25000	EDIT HAPUS

Figure 5. Preview of Direct Transaction’s Basket Page

The main menu of Pasti Makan Ordering Application System, contains menu recommendations, store recommendations, categories and direct or subscription transaction buttons.

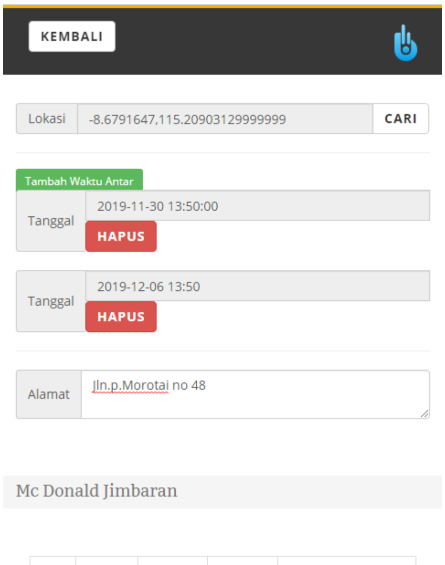
4.2 Page of Direct Transaction’s Basket

The direct transaction’s basket is a page to view menus ordered by buyers when making direct transactions. The direct transaction’s basket page is specifically for buyers who are in the process of purchasing direct transactions.

The direct transaction basket page in Figure 5 is the page where buyers place orders directly on the system. The buyer selects a store, and then chooses the menu to be ordered along with the amount. The ordered menu is entered on the basket page. This basket page is in the subscription order process and directly to accommodate the ordered menu data. Data needed to conduct transactions is to mark the location of delivery using Google Maps, delivery hours and delivery addresses.

4.3 Page of Subscription Transaction's Basket

Page of subscription transaction's basket is a page to see the menu ordered by the buyer when making a subscription transaction. This page is specifically for buyers who are in the process of purchasing a subscription transaction.



Lokasi -8.6791647,115.20903129999999

Tanggal 2019-11-30 13:50:00

Tanggal 2019-12-06 13:50

Alamat Jln.p.Morotai no 48

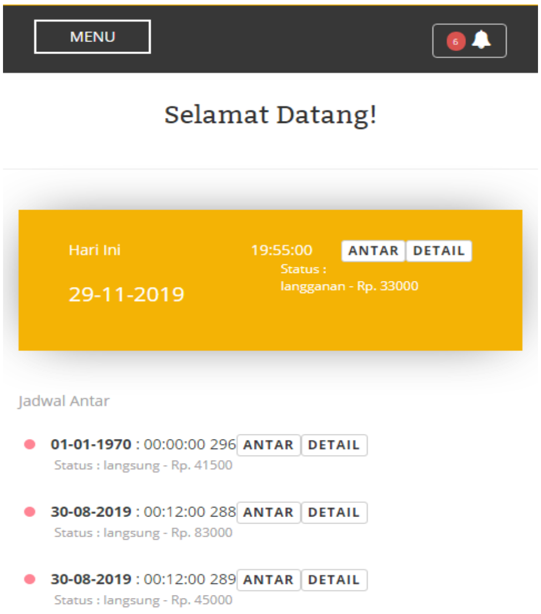
Mc Donald Jimbaran

Figure 6. Preview of Subscription Transaction's Basket Page

Page of subscription transaction's basket in Figure 6 is the page where the buyer places a subscription on the system. The buyer selects a store, then chooses the menu to be ordered along with the amount. The ordered menu is entered on the basket page. This basket page is in the subscription order process and directly to accommodate the ordered menu data. The data required to make a transaction is to mark the location of delivery using Google Maps, the date and time that the subscription order will be delivered.

4.4 Page of Seller Delivery Schedule

The inter-seller schedule page is the main page of the seller's user to view the delivery schedules.



Selamat Datang!

Hari Ini 29-11-2019

19:55:00 Status : langganan - Rp. 33000

Jadwal Antar

- 01-01-1970 : 00:00:00 296 Status : langsung - Rp. 41500
- 30-08-2019 : 00:12:00 288 Status : langsung - Rp. 83000
- 30-08-2019 : 00:12:00 289 Status : langsung - Rp. 45000

Figure 7. Preview of Seller Delivery Schedule Page

Figure 7 is a preview that presents information about the delivery schedule of orders to be delivered by the seller. Information presented in the form of date, time and total transaction of the order. The information presented is schedule information between today and delivery schedule information for the following days.

5. CONCLUSION

The Website Application System of Pasti Makan is produced using cloud computing technology that makes buyers and partners can access the system anywhere and anytime without installing the application. The scheduling system, which is the main feature of this system, can solve the problem of productive people who have little time to get orders by delivering orders on time by providing information in the form of delivery location, delivery date and delivery schedule.

6. REFERENCES

[1] Swastika, I. P., Widiatmika, I. M., & Wiadi, P. E. 2010. Rancang Bangun Sistem Informasi Geografis Penguasaan Pemilikan Penggunaan dan Pemanfaatan Tanah (P4t) Kabupaten Jembrana Berbasis Web. Lontar Komputer, 1(1), 4-16. Darmawan, I. P., Piarsa, I. N., & Dharmadi, I. P. (2017).

[2] Darmawan, I. P., Piarsa, I. N., & Dharmadi, I. P. 2017. Ekstrak Hirarki Data dari Situs Web A-Z Animals Menggunakan Web Scraping. Lontar Komputer, 8(3), 166-177.

[3] Nurastryana, K. W. 2003. Implementasi Tabel Federasi dalam Komunikasi Server pada Penjadwalan Seminar Tugas Akhir. Merpati, 1(2), 1-7.

[4] Frediyatma, S. Y. 2014. Aplikasi Pemesanan Makanan Berbasis Cloud dengan Platform Android. Merpati, 2(1), 118-126.

[5] Khairunnisa, & Damayanti, F. 2018. Pengolahan Bisnis Catering Ummi Nisa Medan Berbasis Web. Jurnal Sistem Informasi, 2(1), 63-71.

[6] Prawiranata, H., & Rahmawati, D. 2018. Pengaruh Kualitas Sistem Informasi, Harga dan Kualitas Pelayanan Terhadap Kepuasan Pelanggan Pada Jasa Gojek di Yogyakarta. Jurnal Pendidikan Akutansi, 6(4), 1-22.

[7] Mahardika, I. G. 2014. Aplikasi Back End Manajemen Restoran Berbasis Cloud. Merpati, 2(1), 98-105.

[8] Saputra, P. S., Sukarsa, I. M., & Bayupati, I. P. 2017. Sistem Informasi Monitoring Perkembangan Anak di Sekolah Taman Kanak – kanak Berbasis Cloud. Lontar Komputer, 8(2), 112-123.

[9] Dharma, I. G., Sukarsa, I. M., & Sutramiani, N. P. 2019. Rancang Bangun Sistem E-Commerce Marketplace Gypsum Berbasis Cloud Computing. Merpati, 7(1), 37-48.

[10] Sudiartha, I. K. 2013. Sistem Informasi Geografis Pura di Pulau Bali Pada Platform Blackberry. Merpati, 1(2), 1-10.

[11] Gautama, I. W., Putra, I. K., & Sukarsa, I. M. 2016. Aplikasi Pemetaan Objek Wisata Pantai Bali Selatan Berbasis Android. Merpati, 4(1), 43-51.

[12] Utama, F. F., & Johar, A. 2016. Minuman Restaurant Berbasis Client Server Dengan Platform Android. Jurnal Rekursif, 4, 288-300.

[13] Masruroh, N. 2011. Analisa Penjadwalan Produksi dengan Menggunakan Metode Ampbell Dudeck Smith, Palmer, dan Dannenbring di PT. Loka Refraktoris Surabaya. Tekmapro, 1(1), 158-171.

[14] Pursana, P. E. 2014. Sistem Informasi Koperasi Modul Simpanan Berbasis Android Terintegrasi Berbasis Web. Merpati, 2(1), 67-78.

[15] Putra, M. S., Piarsa, I. N., & Rusjayanthi, N. K. 2018. Rancang Bangun Sistem Informasi Web-Based Travel Assistant untuk Membantu Perjalanan Wisatawan. Merpati, 6(3), 214-224.

[16] Saputra, I. G., Sasmita, G. M., & Wiranatha, A. A. 2017. Pengembangan Sistem Keamanan untuk E-Commerce. Merpati, 5(1), 17-28.

[17] Purba, I. R., Purnawan, I. K., & Sasmita, I. G. 2016. Sistem Antrean Pelayanan Medis Praktik Dokter Bersama Berbasis Web. Merpati, 4(3), 248-258.

[18] Putra, I. M., Piarsa, I. N., & Mandenni, N. M. 2015. Sistem Informasi Geografis Pemetaan Wilayah Berdasarkan Kualitas Pendidikan di Provinsi Bali. Merpati, 3(3), 108-119.

[19] Harmadya, M., Sasmita, G. M., & Wirdiani, N. K. 2015. Rancang Bangun Aplikasi Tryout Ujian Nasional Sekolah Menengah Pertama (SMP) Berbasis Android. Lontar Komputer, 6(2), 108-119.

A Duplexer Design Technology Based on Co-simulation of HFSS and ADS

Xiong Rong
Chengdu University of Information Technology
Institute of Communication Engineering
Chengdu, China

Yang Huan
Chengdu University of Information Technology
Institute of Communication Engineering
Chengdu, China

Abstract: A miniaturized duplexer with high isolation is designed in this paper. The coupling coefficient of branch filter is solved by Generalized Chebyshev function, and the circuit model is simulated and optimized by MATLAB and ADS. Finally, the physical dimensions are extracted by HFSS, and the simulation model of combiner is established and optimized. The experimental results show that the loss of the miniaturized combiner is less than 1.5dB, and the out of band rejection is greater than 40dB@1550-1580MHz&1600-1625MHz .The simulation results verify the rationality and feasibility of the design.

Keywords: Combiner; Common cavity; Coupling matrix; Model simulation

1. INTRODUCTION

With the rapid development of mobile communications, the demand for a compact and highly selective combiner is increasing. Combining two or more circuit system signals into one circuit not only saves material, but also eliminates the need to switch antennas during signal transmission. Compared with the traditional combiner, the current cavity combiner has many advantages, such as low insertion loss, high out-of-band suppression and large power capacity, which are widely used in broadcasting, radar and satellite communication systems [1-3]. Miniaturization of multi - frequency combiner has become one of the important development directions of combiner development [4].

The combiner is composed of multiple signals, so in order to ensure the communication quality of the system and prevent mutual interference between the systems, the separation of the combiner must have a high requirement. Therefore, how to design a compact and highly isolated combiner is always a problem studied by designers [5].

The current combiner is composed of two filter elements with multiple cross-coupling designs and a common cavity at the combiner end. A three-port combiner with operating frequencies ranging from 1550-1580MHz to 1600-1625MHz is designed according to this structure. In this thesis, Chebyshev filter design is adopted to obtain the coupling coefficient of two-pass filter and the number of order and zero of single-pass filter [6]. Finally, the simulation test of this combiner is carried out, and the dimensions of the combiner are reduced while the parameters of the combiner are guaranteed. This design not only makes the circuit breaker have high isolation degree, but also realizes the design structure of the combiner miniaturization to adapt to different scene requirements, which has a good application prospect.

2. RELATED WORK

A miniaturized duplexer method based on ads and HFSS co-simulation is designed. In the design process, in order to ensure that each channel can achieve the corresponding indicators, the coupling coefficients of the two channels are simulated in ads. Then, the coupling coefficients of each path are transformed into corresponding physical dimensions in HFSS. Finally, the performance of the diplexer model is compared with the requirements, and it is found that it meets the design requirements. This paper mainly includes the following contents:

- The S-parameter pre simulation is carried out by MATLAB software, and two independent coupling coefficients are obtained;

- The circuit model is simulated in ads with the coupling coefficient, and the external Q value and topology are obtained;

- The obtained structure and parameters are transformed into specific physical dimensions in HFSS, and modeling and simulation optimization are carried out

Finally, the S parameters of the optimized diplexer are compared with the expected indexes.

3. THE CIRCUIT MODEL OF THE COMBINER

The present paper designs a three-port combiner with two channels consisting of a Bandpass filter with cross-coupling design. The main indexes are shown in Table 1.

Table 1 Design indexes of combiner

Parameter	GPS	$L_{\text{short message}}$
Frequency Range/MHz	1550-1580	1600-1625
Band-width/MHz	30	25
IL at BW/dB	≤1.2	≤1.5
Return loss at BW/dB	≤-20	≤-20
Isolation/dB	≥40	≥40

In order to minimize the attenuation in the Passband, Chebyshev filter is designed according to the formula:

$$k_{i,j}^{k} = B_n \times M_K(i,j) \tag{1}$$

$$Q_e^k = 1/\{B_n[M_k(n_{Pk}+1, n_{Pk}+2)]^2\} \tag{2}$$

$$B_n = B/f_0 \tag{3}$$

$$f_{ris,i}^{k} = f_0[\sqrt{1+\left(\frac{B_n M_{i,i}^{k}}{2}\right)^2} - \frac{B_n M_{i,i}^{k}}{2}] \tag{4}$$

$$k_{0,1}^{k} = B \times (M_k(1,2)/\sqrt{c_0}) \tag{5}$$

$$Q_e = c_0/B_n \tag{6}$$

$$f_{ris,0} = f_0[\sqrt{1+\left(\frac{B_n b_0}{2}\right)^2} - \frac{B_n b_0}{2}] \tag{7}$$

Where, equations (1) - (4) are applied to non-common cavities, and equations (5) - (7) are applied to common cavities. $k_{i,j}^{k}$ is the internal coupling coefficient of the filter, Q_e^k is the Q value of filter, $f_{ris,i}^{k}$ is the resonant frequency of the resonator, B is the filter bandwidth, B_n is relative bandwidth, M_k is the coupling matrix, b_0 is the normalized admittance of the common cavity, c_0 is normalized capacitance, f_0 is the center frequency, n_{Pk} is the order of the *K-th* filter[7].

According to the design theory of coupled resonator Bandpass filter and the above formula, the S-parameter obtained by Matlab simulation is shown in Figure 1, where the coupling coefficient of each branch, external *Q* value and other initial values are shown in Table 2. Each a separate filter circuit is Bandpass filter with transmission zeros equivalent circuit, the two channel filter through a public space together, after the synthesis of duplexers topology structure as shown in figure 2, the hollow circle represents the source and load, solid lines represent the main coupling, dashed line represents the cross coupling.

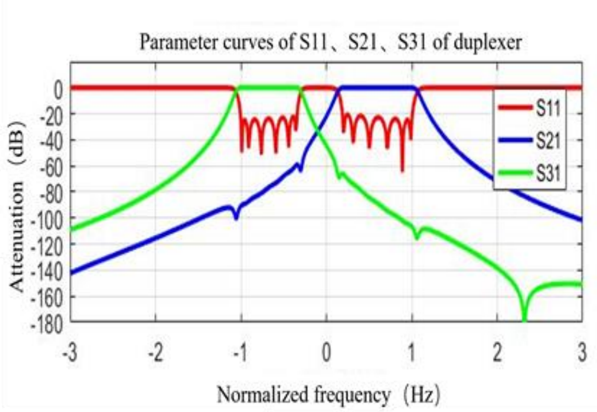


Figure 1 Matlab simulation of S parameters

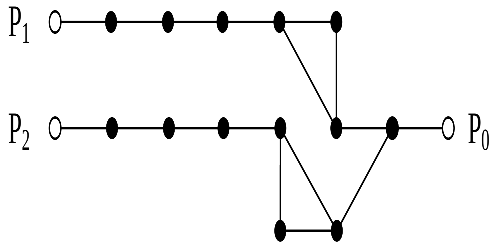


Figure 2 Topology diagram of combiner

Table 2 The coupling coefficient and the external Q value of the duplex

K	Band 1	Band 2
K12	0.0425	0.0393
K23	0.0147	0.0123
K34	0.0119	0.0097
K45	0.0114	0.0093
K56	0.0121	0.0099
K67	0.0171	0.0141
K13	0.0009	-0.0009
Q_e	46.6643	55.6215
Common port Q_e	9.0668	

Figure 2 shows that the order of the band-pass filter with a Passband frequency of 1550-1580MHz is 7, and that of the band-pass filter with a Passband frequency of 1600-1625MHz is 7. In the design, the bandpass filter adopts cross coupling structure, and the circuit model of duplexer established by the topology structure in Figure 2 is shown in Figure 3 by using the circuit simulation design software ADS.

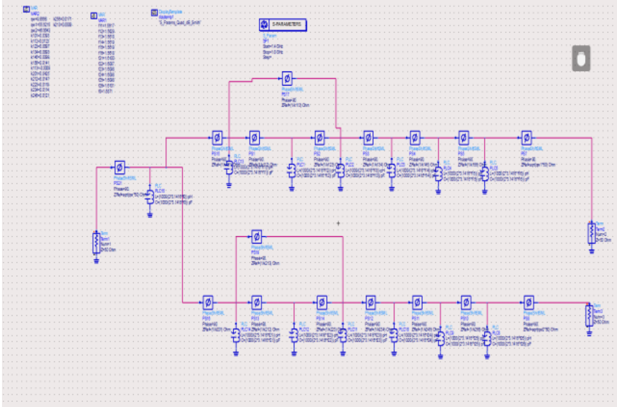


Figure 3 Circuit model of combiner

4. DUPLEX MODEL SIMULATION

HFSS software was used to simulate the combiner, and according to the optimized design parameters extracted from the ADS circuit model, the common end coupling structure, the coupling dimensions of each input end and the tap position were calculated. The structure of a single cavity can be determined by the resonant frequency through single cavity simulation by HFSS. For the extraction of the coupling coefficient between resonators, the eigen mode of HFSS is needed. In the solution environment of the eigen mode, the size of the coupling coefficient can be obtained, and the formula is as follows:

$$K = \frac{f_m^2 - f_e^2}{f_m^2 + f_e^2} \tag{8}$$

Where f_m and f_e are two single cavity resonant frequencies.

The coupling modes of tap structure include direct coupling, capacitive coupling and inductive coupling. The dimension of the multiplexer is determined by simulation of variables sensitive to the influence of circuit parameters, in which the resonant frequency is determined by the depth of the frequency modulation screw *d*, the coupling coefficient between adjacent cavities is determined by the width of the side wall window *w*, and the on-load *Q* value is determined by the height of the solder joint position *H*. Then select appropriate scanning interval and step length to scan each parameter one by one, extract the circuit parameters in the 3D model, and establish the quantitative relationship between the circuit parameters and the corresponding model size. Finally, the optimized model size parameters in HFSS are shown in Table 3.The simulation model is shown in Figure 4.

Table 3 Duplex dimension Table (unit: mm)

GPS channel		Lshort message channel	
Frequency modulation screw coupling window width and solder joint position			
d1 = 1.8	w12 = 7.6	d1 = 2.1	w12 = 7.3
d2 = 2.0	w23 = 6.4	d2 = 1.6	w23 = 6.3
d3 = 1.5	w34 = 6.2	d3 = 1.3	w34 = 6.2
d4 = 1.5	w45 = 6.5	d4 = 1.2	w45 = 6.3
d5 = 2.0	w56 = 7.0	d5 = 1.3	w56 = 6.8
d6 = 1.8		d6 = 1.8	
H = 8.0		H = 8.0	

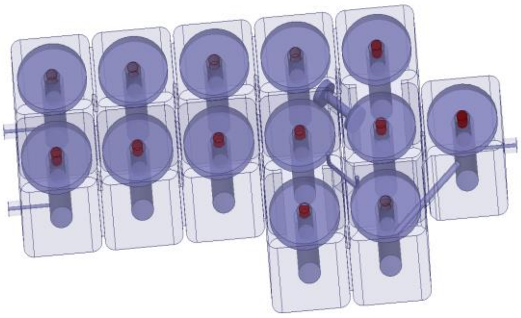


Figure 4 The simulation model of combiner

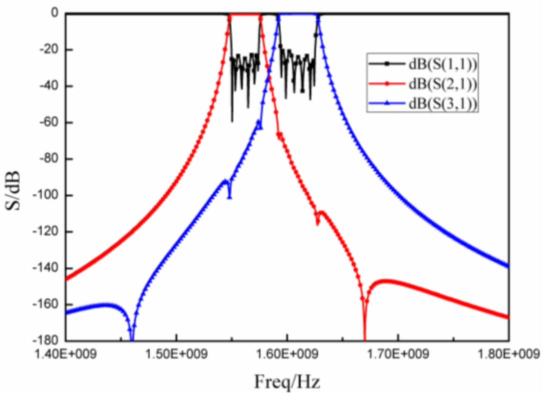


Figure 5 The optimization results of combiner

The overall model of the duplex as shown in Figure 4 is simulated, and the final result is shown in Figure 5.

Figure 5 shows that the simulation return loss in the frequency band 1550-1580 MHz is less than -20dB, the insertion loss is less than 0.5dB, and the return loss in the frequency band 1600-1625MHz is less than -20dB, and the insertion loss is less than 0.5dB, which is consistent with the expected results and proves the previous design theory.

5. CONCLUSION

In this thesis, we design a miniaturized dual combiner, based on the basic theory of filter Matlab the coupling coefficients of draw two line filter parameters, such as using circuit simulation software ANSOFT DESIGNER for circuit simulation and optimization, and through the simulation software HFSS design meets all design index, a simulation model of each filter all the way to make good, restraint outside the band can improve the isolation. The design efficiency is high, the engineering practicality is strong, can satisfy the modern communication system to combiner miniaturization, the high isolation degree request.

6. ACKNOWLEDGMENTS

First of all, I would like to thank my research tutor for his guidance, who pointed out the problems existing in the thesis for me. Secondly, Yang Huan helped me to successfully complete the measurement of the experimental results. I really thank them for helping me to finish this paper successfully.

7. REFERENCES

[1] Macchiarella G, Tamiazzo S. Synthesis of Star-Junction Multiplex-ers [J]. IEEE Transactions on Microwave Theory and Techniques, 2010, 58(12), 3732-3741.

[2] Kim B, Kim Y S. Mixed Coupling Structure for the Cross Couplingof Combline Filters [J]. Microwave and Optical Technology Let-ters, 2002, 35（1）: 20-23.

[3] Hong J S, Lancaster M J. Couplings of Microstrip Square Open-Loop Resonators for Cross-Coupled Planar Microwave Filter[J]. IEEE Transactions on Microwave Theory and Techniques, 1996, 44（11）： 2099-2109.

[4] Feng Jianhua. Miniaturization design of multifrequency coaxial cavity combiner [J]. Mobile Communication, 2016,40(19):60-65.

[5] Chen Qihao, Ye Qiang, Feng JianHua. Design of a high-isolation cavity dual-frequency comb-solver [J]. Electronics,2016,39(02):276-279.

[6] Cameron R J. General Coupling Matrix Synthesis Methods for Chebyshev Filtering Functions [J]. Microwave Theory and Techniques, IEEE Transactions on, 1999, 47（4）: 433-442.

[7] Chen Qihao, YE Qiang, Feudal Hua. Design of a high-isolation cavity dual-frequency combine [J]. Journal of Electron Device, 2016, 39(02): 276-279.

QuMANET- changes in Mobile ad-hoc network with quantum bits for reliability

Shruti Mishra
Research Scholar
Bhagwant University,
Ajmer (Rajasthan), India

Dr. B. Dhanasekaran
Professor
SJIET
Chennai, India

Abstract: Due to the mobility, service discovery and service selection in Mobile Ad hoc Network (MANET), the routing protocol of MANET must adapt to the rapid changes of the network structure, and ensure that the services will be available to the users as quickly as possible. This paper proposes a kind of new quantum based MANET with the help of which service discovery will be fast enough that the user will not get a break in connection and swap over multiple services very fast with the help of quantum bits. By embedding MANET with quantum computing bits, it can effectively reduce the time for service selection, improve the delivery rate of data packets, and reduce the time delay of switching between services. The goal of this work is to shed light on the challenges and the open problems of the Quantum bits dealing with MANET design. To this aim, we first introduce some basic knowledge of Quantum mechanics, MANET needed to understand the differences between a MANET and the design we are thinking quantum MANET. Then, we introduce quantum bits as the key strategy for service swapping without physically transferring the particle that stores the quantum information. Finally, we will introduce the challenges that can be faced while transferring data over quantum bits.

Keywords: MANET, Quantum bits, mobility, service discovery, Qubits, QUMANET.

1. INTRODUCTION

Nowadays, Researchers are working with quantum computers and proposing new ideas about how successfully we can deal with quantum bits (qubits). On Jan 3,2020 Researchers at the Department of Energy's Oak Ridge National Laboratory simulated the performance of quantum devices. On May 18, 2020 university of California has set a new record for preparing and measuring the quantum bits, or qubits, inside of a quantum computer without error. The techniques they have developed make it easier to build quantum computers that outperform classical computers for important tasks [1]. Researchers are also working on quantum mechanical interactions. We can use networking in multiple quantum devices. Quantum MANET can be visualized to open a new way to provide fast services to the users.

All the researchers are in a race to work upon quantum computers and provide the best results in different fields. They are working on affordable quantum computers for real world business and government applications. [2] This goal achievement is crucial as to control interaction between different quantum systems in MANET remains extremely challenging.

We can create complementary approach towards mobile network to connect multiple clouds through wireless links. These small clouds will enable signals to interact with qubits even while supporting mobility.

Typically, MANET is built with no fixed infrastructure and routers. These are developed for temporary network connections. A mobile ad-hoc network (MANET) is a network supporting multi-hop with no fixed topology and offer different services to all the nodes. [3]

In this disquisition, we will focus on employment of qubits in MANET with the objective of fastest services that we can provide to the users.

The rest of the paper is structured as follows. In section 2 we will discuss the mechanics of quantum bits, their measurement, and quantum entanglement. We continue by offering a brief historical perspective of quantum computing and review the measurement of bits. In section 3, we will take a brief introduction of mobile ad-hoc network. In section 4, we will describe the advancement in present mobile ad-hoc network with advent of qubits. Finally, we will do comparison in section 5 and study the various challenges in section 6 and then we will conclude in section 7.

2. QUANTUM MECHANICS

Quantum computers are based on quantum mechanics and qubits increase the value of threshold at which information can be shared and processed. In classical computers, we were using binary bits 0 or 1 while in quantum computers we make use of both 0 and 1 by encoding them with in between values thus, providing large number of possibilities for storing data. As we know, with the help of 0 and 1 we can make only 4 combinations. But with the help of Qubits we can make 2^n combinations and we can do parallel processing practically with quantum computers.

Multiple quantum algorithms [4] are implemented to find a better solution to remove the various problems like delay in networks etc. However, dealing with superposition and entanglement is only one aspect of quantum computing. An equal challenge is to control the exponential bits in MANET which will be exciting time for breakthroughs in ad-hoc networks. This new technology can advance quickly if discovery of quantum internet will continue to push forward quantum technologies.

2.1 Quantum Entanglement

It is a special case of superposition in which we define qubits as a reference to each other where the particles can be shared or they can be spatially separated. We can define physical properties of the system with the help of these qubits where one can influence other. For example, if two systems are connected to each other in a network and if we create influencing relationship between the two with the help of qubits where processing of one system influence other with the help of qubits then all the system will be connected with the help of quantum entanglement and MANET will work properly there.

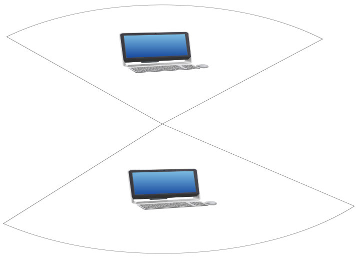


Figure 1 Quantum entanglement

Entanglement enables multiple applications since it helps to exchange of information very fast between qubits.

2.2 Quantum measurement

Quantum measurement can be considered as measuring the value of qubits that can be either 0 or 1. We consider that qubits collapse towards either 0 or 1. These outcomes are probabilistic. Superposition principle helps in determining the value of quantum states. This principle can be taken as a basis for quantum theory for determining states. We can add or superimpose multiple states to find another valid quantum state. This quantum state is a linear combination of multiple quantum states and these states can be either |0⟩ or |1⟩. These states are convertible with equal probability of 50% with 2^n states.

3. MANET

Mobile ad-hoc network is a type of network that uses the concept of decentralization. There can be direct or indirect type of communication. When the nodes are in reach of one another, they can create direct communication while if they are located far from each other, they create communication through other nodes i.e indirect communication. There are multiple applications of MANET like military, health care etc. In MANET all the nodes are free to join and leave the network. All the nodes that are joining can take part in communication either as source or destination. Bandwidth is an important network property of wireless network because it is much lower in wireless link than that of wired links.

In this paper we are going to transform this MANET into QUMANET. This transformation is required as users of wireless network are increasing. As the pandemic COVID-19 has hit all the countries, uses of MANET have increased day by day. Users like students, teachers, company employees etc. all are accessing their facility from their home. So, it is the basic need of today that we should provide fast services to all the users

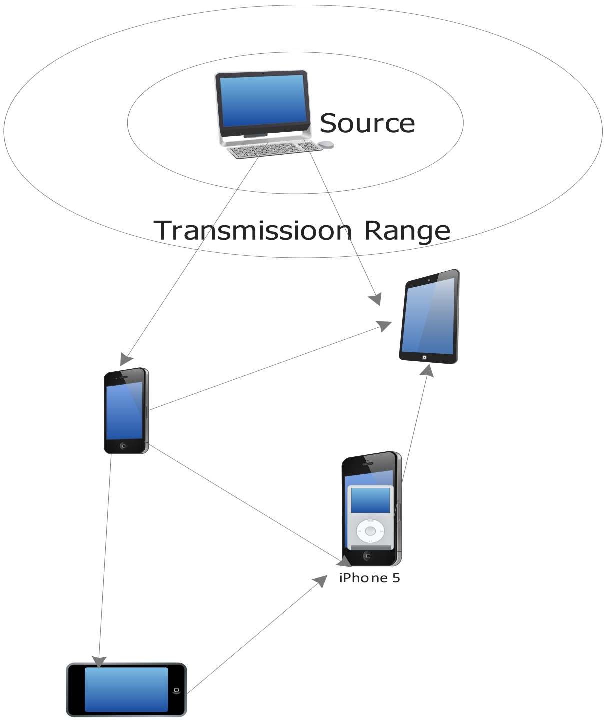


Figure 2: Mobile ad-hoc network

4. OVERVIEW OF QUMANET

In a classical MANET, the network provides QoS with the help of 0 and 1. Whenever a node is moving from one point to another and wants to access some service with the help of service provider (SP), these services may be lost due to mobility. But with the advent of quantum adaptation, some applications like quantum key distribution (QKD) and superdense coding [5] are used. Quantum MANET can enhance the concept of quantum communication by using EPR pair also called bell states in superimposed state. If we try to measure these superimposed states qubits then we can find the equal distribution of 0or 1 with equal probability. This probability will help to distributed services in a network. Thus, we will be able to transfer messages with the help of these qubits. The EPR paradox used by Einstein and his colleagues helps to identify quantum entanglement behavior of qubits.

5. COMPARISON

There are different researches [8]-[10] that show Quantum internet can be used for long- distance communication of quantum and classical information. We can use this type of network in recent technology i.e MANET.

Services that are provided by different service providers and the number of services will be more in QUMANET as compared to classical MANET because the number of states will be increased in QUMANET.

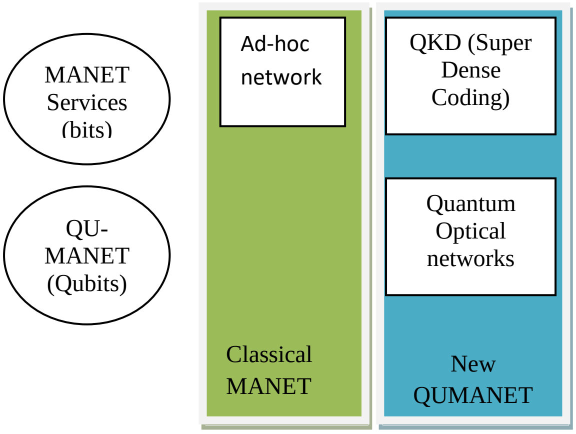


Figure 3: Classical MANET versus QUMANET

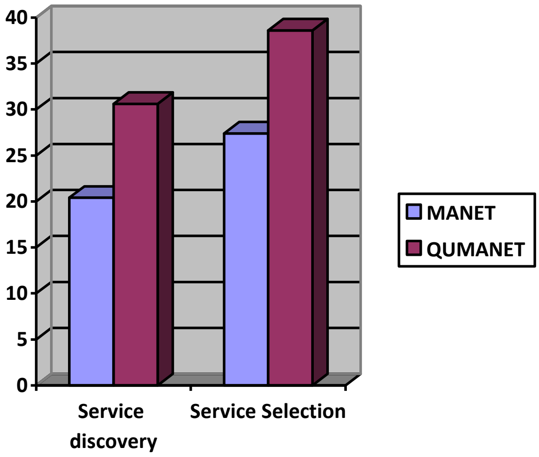


Figure 4: Comparison of MANET and QUMANET: Service discovery and selection with the help of QUMANET will be fast as compared to Classical MANET

6. CHALLENGES

Quantum MANET is still a complicated project to implement in real world as we have to find solutions of some challenges:

- ✓ Implementation of QuMANET will require realization of qubits in small clusters for temporary time.
- ✓ Each time a node access the service with the help of quantum bits, it must be entangled to some states providing services to the users with security.
- ✓ Some protocol must be implemented in ad-hoc network that must be followed for managing information required by different users.

7. CONCLUSION

QuMANET can be implemented by solving all these open problems. This is a very interesting concept including set of ideas that will help to improve speed, security, and facilities provided to the users. Quantum MANET will also help in implementation of advance networks like 6G and more.

REFERENCES

1. University of California - Los Angeles. "Physicists develop world's best quantum bits." ScienceDaily. ScienceDaily, 18 May 2020.
2. University of New South Wales. "Hot qubits break one of the biggest constraints to practical quantum computers." ScienceDaily. ScienceDaily, 15 April 2020.
3. Sharma, R.K.; Sharma, A.K.; Jain, V. Genetic Algorithm-Based Routing Protocol for Energy Efficient Routing in MANETs. In Next-Generation Networks; Lobiyal, D.K., Mansotra, V., Singh, U., Eds.; Springer: Singapore, 2018; Volume 638, pp. 33–40. ISBN 978-981-10-6004-5
4. Kindem, J.M., Ruskuc, A., Bartholomew, J.G. *et al.* Control and single-shot readout of an ion embedded in a nanophotonic cavity. *Nature* (2020).
5. M. A. Nielsen and I. L. Chuang, Quantum Computation and QuantumInformation, 10th ed. Cambridge University Press, 2011.
6. R. J. Lipton and K. W. Regan, Quantum Algorithms via Linear Algebra: A Primer. MIT Press, 2014.
7. Jozsa R, Linden N. On the role of entanglement in quantum computational speed-up. Proc Roy Soc Lond A, 2003, 459: 2011–2032
8. M. Caleffi, A. S. Cacciapuoti, and G. Bianchi, “Quantum internet: from communication to distributed computing!” in Proc. of IEEE/ACM ANOCOM, 2018, invited paper.
9. S. Pirandola and S. L. Braunstein, “Physics: Unite to build a quantum Internet,” Nature, vol. 532, no. 7598, pp. 169–171, Apr. 2016.
10. S. Wehner, D. Elkouss, and R. Hanson, “Quantum internet: A vision for the road ahead,” Science, vol. 362, no. 6412, Oct. 2018.

International Journal of Computer Applications Technology and Research
Volume 9–Issue 08, 240-245, 2020, ISSN:-2319–8656

Document Summarization using Graph Based Methodology

Aditya Jeswani
Student
Dwarkadas J. Sanghvi College of Engineering
Mumbai 400056, India

Shruti More
Student
Dwarkadas J. Sanghvi College of Engineering
Mumbai 400056, India

Kabir Kapoor
Student
Dwarkadas J. Sanghvi College of Engineering
Mumbai 400056, India

Sifat Sheikh
Student
Dwarkadas J. Sanghvi College of Engineering
Mumbai 400056, India

Ramchandra Mangrulkar
Associate Professor
Dwarkadas J. Sanghvi College of Engineering
Mumbai 400056, India

Abstract: This paper works towards constructing a short summary of documents with the help of natural language processing techniques. The authors goal is to identify the important aspects of a large piece of textual information, extract it and present it in a concise manner such that it conveys the information in a more efficiently and precisely. The proposed approach will generate a simple summarization of one or more documents which will help the readers to understand what the documents offer to them and identify their context without reading through them entirely. The existing methods for this work focus on different aspects of the text involved but the efficiency of these methods largely varies. The proposed methodology makes use of a combination of multiple aspects of text instead of a single aspect in order to improve the efficiency of summarization systems. The authors present a qualitative and quantitative analysis of their system as compared to the existing base-lines and demonstrate our system for a relevant application like news snippet generation.

Keywords: Extractive summarization, Multi-document summarization, Key phrase extraction, Shortest path algorithm, Textrank algorithm, GloVe embeddings, Cosine similarity.

1. INTRODUCTION

With the advent of the Internet, there has been a voluminous increase in the amount of data available for humans to read and understand. Going through all of the data is a mammoth task and often required more effort than the value provided by the goal being achieved. Thus, there is a need for a system which can concisely convey all the information that is available in different sources.

Document summarization is one of the most widely researched fields of Natural Language Processing. The task involves using a single or multiple document(s) belonging to a particular domain, understanding the contents of the document and then generating a paragraph which conveys the information in a concise and human readable format. The task of summarization can be carried out in two different ways – extractive and abstractive.

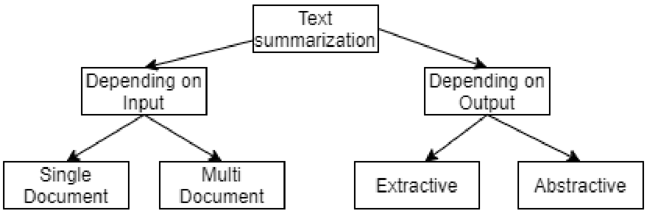


Figure 1. Text Summarization Overview

In extractive document summarization, generation of the summary uses sentences which are present in the document set provided. In this mechanism, the different sentences in the documents are analyzed for their relevance to the main idea of the document cluster, assigned a score and a rank. On the basis of the rank, the top sentences are extracted according to the length required and are then presented to the user as a summary.

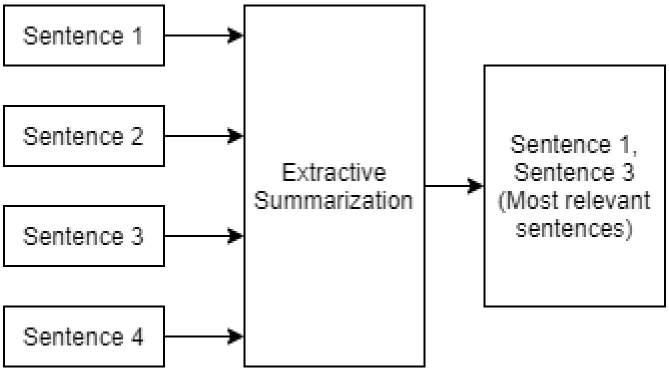


Figure 2. Extractive summarization

On the other hand, an abstractive summarizer works differently. While the initial stages are similar where the document contents are analyzed in order to identify the main idea, this summarizer does not directly pick up sentences from the document. Instead, the model uses the information and knowledge gained in order to generate new sentences on its own to create the summary. The abstractive summarizers more closely resemble how humans generate summaries, by understanding the meaning being conveyed rather than simply picking up sentences from the given documents.

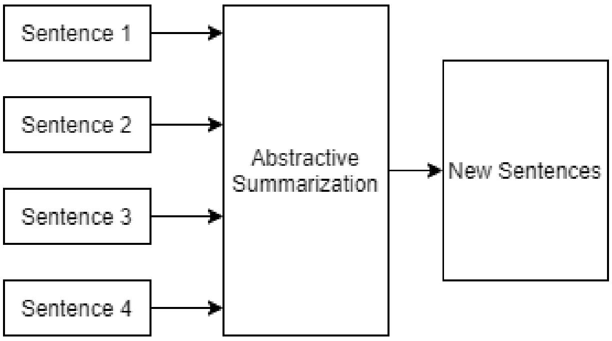


Figure 3. Abstractive Summarization

Graph-based summarization methods have yielded strong and promising results. In these approaches, sentences are treated as nodes and the relationships between them are represented by edge weights. The significance score of each sentence in a document is assumed to be related to other sentences. An edge (or link) with a corresponding weight is created if there is a relation between two sentences, while the weight between sentences in a document is used to provide a score for each sentence. In this paper, the authors propose a technique which combines the uses of sentence relations and the importance of keyphrases to carry out the task of extractive summarization.

2. LITERATURE REVIEW

Akash Ajampura Natesh et al [1] explores an approach different from traditional graph summarization techniques. Since nouns form an important part of the sentence, they create a graph of the available phrases where the nouns form the nodes in the graph. Pronouns in the sentences are assigned name or object references by analyzing the preceding sentence. An edge is added between the 2 nodes if the nouns occur together in the sentence, with the edge weight being the distance between the nouns. Sentence scores are assigned on the basis of noun scores for nouns in the sentence and the top sentences are selected to form the summary for the document. A special feature of their methodology is the use of pronoun resolution to ensure noun occurrences in the sentences, ensuring that a valid score for the sentence can be generated.

Shikhar Sharma et al [2] adopts a different technique for summarizing documents. After breaking the document into individual sentences, these sentences are used to form the graph. A distortion measure based on the "Squared Difference" technique is used to calculate the semantic dissimilarity between the sentences. The dissimilarity is then subtracted from 1 in order to obtain the semantic similarity between the sentences. These values are used to initialize the graph weights in order to avoid random initialization. Once the graph is created, it is passed onto the Textrank algorithm to obtain sentence scores.

Chirantana Malik et al. [3] implements a graph-based approach with a modified Textrank algorithm. Each sentence corresponds to a node in the graph. Modified Cosine similarity is used to give weights to edges which takes into account different levels of importance of words in each sentence. Textrank score is calculated for each node of the graph by considering the average weight of the edges incident to it for giving importance to the weights associated with the edges. Summarization is done here by selecting top 'n' number of sentences based on their Textrank value and then arranging them by their index.

Kang Yang et al. [4] proposes a methodology based on an integrated graph model used along with Textrank. POS tagging is performed for each word, forming word-POS pairs. For context analysis, bigrams and trigrams are constructed. Thus, three separate structures are created- word-POS, bigram and trigram. Then three undirected weighted graphs are built for the sentences in a document which correspond to the three structures constructed earlier. Graphs from different sources are integrated in a Naive Bayesian fashion. Textrank is performed to calculate score of each node or sentence. Sentences with the highest score are selected as per the compression rate.

Hakim et al [5] presents new ways to expand and try to improve graph-based multi-document extractive summarization models by exploring how key phrases can be used in the process of text summarization to produce better summaries. The intuition behind this approach is that the key phrases of a document cluster represent the core ideas and topics of the cluster. Therefore, by taking into account those key phrases, and more specifically, by considering the similarity between those key phrases and the various sentences in the cluster, it can evaluate better that which sentences are the most important.

Erkan et al [6] propose a multi-document extractive summarization system which serves as the baseline model. The centrality score is computed using the LexRank method, then this score is modified to include a different key phrase score that represents the sentence's similarity to the key phrases in the document cluster. The key phrase score is computed using 3 approaches. First using only the key phrases that is equal to the number of phrases present in the given sentence, second using the key phrases equal to the sum of the cosine similarity between the TF-IDF representations of each key phrase (Feinerer [7] et al.); and third, by calculating the key phrase's importance using the scores/ranking provided by the pke package for each key phrase [7]. After computing the final modified score, the author has used ROUGE metric for evaluation.

Jonas et al [8] provides us with a method that results in a smooth summary. Most of the graph-based summarization techniques suffers from sudden topic shifts. This problem could be solved by using Shortest Path Algorithm suggest by the author in the paper 'Extraction based summarization using a shortest path algorithm'. The method first divides the entire documents such that sentences form the nodes of the graph. Costs of edges between nodes are based on number of overlapping words between two sentences, more similar the sentence implies less cost. A special feature of this method is that a node has an edge to its following sentence too, so this might result in smooth summary. For constructing summary, chose a path from the first sentence to last sentence and include all sentences in the path. This method results in smooth summary and summaries of varying length. After computing the model, the author has used ROUGE metric for evaluation.

Madhurima et al [9] suggest a different approach for graph-based summarization. Instead of using Textrank algorithm, the authors use clustering technique. After POS tagging, pronoun resolution and stop word removal, each sentence acts as the node of the graph. Cosine similarity is used to assign weight to the edges between two nodes. After graph construction, clustering coefficient and average clustering coefficient is computed for each node. The special feature of this methodology is applying info map clustering algorithm to partitioning graph into subgraphs and selecting subgraphs having coefficient greater than the average clustering coefficient.

Flourian Boudin et al [10] describes pke, an open source python-based keyphrase extraction toolkit. It provides an end-to-end keyphrase extraction pipeline in which each component can be easily modified or extended to develop new approaches.

3. PROPOSED METHODOLOGY

The approach combines two aspects of making document summaries useful – keyword extraction [5] and graph representation of document contents – to build a Multi-Document extractive summarization model. While both graph summarization and keyword extraction have been implemented in the past, a model which combines the power of both the techniques has not been examined in detail. The complete methodology can be broken down into three stages:

3.1.1 Stage 1: Pre-processing

In the first stage, the model reads the contents of all (or the single) documents(s) which are given by the user, such that the documents belong to the same domain. Once all the input documents are read, the contents of the documents are split from the paragraphs into the corresponding individual sentences. Once the application obtains a list of sentences, they are pre-processed to remove stopwords and resolve pronouns in consecutive sentences.

3.1.2 Stage 2: Graph Builder

Keywords are extracted using an open-source tool, pke. Shortest path algorithm is generated to ensure summary generated is smooth and it also reduces the corpus size. The graph constructed consists of sentences given as output from the shortest path algorithm as the nodes and edge weights as the sum of sentence-sentence similarity and sentence-keyword similarity.

3.1.3 Stage 3: Summary Generation

The weighted graph is passed to the Textrank module to get sentence importance scores and top sentences are extracted to form the summary.

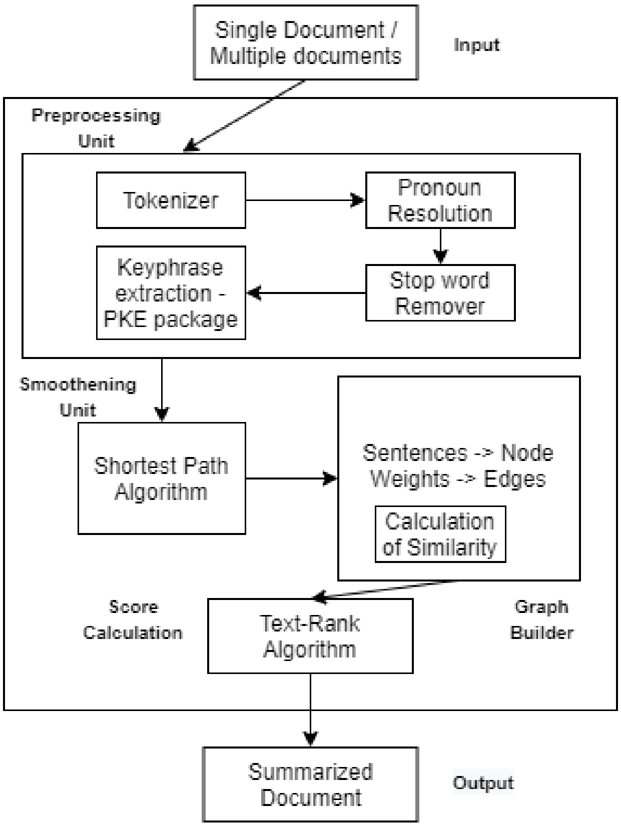


Figure 4. Proposed Model

4. IMPLEMENTATION

The different modules involved right from reading the input files to ranking sentences according to content are described in the following subsections.

4.1 Algorithms Used

4.1.1 TextRank

The TextRank algorithm determines the similarity of each sentence with other sentences in a given text. Based on this, TextRank scores are given to each sentence. Every sentence is stored as a node of a graph. The values are iterated over multiple times until they converge. The sentence with the highest score is the sentence which is the most similar to other sentences. Cosine similarity is used as a similarity measure for TextRank.

4.1.2 Dijkstra's Shortest Path Algorithm

Dijkstra's Shortest Path algorithm finds the shortest path between nodes of a graph. It uses a queue for storing and querying partial solutions sorted by distance from the start. Time complexity of this algorithm with a min-Priority queue is

$$O(|V| + |E| * \log|V|)$$

where |V| is the number of vertices and |E| is the number of edges.

This algorithm is used in the project to smoothen the flow of sentences. It selects the sentences in the sequence in which they were present in the input given by the user.

4.2 Pre-processing

In the first stage, the model reads the contents of all (or the single) document(s) which are given by the user, such that the documents belong to the same domain. A single domain for all the documents ensures that the summary generated is meaningful and comprehensible by the user. Once all the input documents are read, the text from the documents is extracted. The text is then cleaned i.e. unnecessary white spaces and lines are removed. The neuralcoref and spaCy libraries are used to perform pronoun resolution. Anaphoric ambiguities are removed by using pronoun resolution. The text obtained after pronoun resolution is tokenized. The sentences are then processed to remove stopwords which do not contribute to the meaning of the document content. The stopwords used for reference are from the NLTK corpus. Once tokenized text is obtained without stopwords, it is passed to TextRank module.

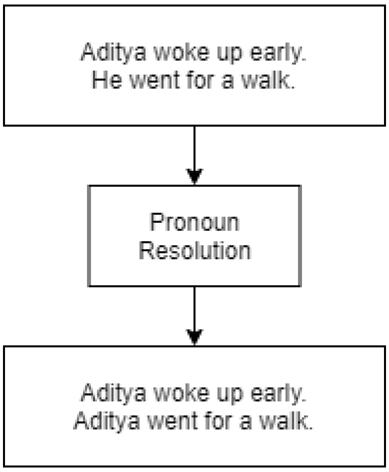


Figure 5. Pronoun Resolution

4.3 Keyphrase Extraction

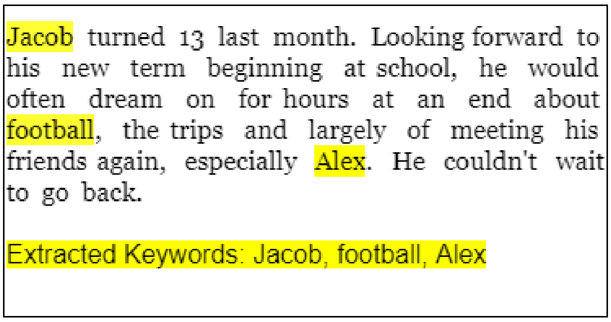
The next step is keyphrase extraction. Keyphrases have an important role in document summarization as they convey the essence of the document with the help of clear, concise and direct words. The extraction is carried out with the help of a pre-trained keyphrase extraction toolkit, pke [10]. It is an open-source toolkit which provides an end-to-end keyphrase extraction pipeline in which each component can be easily modified or extended to develop new models [10]. A similarity matrix is developed by considering sentence-sentence similarity as well as sentence-keyword similarity. To obtain keyphrases, the cleaned text obtained after pronoun resolution is used. Keyphrase extraction is performed using two methods, YAKE and TextRank, with the three best keyphrases being selected in both cases.

4.3.1 YAKE

One implementation is by using YAKE library. The candidate selection is done using bigrams to analyse keyphrases. A one-word window is used to refer to a single word before and after the current word to understand the context. This implementation does not use stemming of words and the stoplist used is the collection of stopwords from NLTK library. The threshold for extraction is set at 0.3 which is required to set a limit for redundant phrases.

4.3.2 TextRank

The second implementation is the TextRank extractor of pke. The sentences are stemmed in this implementation and the window size is set to 3. This implementation uses Parts-Of-Speech tagging to analyse the words in the given text. For selection of the keywords, we analyse only the top 60% of the scored words to limit the processing needed.


Jacob turned 13 last month. Looking forward to his new term beginning at school, he would often dream on for hours at an end about football, the trips and largely of meeting his friends again, especially Alex. He couldn't wait to go back.

Extracted Keywords: Jacob, football, Alex

Figure 6. Keyphrase Extraction

4.4 Graph Building

Once the keyphrases for the document have been extracted, these can be used to evaluate sentence importance and calculate edge weights for the graphs. The sentences which were extracted from the document are represented in the form of a graph structure. The nodes of the graph represent the different sentences obtained and the edge weights are symbolic of the relation between the sentences. The weights are a combination of two aspects – sentence semantic similarity and coherence with keyphrases. The similarity measure used is the cosine similarity which uses a vector representation of sentences and keyphrases. We use GloVE embeddings to vectorize the textual aspects.

4.5 Shortest Path

Simply extracting the important and relevant sentences however does not always guarantee a summary that can be easily understood by the reader. Hence, the authors pass the processed corpus to the Shortest Path module. It helps to reduce the size of the input corpus. It also helps ensure we have a smooth flow from one sentence to another. The graph fed to the algorithm consists of sentences as the nodes with similarity values calculated as edge weights. The authors use Modified Dijkstra's Algorithm, where higher edge weights are considered to find a smooth path from the first to the last sentence. In this manner, we get a set of reduced sentences, which is then passed to the TextRank module.

4.6 TextRank

The sentences obtained from the Shortest Path module are used to re-initialize the graph weights, having an advantage over random initialization for TextRank. After re-building the graph using cosine similarity for edge weights, it is passed to the TextRank module which continuously iterates till convergence to obtain a ranking of sentences according to their importance for summary generation.

4.7 Dataset

For training the model to understand optimal parameters, the authors made use of the DUC 2004 dataset provided by NIST. The data provided consisted of a number of tracks, each track consisting of a cluster of 50 folder and each folder in turn consisted of 8-10 documents. The data used for the purpose of testing and deriving parameters belonged to Task 1 and 2 of the dataset. These tasks had corresponding model summaries written by multiple authors, where every author had written summaries for not all but a few of the document clusters. Hence, the processing of the dataset before any kind of testing required the authors to extract the relevant text data from the documents, prepare it in a format suitable to be parsed and map it to the summary by an author to carry out effective evaluation of the model parameters. Since the end model is unsupervised, the data was required to see how the summaries respond to the tuning of parameters and in turn helped establish the final parameters in the unsupervised model.

5. EXPERIMENTATION, RESULTS AND DISCUSSIONS

5.1 Performance Metric

The authors have used the ROUGE metric for evaluation of the generated summaries. Lin introduced a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to automatically determine the accuracy of a summary by comparing it to a reference summary. Various Rouge metrics for evaluation are as follows:

5.1.1 ROUGE-N

Rouge-N metric is a measure of overlapping words which considers N-grams between a model generated summary and a reference summary. The value of N can be 1 i.e. ROUGE-1 represents unigram overlap between the 2 summaries, a value of N = 2 represents bigram overlap and so on.

5.1.2 ROUGE-L

In this evaluation scheme, the longest common subsequence is identified between the reference and the model generated summary. Rouge-L is more flexible as compared to Rouge-N, but has the drawback that it requires all the N-grams in consecutive positions.

5.2 Evaluation

For evaluating on the proposed summary generation scheme, the authors utilized the DUC 2004 dataset, specifically the data limited to Task 1 and 2. The dataset consisted of a total of 50 document clusters pertaining to a news published in the media. Testing on the dataset involved tuning of different parameters and the results obtained over the 50 sets is as mentioned in the table below.

For evaluation 2 models were used, YAKE and TextRank, provided by PKE. YAKE was the first model to be tried and it yielded the ROUGE values as shown in Table 1.

Table 1. ROUGE values for YAKE keyphrase extraction

	Precision	Recall	F-Measure
ROUGE-1	0.4427	0.1731	0.2472
ROUGE-2	0.0721	0.0275	0.0395
ROUGE-L	0.2207	0.0862	0.1232

Since the values obtained through YAKE were not optimal, the authors implemented keyphrase extraction using PKE’s TextRank model which showed a marked improvement over the previous extraction model. The results obtained as a result are highlighted in Table 2.

Table 2. ROUGE values for TextRank keyphrase extraction

	Precision	Recall	F-Measure
ROUGE-1	0.5138	0.1668	0.2505
ROUGE-2	0.1064	0.0346	0.052
ROUGE-L	0.2626	0.0853	0.1281

The comparison between the 2 models with different parameters is further highlighted through the graphs below. The graphs reflect how the evaluation metric values changed in correspondence to the change in parameters used in the model.

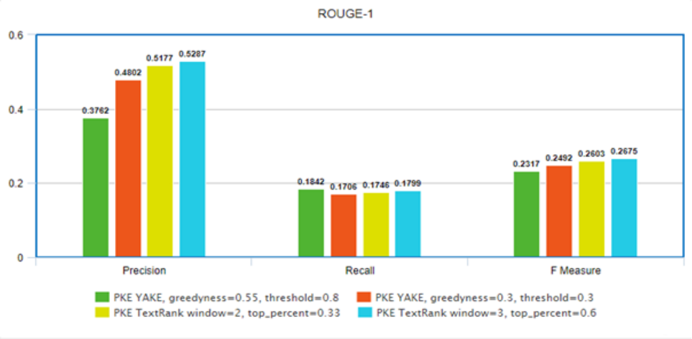


Figure 7. Comparison of ROUGE-1 for YAKE and TextRank

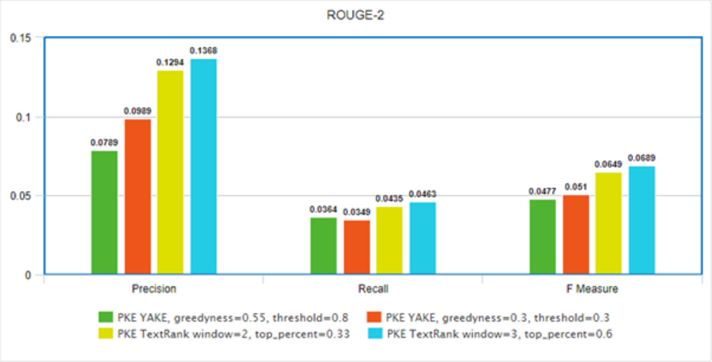


Figure 8. Comparison of ROUGE-2 for YAKE and TextRank

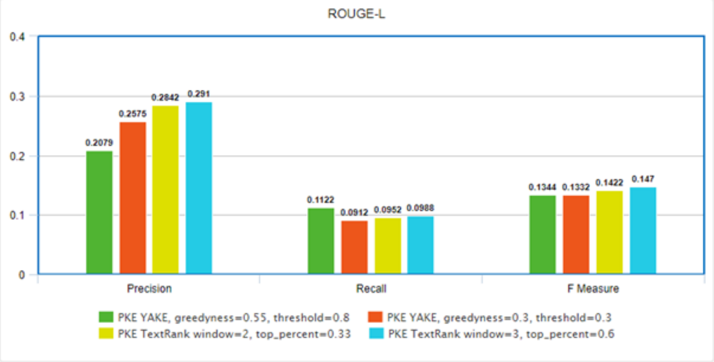


Figure 9. Comparison of ROUGE-L for YAKE and TextRank

6. CONCLUSION AND FUTURE SCOPE

Due to the information overload in the internet, news articles, contents on social Medias, automatic document summarization has received a great deal of attention. The proposed model serves the purpose of summarizing a document or a set of documents in a concise manner. The model combines the advantages of different techniques to generate accurate summaries for different documents belonging to the same domain given by the user. Unlike other text summarizers available, the authors have successfully implemented Shortest Path Algorithm in the model to provide crisp summaries. The combined implementation of TextRank, Keyphrase Generator and Shortest Path Algorithm has helped to achieve greater accuracy in generating concise summaries. The proposed methodology is able to cover up drawbacks of traditional summarization techniques and produce good results.

Automatic summarization evaluation is still a very promising research area with numerous challenges ahead. The project can be extended to include features like reading and writing from PDF's and generating summaries as per user specified lengths. An application of this is convenient text-to-speech for blind people; the idea here is to scan and examine over a page from a book, and then read a summary of the page rather than the entire text. This is an effective way to provide page by page synopsis rather than the whole book. The implemented system can be used to provide summaries in different languages in future. Document Visualization is also another topic for research regarding automatic text summarization. Integration into a document visualization tool can be done to visualize documents or document clusters in a number of ways, including as points on a graph. Moreover, the development of more focused summaries using Abstractive Summarization can be applied to achieve more consistent evaluation and to a better convergence between human and automatic evaluation strategies.

7. ACKNOWLEDGMENTS

The authors would like to express their gratitude towards their guide, Dr. Ramchandra Mangrulkar, the Head of Department, Dr. Meera Narvekar and the Principal, Dr. Hari Vasudevan for the opportunity provided to them in order to carry out implementation and research on this topic.

8. REFERENCES

[1] Akash Ajampura Natesh, Somaiah Thimmaiah Balekuttira and Annapurna P Patil. Graph Based Approach for Automatic Text Summarization in International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Special Issue 2, October 2016.

[2] Agrawal Nitin, Shikhar Sharma, Prashant Sinha, and Shobha Bagai. "A graph based ranking strategy for automated text summarization." DU J. Undergrad. Res. Innov 1, no. 1 (2015).

[3] Mallick, Chirantana, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, and Apurba Sarkar. "Graph-based text summarization using modified TextRank." In Soft Computing in Data Analytics, pp. 137-146. Springer, Singapore, 2019.

[4] Yang, Kang, Kamal Al-Sabahi, Yanmin Xiang, and Zuping Zhang. "An integrated graph model for document summarization." Information 9, no. 9 (2018): 232.

[5] Dunia Hakim, The Role of Key Phrases in Extractive Graph-Based Multi-Document Text Summarization, Stanford University Department of Computer Science dunia@stanford.edu.

[6] Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22:457–479.

[7] Ingo Feinerer and Kurt Hornik. 2017. wordnet: WordNet Interface. R package version 0.1-14.

[8] Jonas Sjobergh, Kenji Araki, "Extraction based summarization using a shortest path algorithm", Proceedings of the Annual Meeting of the Association for Natural Language Processing, 2006.

[9] Dutta M., Das A.K., Mallick C., Sarkar A., Das A.K. (2019) A Graph Based Approach on Extractive Summarization, Emerging Technologies in Data Mining and Information Security: Advances in Intelligent Systems and Computing, vol 813. Springer, Singapore.

[10] Boudin and Florian, pke: an open source python-based keyphrase extraction toolkit, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, December 2016.

[11] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411. 2004.

[12] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics. 28(4):399–408. [1, 2]

Application and Research of Naive Bayes Algorithm in Spam Filtering

Jun Li
School of Software Engineering
Chengdu University of Information Technology
Chengdu,China

Abstract: With the advent of the 5G era, the scope of e-mail applications has become more and more extensive, but the various spam messages that follow have also caused more and more serious problems. Among the many existing methods for filtering spam, the probability-based Bayesian classification algorithm is simple and efficient, and the accuracy rate can reach about 90%. This article briefly introduces the Bayesian model, gives an email filtering method based on the naive Bayesian classification model, and briefly analyzes its advantages and disadvantages. Finally, its effectiveness is verified through experiments.

Keywords: 5G; Email; Spam Filtering; Naive Bayesian Classification Model

1 INTRODUCTION

With the popularization of the Internet, e-mail has been loved by many netizens due to its low price and convenient use. However, it has also caused a large number of spam emails to affect normal communication. According to Kaspersky Lab, in 2019, the third In the quarter, the average proportion of spam in global mail traffic was 56.26%. Among them, the top 5 spam source countries: China ranked first (20.43%), followed by the United States (13.37%) and Russia (5.60%). Fourth place is Brazil (5.14%) and fifth place is France (3.35%). It can be seen that the form of spam processing in my country is still not optimistic.

Many countries have formulated anti-spam laws and regulations, and my country has also formulated relevant legal provisions. However, due to interest-driven, currently spam has not been effectively curbed, but has a growing trend. In addition to national prevention and control, many mail servers use technical methods to filter spam [1], such as adding blacklists, adopting sensitive word filtering rules, and using whitelists. At present, the more popular spam filtering methods include decision tree[2], Boosting[3], K nearest neighbor[4], support vector machine[5], Bayesian principle, etc.[6].

This article mainly introduces the use of Naive Bayes algorithm to filter spam, and combines Adaboost to improve the algorithm [7].

2 NAIVE BAYES MODEL AND RELATED PRINCIPLES

2.1 Bayesian Principle

Bayesian principle is a method proposed by British scholar Bayes as early as the 18th century to apply observed phenomena to correct subjective judgments about probability distribution [8]. The theorem states that the probability of something happening in the future can be estimated by calculating how often it has happened. Using Bayesian algorithm to filter spam, first we prepared 5574 samples, used cross-validation, randomly selected 4574 as training samples, generated a vocabulary list (corpus), tested and calculated the average error rate of classification for 1000 test samples .

2.2 Bayesian Classifier

The naive Bayes classifier uses the "attribute conditional independence hypothesis": assuming that all attributes are independent of each other, based on the attribute conditional independence hypothesis. Assume that the words contained in the content of the email are Wi, Spam, and ham. To judge an email, the word contained in the content is Wi, to judge whether the email is spam, that is, to calculate the conditional probability of P(S|Wi). According to Bayesian formula:

$$\Pr(S|W) = \frac{\Pr(W|S)\cdot\Pr(S)}{\Pr(W|S)\cdot\Pr(S)+\Pr(W|H)\cdot\Pr(H)}$$

among them:

- Pr(S|Wi) The conditional probability that a message with the word Wi appearing is spam (that is, the posterior probability);
- Pr(S) The probability of spam in the mail data set during the training phase, or the probability of spam actually investigated (ie, prior probability);
- Pr(Wi|S) the probability of the word Wi in spam emails;
- Pr(H) The probability of normal mail in the mail data set during the training phase, or the probability of normal mail actually investigated;
- Pr(Wi|H) The probability that the word Wi appears in a normal email;

For all words appearing in the email, considering the independence of each word occurrence event, calculate the joint probability Pr(S|W) of Pr(S|Wi), W={W1, W2,...Wn}:

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1-p_1)(1-p_2)\cdots(1-p_N)}$$

Among them:-P is Pr(S|W), the conditional probability of spam when the word W={W1, W2......Wn} appears;-Pi is Pr(S|Wi), the word Wi appears Is a conditional probability of spam.

3 ALGORITHM IMPROVEMENT

Use Bayesian formula to classify emails, calculate Pr(S|W) and Pr(H|W), compare the size of Pr(S|W) and Pr(H|W), and judge whether it is spam or normal email . We find that Pr(S|W) and Pr(H|W) calculate the same denominator, so we only need to compare the numerators. But there are still two problems: 1. When the vocabulary does not exist, that is, ni=0, at this time Pr(S|Wi) = 0, it will cause P=0, which cannot be compared; 2. When Pr(S| When Wi) is small, multiplying operations will cause underflow problems.

3.1 Solution

To solve these two problems, we have adopted the following solutions: 1. When calculating P(Wi|S) and P(Wi|H), initialize the number of occurrences of all words to 1, and initialize the denominator to 2 (or Adjust the denominator value according to the sample/actual survey results); 2. When calculating P(Wi|S) and P(Wi|H), take the logarithm of the probability.So the final comparison is, P(W1|S)P(W2|S)...P(Wn|S)P(S) and P(W1|H)P(W2|H)....P The size of (Wn|H)P(H).

Test effect: 5574 samples, using cross-validation, randomly selected 4574 as training samples to generate a vocabulary list (corpus), for 1000 test samples, the average error rate of classification is about 2.5%.

3.2 Improve The Algorithm Combined with Adaboost

When we calculate the joint posterior probability of ps and ph, we can introduce an adjustment factor DS, whose function is to adjust the "spamicity" of a word in the vocabulary, where DS is iteratively obtained by the Adaboost algorithm to obtain the best value. The process is as follows:

Step 1 Set the number of adaboost cycles count;

Step 2 Cross validation randomly select 1000 samples;

Step 3 DS is initialized to an all-one vector equal in size to the vocabulary list;

Step 4 Iterate the loop count times:

Step 4.1 Set the minimum classification error rate inf

Step 4.2 For each sample:

Step 4.2.1 Classify the sample under the current DS

Step 4.2.2 If the classification is wrong:

Step 4.2.2.1 Calculate the degree of error, that is, compare the alpha difference between ps and ph

Step 4.2.2.2 If the sample was originally spam, it was classified as ham by mistake:

Step 4.2.2.2.1 DS[Vocabulary contained in sample] = np.abs(DS[Vocabulary contained in sample]-np.exp(alpha) / DS[Vocabulary contained in sample])

Step 4.2.2.3 If the sample was originally ham, it was classified as spam by mistake:

Step 4.2.2.3.1 DS[Vocabulary contained in sample] = DS[Vocabulary contained in sample] + np.exp(alpha) / DS[Vocabulary contained in sample]

Step 4.3 Calculate the error rate

Step 5 Save the minimum error rate and the vocabulary list at this time, P(Wi|S) and P(Wi|H), DS and other information, that is, save the information of the best trained model

Test effect: 5574 samples, get the best model information for Adaboost algorithm training (including vocabulary list, P(Wi|S) and P(Wi|H), DS, etc.), for

1000 test samples, the average error rate of classification Approximately: 0.5%.

4 CONCLUSION

The harm of spam is self-evident, so our treatment of spam is also urgent. This article introduces the application of the Naive Bayes algorithm in spam processing, introduces the content and working principle of the Naive Bayes algorithm model in detail, and finds that underflow and probability of 0 occur during the construction of the algorithm model In response to the discovered problems, we found out the corresponding solutions, and finally implemented the coding combined with the improved Adaboost algorithm to greatly reduce the average error rate of spam classification.

5 REFERENCES

[1] Yang Shan, He Yue, Yan Jinjiang. Discussion on anti-spam technology based on Bayesian [J]. Network Security Technology and Application, 2007 (08): 54-56.

[2] Zhang Fuzhi, Wu Chaohui, Yao Fang, et al. Research and improvement of spam filtering technology based on Bayesian algorithm [J]. Journal of Yanshan University, 2009, 33(1): 47-52.

[3] Wang Qingsong, Wei Ruyu. Phrase-based Bayesian Chinese spam filtering method [J]. Computer Science, 2016, 43 (04): 256-259+269.

[4] Qu Meiting, Yu Jingxiao, Bu Wei, et al. Interactive architectural design model based on semantically guided Bayesian algorithm[J]. Bulletin of Science and Technology, 2016, 32(5): 133-136.

[5] Zheng Wei, Shen Wen, Zhang Yingpeng, etc. Research on Spam Filter Based on Improved Naive Bayes Algorithm [J]. Journal of Northwestern Polytechnical University, 2010, 28 (4): 622-627.

[6] ZHAN Chuan，LU Xianliang，ZHOU Xu，et al. Spam filtering method based on Bayesian formula［J］. Computer Science，2005，2（32）：73-75.

[7] Cao Ying, Miao Qiguang, Liu Jiachen, et al. Research progress and prospect of AdaBoost algorithm ［J］. Acta Automatica Sinica, 2013, 39(6): 745-758.

[8] Ma Xiaolong. Research on an improved Bayesian algorithm in spam filtering[J]. Application Research of Computers, 2012, 29(3): 1091-1094.

Android Based Dictionary Application of Agricultural Terms in Bali

Dwi Putra Githa
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Desy Purnami Singgih Putri
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

IGD. Gita Purnama Arsa Putra
Department of Balinese Language
Faculty of Arts
Udayana University
Denpasar, Bali, Indonesia

Abstract: Bali is one of the islands in Indonesia which is famous for its natural and cultural beauty. Cultural preservation, especially in Bali, is needed to keep the ancestral heritage and Balinese identity. Culture in agriculture is one of the cultures in Bali that must be preserved. The agricultural terms in Bali is diverse and has its own uniqueness. One way to preserve culture in agriculture is to develop a dictionary application for the agricultural terms in Bali based on Android. It is hoped that this application can introduce and preserve culture especially in agriculture in Bali to the general public. Considering the technological developments that most people already use smartphones, an android based application is developed so that it can facilitate access to applications via a smartphone.

Keywords: Android, Culture, Cultural Conservation, Dictionary, Agricultural Terms.

1. INTRODUCTION

Culture is a complex which includes knowledge, beliefs, art, morals, customs and other abilities and habits possessed by humans as part of society [1]. Cultural preservation, especially in Bali, is needed to maintain the ancestral heritage and identity of Bali and the Indonesian people more broadly.

Culture in agriculture is one of the cultures in Bali that must be preserved. Bali is an island in Indonesia where most of the land is used in agriculture. According to data from the Central Statistics Agency Bali, 2018 regarding land area according to its use in the Province of Bali in 2017, the area of agricultural land in Bali is 407534 ha, 72% of the total land in Bali [2]. The terms in agriculture in Bali are diverse and have their own uniqueness. One way to preserve culture in agriculture is to develop an Android-based dictionary of agricultural terms in Bali.

Android is an operating system based on Linux for cellular phones such as smartphones and tablet computers. Android provides an open platform for developers to create their own applications for use by a variety of mobile devices. At present most smartphone vendors have produced Android-based smartphones, including HTC, Motorola, Samsung, LG, Sony Ericsson, Acer, Nexus, Oppo, and Vivo.

From the above background, research was conducted to design and build a dictionary of agricultural terms in Bali based on Android. It is expected that this application can introduce and preserve culture, especially in agriculture in Bali to the general public. Considering the development of technology that most of the people already use smartphones, an android-based application was developed to facilitate access to application.

2. LITERATURE REVIEWS

Based on previous research related to the Android-based dictionary application from a journal entitled "Android-based Mobile Application Dictionary of Computer Terms" in 2014 compiled by Herlan Mulyana and Maimunah stated that the Computer Glossary Dictionary was created because the need for information is very important and a difficult time when having to look for meaning words or terms using conventional print dictionaries [3]. The 2016 "Android Based Geography Dictionary Application Research" compiled by Winda Yormala and Kurnia Setiawati created an application that has the term geography, equipped with solar system information, exercises according to solar system material and admin to update dictionaries and materials [4].

3. RESEARCH METHODS

3.1 Research Flow

Figure 1 is the stage of the research conducted. The first stage of this research is defining the problems want to solve. After defining the problems, the next step is to collect data to support the resolution of the existing problems. After the required data is collected, the data is analyzed as a basis for making an application. The stages of making an application consist of designing databases, interfaces, and making program code. The next step is to inputing sample data to test the system. If the system produces outputs that are not as expected, data analysis will be performed again. If the results are as expected, the research phase has been completed.

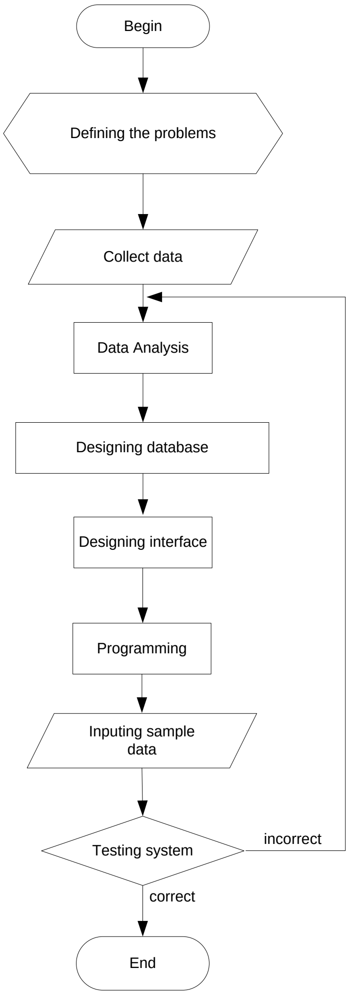


Figure. 1 Research flow

3.2 Overview

Figure 2 is a overview of the system being built. The initial input of the system is the agricultural terms in Bali, then the system will request the agricultural term data in the database. Data about the agricultural term requested by the user will be forwarded to the dictionary application of agricultural terms in Bali. Information on agricultural terms in Bali is accepted by users in Indonesian and English.

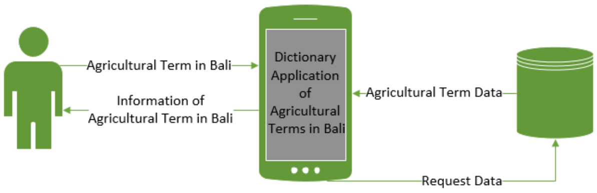

Agricultural Term in Bali
Information of Agricultural Term in Bali
Dictionary Application of Agricultural Terms in Bali
Agricultural Term Data
Request Data

Figure. 2 Overview system

4. CONCEPTS AND THEORIES

This section contains concepts and theories that support theresearch. They are including Android, Android Studio, PHP, and MySQL.

4.1 Android

Android is a developed operating system for Linux-based mobile devices such as smart phones and tablet computers [5]. Android includes an operating system, middleware, and applications. Android Inc. is a name of a company engaged in the world of information and communication technology, the company was bought by a giant company, namely Google Inc. and the Handset Alliance was formed, a consortium of 34 hardware, software and telecommunications companies including: Google, HTC, Intel, Motorola, Qualcomm, T-Mobile and Nvidia.

4.2 Android Studio

Android Studio is the official Integrated Development Environment (IDE) for Android app development, based on IntelliJ IDEA . On top of IntelliJ's powerful code editor and developer tools, Android Studio offers even more features that enhance your productivity when building Android apps.[6]

4.3 PHP

PHP is a scripting language that can be embedded or inserted into HTML. PHP is widely used for dynamic website programming [7]. PHP stands for PHP hypertext preprocessor which is used as a server-side script language in web development that is inserted in an html document [8]. The use of PHP allows the web to be made dynamic so that the maintenance of the website becomes easier and more efficient.

4.4 MySQL

MySQL is an open source DBMS (Database Management System). The advantage of MySQL is that the database can work on various platforms and is easy to access [9]. MySQL uses the standard query language SQL (Structure Query Language).

5. RESULT AND IMPLEMENTATION

5.1 Requirements Analysis

Requirement analysis is carried out to determine the functional requirements of the Dictionary of Agricultural Terms in Bali. The following in Table 1 are system requirements.

Table 1. System requirements

No.	Requirements
1	The system can display a list of agricultural terms.
2	The system can display information and pictures / photos (if any) about agricultural terms.
3	The system can search agricultural terms according to user input.
4	The system can provide a system admin menu to manage data in agricultural terms.
5	The system can validate the system admin username and password.

5.2 Data

Based on observations and documentation, 200 data on agricultural terms in Bali are obtained, Table 2 is an example of agricultural term data in Bali.

Table 2. Example of agricultural terms data in Bali

No.	Terms	Description
1.	**Abangan/ Blangu**	Conduit made of bamboo, areca nut or areca palm tree trunks; function to channel water from one place to another; tree trunks used depend on the volume of water, the greater the volume the greater the trunk is used.
2.	**Aét**	Clue when people plow so the cow turns left.
3.	**Andog**	Fertile soil; usually still flowing water, muddy, and rarely dry.
4.	**Andungan**	The first water channel from a water source (Temuku) that flows through the fields.
5.	**Anggapan**	Knives for cutting rice or cutting during harvest; flat shaped; the size of a hand grip; the edges are filled with wood or bamboo which functions to hold the tool.
6.	**Anggas**	The name of the poison grasshopper; thorns or barriers that are placed in such a way as to close access and protect the coconut or other plants from being stolen.
7.	**Bangkil**	Tools used to harvest rice in fields.
8.	**Camok**	Mouth cover animals such as cow or buffalo.
9.	**Dapak**	Cutting tools for branches, wood, twigs, and roots; made of iron
10.	**Gabag**	Agricultural tools in the form of rakes that are used in rice fields, made of wood, function to destroy or crush the soil to become smooth, as well as leveling the soil so that water can be flooded evenly on each plot of paddy.

5.3 Use Case Diagram

The Dictionary of Agricultural Terms in Bali involves two actors: system administrator and user. Application functionality consists of login, managing data on agricultural terms, displaying a list of agricultural terms, searching for agricultural terms, and displaying the information of agricultural terms. The use case diagram can be seen in Figure 3.

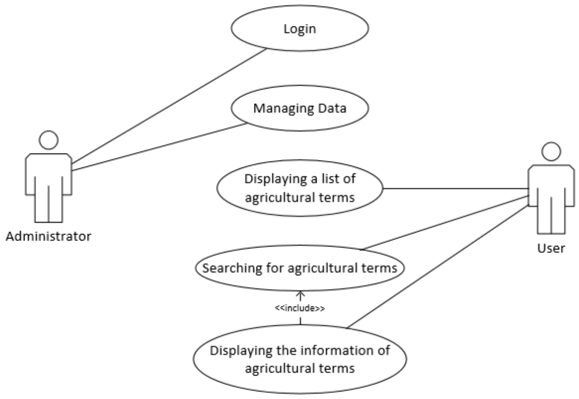

Login
Managing Data
Displaying a list of agricultural terms
Administrator
User
Searching for agricultural terms
<<include>>
Displaying the information of agricultural terms

Figure. 3 Use case diagram

5.4 Database Design

The Dictionary of Agriculture Terms in Bali uses a table in the database used. The table is used to store data on the name of the term, the description of the term in Indonesian, the term description in English, the picture of the term and the publish status of the term. The table structure used in the database can be seen in Table 3.

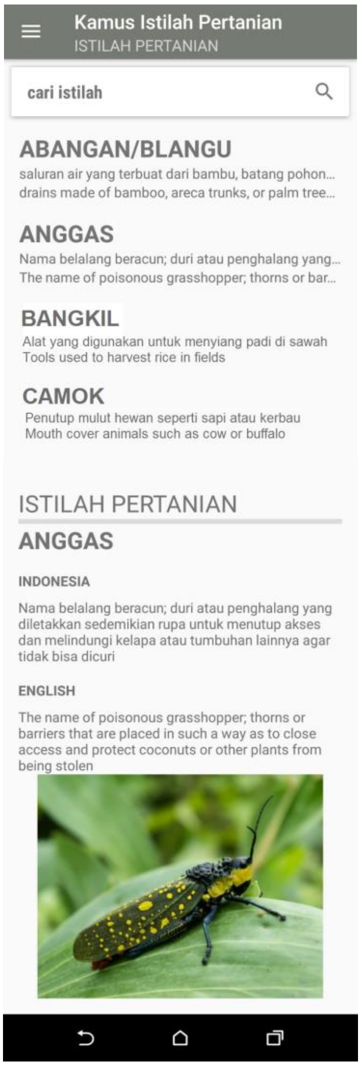
Table 3. Table structure

Field	Data Type
id	int
bali_word	varchar(50)
ina_desc	text
eng_desc	text
img	varchar(50)
is_publish	tinyint

5.5 System Interface

5.5.1 User

The application can be accessed by users using smartphones that use the Android operating system. The user application display can be seen in Figure 4. The user in using the application chooses one of the terms available to see the description and picture of the agricultural term. Users can also search for agricultural terms by typing keywords in the search field.



ABANGAN/BLANGU
saluran air yang terbuat dari bambu, batang pohon...
drains made of bamboo, areca trunks, or palm tree...

ANGGAS
Nama belalang beracun; duri atau penghalang yang...
The name of poisonous grasshopper; thorns or bar...

BANGKIL
Alat yang digunakan untuk menyiang padi di sawah
Tools used to harvest rice in fields

CAMOK
Penutup mulut hewan seperti sapi atau kerbau
Mouth cover animals such as cow or buffalo

ISTILAH PERTANIAN

ANGGAS

INDONESIA

Nama belalang beracun; duri atau penghalang yang diletakkan sedemikian rupa untuk menutup akses dan melindungi kelapa atau tumbuhan lainnya agar tidak bisa dicuri

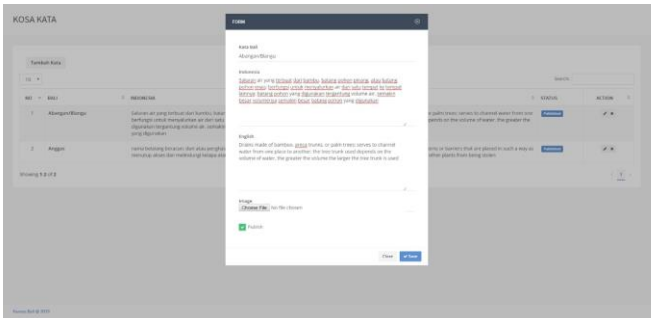
ENGLISH

The name of poisonous grasshopper; thorns or barriers that are placed in such a way as to close access and protect coconuts or other plants from being stolen

Figure. 4 User interface

5.5.2 *Administrator*

The process of updating data can be done by the system administrator using the administrator menu. On the administrator menu, the administrator can add the agricultural terms data, change the agricultural terms data and delete the agricultural terms data. Data will be synchronized with data on the Android application. The administrator menu is created using PHP so it is web based. The administrator display menu can be seen in Figure 5.



Kata Bali
Abangan/Blangu

Indonesia
Saluran air yang terbuat dari bambu, batang pohon pinang, atau batang pohon enau berfungsi untuk menyalurkan air dari satu tempat ke tempat lainnya. batang pohon yang digunakan tergantung volume air. semakin besar volumenya semakin besar batang pohon yang digunakan

English
Drains made of bamboo, areca trunks, or palm trees serves to channel water from one place to another. the tree trunk used depends on the volume of water. the greater the volume the larger the tree trunk is used

Image
Choose File No file chosen

☑ Publish

NO	BALI	INDONESIA	STATUS
1	Abangan/Blangu	[illegible]	[illegible]
2	Anggas	[illegible]	[illegible]

Showing 1-2 of 2

Figure. 5 Administrator interface

5.6 Black Box Testing

Black box testing is done by trying the functions / menus provided by the application. Details of the test can be seen in Table 4.

Table 4. Black box testing

No.	Functional	Scenario	Result	Conclusion
1.	*Login*	User enters correct username and password.	User successfully logged in as administrator.	Correct
2.	*Login*	User entered wrong username or password.	The system displays an incorrect username or password message. The user failed to log on as an administrator.	Correct
3.	Displaying a list of agricultural terms.	The user enters the application.	The system displays a list of agricultural terms.	Correct
4.	Displaying information about the explanation of agricultural terms.	The user chooses one agricultural term.	The system displays information on selected agricultural terms in Indonesian and English.	Correct
5.	Searching for agricultural terms.	The user enters the term he wants to search for.	The system displays the agricultural term sought.	Correct
6.	Adding agricultural term data.	Administrator clicks the 'Tambah Kata' button, then enters all data fields and clicks the 'Save' button.	The system stores data on agricultural terms according to what the administrator input into the database.	Correct
7.	Changing the agricultural term data.	The administrator clicks the pencil marked button on one of the agricultural terms, then changes the data and clicks the 'Save'	The system changes the agricultural term data in the database.	Correct

		button.		
8.	Deleting agricultural term data.	The administrator clicks the button marked with a cross on one of the data agricultural terms.	The system deletes agricultural term data from the database.	Correct

6. CONCLUSION

The design and build of an Android-based Dictionary of Agricultural Terms in Bali has been successfully implemented. All application functionality is running as expected. Application design and development begins with analyzing application requirements, then making application designs using use case diagrams, making database designs according to application requirements and creating interface designs. After the design is finished, proceed with making the program code, then entering the sample data and testing all the functionalities of the agricultural term dictionary application in Bali.

7. REFERENCES

[1] Hawkins, P. 2012. Creating a Coaching Culture. New York: Open University Press.

[2] https://bali.bps.go.id/statictable/2018/04/11/72/luas-lahan-per-kabupaten-kota-menurut-penggunaannya-di-provinsi-bali-2017.html, diakses pada tanggal 1 Maret 2019.

[3] Mulyana, Herlan & Maimunah. 2014. Aplikasi Mobile Kamus Istilah Komputer Berbasis Android. Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, Vol. 1, No.2, Hal. 27-34.

[4] Yormala, Winda & Setiawati, Kurnia. 2016. Perancangan Aplikasi Kamus Geografi Berbasis Android. Jurnal TEKNOIF, Vol. 4, No. 1, Hal. 48-56.

[5] I W.A. Krisna, I N. Piarsa, and P. W. Buana. 2019. Android-Based High School Management Information System. International Journal of Computer Applications Technology and Research, Volume 8–Issue 11, 415-419.

[6] https://developer.android.com/studio/intro, diakses pada tanggal 1 Agustus 2020.

[7] Pursana, P. E. 2014. Sistem Informasi Koperasi Modul Simpanan Berbasis Android Terintegrasi Berbasis Web. Merpati, 2(1), 67-78.

[8] Peranginangin, Kasiman. 2006. Aplikasi WEB dengan PHP dan MySQL. Yogyakarta: Andi Offset.

[9] A. Hanafi, I. M. Sukarsa, and A. A. K. A. C. Wiranatha. 2017. Pertukaran Data Antar Database dengan Menggunakan Teknologi API. Lontar Komputer, vol. 8, no. 1, pp. 22–30.