A

Project Report

on

INFLUENCE MAXIMIZATION OF VIRAL MARKETING ON SOCIAL NETWORK

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Sanchita Wagh
Hemangini Chaudhari
Juili Kulkarni
Payal Jagtap

Under the Guidance of

Mr. Nitin Y. Suryawanshi

DEPARTMENT OF COMPUTER ENGINEERING
SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)
2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the project entitled *Influence maximization of viral marketing on social network*, submitted by

Sanchita Wagh
Hemangini Chaudhari
Juili Kulkarni
Payal Jagtap

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: April 8, 2017
Place: Jalgaon

Mr. Nitin Y. Suryawanshi
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

We would like to express our deep gratitude and sincere thanks to all who helped us to complete this project work successfully. Our sincere thanks to our principal Prof. Dr. K. S. Wani. Our gratitude goes to our Head of the Department, Prof. Dr. G. K. Patnaik for granting us opportunity to conduct this project work. We sincerely thankful to our project guide, Mr. Nitin Y. Suryawanshi and project in charge Mr. Aakash Waghmare, for their valuable suggestions and guidance at the time of need. Also we would like to thanks our parents for always encouraging us.

Sanchita Wagh
Hemangini Chaudhari
Juili Kulkarni
Payal Jagtap

Contents

List of Figures

Abstract

Influence maximization is introduced to maximize the profit of viral marketing in social networks.Impact amplification is used to augment the benefit of viral promoting in informal organizations. The shortcoming of impact expansion is that it doesn't recognize particular clients from others, regardless of the possibility that a some things might be helpful for the particular clients. For such things, it is a superior system to concentrate on boosting the impact on the particular clients. The detail impact boost issue as question handling to recognize particular clients from others is given. We formulate an influence maximization problem as query processing to distinguish users which are specific from other users. The problem of query processing is NP-hard and its objective function is submodular.We propose an expectation model for the value of the objective function and expectation model is used for fast greedy-based approximation method. For detail of expectation model, we investigate a relationship of paths between users. We work out an efficient incremental updating of the marginal gain to our objective function for the greedy method. We propose a desire model for the estimation of the target capacity and a quick covetous based and the desire model is utilized by close estimation strategy. For detail of desire model, we explore a relationship of ways between clients. Their lead trials to the genuine datasets which are assess the proposed technique, and contrast the outcomes and those of existing systems that are adjusted to the issue.The measure of website effectiveness should be the satisfaction of the users rather than that of the developers.Thus, Web pages should be organized in a way that generally matches the users model of how pages should be organized.

Chapter 1

Introduction

The advent of the Internet has provided an unprecedented platform for people to acquire knowledge and explore information. There are 3.77 billion Internet users in the world. The fast-growing number of Internet users also presents huge business opportunities to firms.In order to satisfy the increasing demands from online customers, firms are heavily investing in the development and maintenance of their websites.

Despite the heavy and increasing investments in website design, it is still revealed, however, that finding desired information in a website is not easy and designing effective websites is not a trivial task. Galletta et al. indicate that online sales lag far behind those of brick and mortar stores and at least part of the gap might be explained by a major difficulty users encounter when browsing online stores. Palmer highlights that poor website design has been a key element in a number of high profile site failures. McKinney etc. also find that users having difficulty in locating the targets are very likely to leave a website even if its information is of high quality.

A primary cause of poor website design is that the web developers understanding of how a website should be structured can be considerably different from those of the users. Such differences result in cases where users cannot easily locate the desired information in a website. This problem is difficult to avoid because when creating a website, web developers may not have a clear understanding of users preferences and can only organize pages based on their own judgments. However, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Thus, Webpages should be organized in a way that generally matches the users model of how pages should be organized. Previous studies on website has focused on a variety of issues, such as understanding web structures, finding relevant pages of a given page, mining informative structure of a news website, and extracting template from webpages.

Our work, on the other hand, is closely related to the literature that examines how to improve website navigability through the use of user navigation data.Various works have made an effort to address this question and they can be generally classified into two categories : to facilitate a particular user by dynamically reconstituting pages based on his profile and traversal paths, often referred as personalization, and to modify the site structure to ease the navigation for all users, often referred as transformation.we are concerned primarily with transformation approaches. The literature considering transformations approaches mainly focuses on developing methods to completely reorganize the link structure of a website. Although there are advocates for website reorganization approaches, their drawbacks are obvious. First, since a complete reorganization could radically change the location of familiar items, the new website may disorient users . Second, the reorganized website structure is highly unpredictable, and the cost of disorienting users after the changes remains unanalyzed. This is because a websites structure is typically designed by experts and bears business or organizational logic, but this logic may no longer exist in the new structure when the website is completely reorganized. Besides, no prior studies have assessed the usability of a completely reorganized website, leading to doubts on the applicability of the reorganization approaches. Finally, since website reorganization approaches could dramatically change the current structure, they cannot be frequently performed to improve the navigability.

Recognizing the drawbacks of website reorganization approaches, we address the question of how to improve the structure of a website rather than reorganize it substantially Specifically, we develop a mathematical programming (MP) model that facilitates user navigation on a website with minimal changes to its current structure. Our model is particularly appropriate for informational websites whose contents are static and relatively stable over time. Examples of organizations that have informational websites are universities, tourist attractions, hospitals, federal agencies, and sports organizations. Our model, however, may not be appropriate for websites that purely use dynamic pages or have volatile contents.This is because a steady state might never be reached in user access patterns in such websites, so it may not be possible to use the weblog data to improve the site structure.

This chapter is elaborated in following sections: Section 1.1 describes Background of the project. The Motivation of the project is described in Section 1.2. Section 1.3 describes Problem Definition of project. The Scope of project is described in Section 1.4. Section 1.5 describes the Objective of the project. The Organization of the report is described in Section 1.6. Finally, section 1.7 contains Summary.

1.1 Background

Viral marketing is one of the key applications of influence maximization. In viral marketing, an item that a marketer wants to promote is diffused into social networks by word-of-mouth communication. From the perspective of marketing, influence maximization provides how to get the maximum profit from all the users in a social network through viral marketing. However, influence maximization is not always the most effective strategy for viral marketing, because there can be some items that are useful to only specific users. These specific users can be a few people with a common interest in a given item, some or all people in a community, or some or all users in a class. There is no limit for being specific users. For example, consider a marketer that is asked to promote a cosmetic product for women through viral marketing. For the cosmetic product, the specific users are female users who are likely to use it and male users who wish to purchase it as a gift for female users. In this case, the marketer does not need to be concerned about the other users because the cosmetic product is not useful to them. Instead, it is a better strategy to focus on maximizing the number of influenced specific users, but influence maximization has the weakness that it cannot distinguish them from the other users. The number of outward links in a page, the outdegree, is an important factor in modeling web structure. Prior studies typically model it as hard constraints so that pages in the new structure cannot have more links than a specified out-degree threshold,because having too many links in a page can cause information overload to users and is considered undesirable. For instance, Lin uses 6, 8, and 10 as the out-degree threshold in experiments. This modeling approach, however, enforces severe restrictions on the new structure, as it prohibits pages from having more links than a specified threshold, even if adding these links may greatly facilitate user navigation. Our model formulates the out-degree as a cost term in the objective function to penalize pages that have more links than the threshold, so a pages out-degree may exceed the threshold if the cost of adding such links can be justified.

1.2 Motivation

The idea behind viral marketing is that by targeting the most influential users in the network we can activate a chain reaction of influence driven by word-of-mouth, in such a way that with a very small marketing cost we can actually reach a very large portion of the network. Suppose we are given a social network, that is a graph whose nodes are users and links represent social relations among the users. Suppose we are also given the estimates of reciprocal influence between individuals connected in the network, and suppose that we want to push a new product in the market. The mining problem of influence maximization is the following: given such a network with influence estimates, how should one select the set of initial users so that they eventually influence the largest number of users in the social network. This problem has received a good deal of attention by the data mining.

1.3 Problem Definition

Influence maximization problem is query processing which are used to differ particular users from other users. So, for that web personalization is used. The aim of influence maximization is to increase the profit of viral promoting in social networks.But the influence maximization not differ particular user from other users, even if some of items is focused by influence maximization strategy which is focus in this specific user.

Develop a web portal application for influence maximization of viral marketing.Users should be able to search for required information on the portal.Users can register with the portal to get further facilities.Multiple users visiting number of commercial sites these all data is collected in the databases,it can be in the form of logs,cookies,history etc...So, in the influence maximization process these all database logs are sorted and provide Ads, News, information etc. as per their behaviour and Influence diffusion model and target aware viral marketing. First, there are target users who have an interest in item. Second, there are non-target users who can be influence for the item to introduce it to their friends.Finally, there are non-target users who are immune to being influence for the item, because they do not want to introduce to their friends.

The proposed system is going to use data sets and whatever experiments is to be done that is evaluated into datasets and compare these experiment with existing method and from these experimental results.The proposed system is faster than the existing systems also have a high accuracy than the existing methods. Also, the existing methods, suppose even if some items are useful to that particular users still they does not differ that particular users from others.The proposed expectation model which is able to maximizing the influence on the specific users and also it can easily differ particular users from others.

It is easy to see that the influence diffusion model can handle the first case and the second case. However, the influence diffusion model cannot handle the third case, because it does not distinguish such immune nodes from the others.Nevertheless, we can easily modify the influence diffusion model to support the third case by adding one condition to it.Thus, for simplicity we stick to the original influence diffusion model to explain the proposed method.The proposed expectation model which is able to maximizing the influence on the specific users and also it can easily differ particular users from others.

We perform extensive experiments on a data set collected from a real website. The results indicate that our model can significantly improve the site structure with only few changes. Besides, the optimal solutions of the MP model are effectively obtained, suggesting that our model is practical to real-world websites. We also test our model with synthetic data sets that are considerably larger than the real data set and other data sets tested in previous studies addressing website reorganization problem. The solution times are remarkably low for all cases tested, ranging from a fraction of second to up to 34 seconds. Moreover, the solution times are shown to increase reasonably with the size of the website, indicating that the proposed MP model can be easily scaled to a large extent.

1.4 Scope

The scope of this project is to overcome the weakness of influence maximization in viral marketing. It is a better strategy to focus on maximizing the number of influenced specific users, but influence maximization has the weakness that it cannot distinguish them from the other users.Online social networks are increasingly being recognized as an important source of information influencing the adoption and use of products and services. Viral marketing is the tactic of creating a process where interested people can market to each otheris therefore emerging as an important means to spread-the-word and stimulate the trial,adoption,and use of products and services. Consider the case of Hotmail, one of the earliest firms to tap the potential of viral marketing. Based predominantly on publicity from word-of-mouse, the Web-based email service provider garnered one million registered subscribers in its first six months, hit two million subscribers two months later, and passed the eleven million mark in eighteen months.Wired magazine put this growth in perspective in its December 1998 issue: The Hotmail user base grew faster than [that of] any media company in historyfaster than CNN, faster than AOL, even faster than Seinfelds audience. By mid-2000, Hotmail had over 66 million users with 270,000 new accounts being established each day.

1.5 Objective

The aim of Influence maximization is to increase the profit of viral promoting in social networks. But Influence maximization not differ particular user from other users, even if some of the items are useful for that particular user. So these few types of item is focused by influence maximization strategy which is focused on these specific users. Influence maximization problem is a query processing which are used to differ particular users from other users.

To assess the user navigation on the improved website, we partition the entire real data set into training and testing sets. We use the training data to generate improved structures which are evaluated on the testing data using simulations to approximate the real usage. We define two metrics and use them to assess whether user navigation is indeed enhanced on the improved structure. Particularly, the first metric measures whether the average user navigation is facilitated in the improved website, and the second metric measures how many users can benefit from the improved structure. Evaluation results confirm that user navigation on the improved website is greatly enhanced.

First, we explore the problem of improving user navigation on a website with minimal changes to the current structure, an important question that has never been examined in the literature. We show that our MP model not only successfully accomplishes the task but also generates the optimal solutions surprisingly fast. The experiments on synthetic data indicate that our model also scales up very well. Second, we model the out-degree as a cost term in the objective function instead of as hard constraints. This allows a page to have more links than the out-degree threshold if the cost is reasonable and hence offers a good balance between minimizing changes to a website and reducing information overload to users. Third, we propose two evaluation metrics and use them to assess the improved structure to confirm the validity of our model. The evaluation procedure developed in this method provides a framework for evaluating website structures in similar studies.

1.6 Organization of Report

Chapter 1 describes the background of project with its scope and problem definition. Chapter 2 describes system analysis. Chapter 3 describes the algorithms implemented in this project. Chapter 4 describes system requirements specification which includes software, hardware, functional and non functional requirements. Chapter 5 describes system design with the help of various unified modeling language diagrams. chapter 6 describes implementation details and flow of the system development. chapter 7 describes system testing, how to implement it and some of the test cases. Results and analysis are given in chapter 8. Finally, chapter 9 contains conclusion and summary.

1.7 Summary

This chapter covers the Introduction of the project, Problem definition, Scope and Objective of the project. In next chapter, the system analysis is described.

Chapter 2

System Analysis

System analysis is the process of gathering and interpreting facts, diagnosing problems and using the facts to improve the system.System analysis is a process of examining business situation for the purpose of developing system solution to a problem or devising a improvement in such situation. Before the development of any system,a project proposal is prepared by the user of the system or system analyst and submitted to an appropriate managerial structure in the organisation.

This chapter is elaborated in following sections: Section 2.1 describes the Literature Survey. Proposed System is discussed in Section 2.2. Section 2.3 describes Feasibility study. Project Scheduling is described in Section 2.4. Section 2.5 describes Effort Allocation. Finally, section 2.6 contains Summary.

2.1 Literature Survey

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

1. DATA DISSEMINATION: Dissemination is the release to users of information obtained through a statistical activity. Data dissemination consists of distributing or transmitting statistical data to users. Various release media are possible. For example: electronic format including the internet, CD-ROM, paper publications, files available to authorised users or for public use; fax response to a special request, public speeches, press releases.

We highlight two factors that play a key role in determining the nature of influence episodes in viral marketing. The first is the role of the influencerwhether the attempt to influence is passive or actively persuasive. The second is the level of network externalities the additional benefits accruing from broader usage of the product or service being recommended within a user community.There are four types of role of the influencer:

1. Awareness Creation and Benefits Signaling (ACBS): The role of the influencer in persuasion is passive and the network externalities are minimal. Users emailing online greeting cards from Web sites such as Hallmark or Blue Mountain to connected others represent typical instances of ACBS. When a user sends out a card from the site, the recipients get a personalized email message informing them of a greeting created by the sender available at the site and providing the URL to access it. The URL directs visitors to the card on site and once there, he or she is offered the choice to send a greeting to the original sender or to a connected other. In this process, recipients are made aware of the service offered by the site and are persuaded to use it. The role of the influencer is mainly to create awareness and signal benefits to others within their social network and can be particularly influential in encouraging trial and adoption of novel products and services.

2. Targeted Recommendation: The influencer plays an active role in spreading the word and the network externalities are minimal. For example, Honda Motor Europe unveiled the Honda Civic with a lifestyle-oriented campaign Live the Civic at the Paris Motor Show in September 2000. Honda used a viral marketing technique to sustain customer interest and gather data on potential buyers for the car.A set of amusing video clips playable over the Internet were distributed to 300 influencers to kick off a contest where participants chances of winning increased if they forwarded the videos to others. Reports suggest that this was responsible for nearly half a million individuals visiting the Honda Web site for the new car in the first three months with the video clips being forwarded almost 80,000 times.The viral marketing campaign is also credited with generating over 10,000 promising leads for the new car.

3. Signaling Use, Group Membership (SGM): The influencers role is passive but there are significant externalities accruing to both the recipient and the influencer. Instances include the use of specific kinds of products, for example, file compression utilities such as winzip and animation software such as Flash. When a user sends the connected other a file compressed using winzip as an email attachment or makes a Flash animation available on a homepage, the recommenders role in spreading the word about the software is passive.In the initial stages of the lifecycle when a software package is not

widely known and used, early users are generally viewed as being technically advanced.

4. Motivated Evangelism (ME): where recommenders play an active role in influencing connected others and there are significant network externalities accruing to both influencers and recipients.ICQ an instant messaging application and Dialpad an application to place telephone calls over the Internetare instances of motivated evangelism. In these instances, the influencer as well as the recipient need to use the product for either of them to benefit. The structure of benefits motivates early adopters to actively persuade connected others to also try the product so that they can both use the product.

2.2 Proposed System

we propose a mathematical programming model to improve the user navigation on a website while minimizing alterations to its current structure. Results from extensive tests conducted on a publicly available real data set indicate that our model not only significantly improves the user navigation with very few changes, but also can be effectively solved. In addition, we define two evaluation metrics and use them to assess the performance of the improved website using the real data set. Evaluation results confirm that the user navigation on the improved structure is indeed greatly enhanced. More interestingly, we find that heavily disoriented users are more likely to benefit from the improved structure than the less disoriented users.

The basic assumption is that when users see their social contacts performing an action they may decide to perform the action themselves. In truth, when users perform an action, they may have any one of a number of reasons for doing so: they may have heard of it outside of the online social network and may have decided it is worthwhile; the action may be very popular (e.g., buying an iPhone 4S may be such an action) or they may be genuinely influenced by seeing their social contacts perform that action.Generally, the focuss on the important problem of distinguishing real social influence from homophily and other external factors. Homophily is a term coined by sociologists in the 1950s to explain the tendency of individuals to associate and bond with similar others. This is usually expressed by the famous adage birds of a feather flock together. Homophily assumes selection, i.e., the fact that it is the similarity between users to breed connections.

2.3 Feasibility Study

It is the high level capsule version of the entire requirement analysis process. The objective of feasibility study is to determine whether the proposed system can be developed with available resources.Feasibility study aim to uncover the strengths and weaknesses of the existing business or proposed venture opportunities and threats as presented by the environment. The results of feasibility study are used to make a decision whether to proceed with the project or whether to table it. There are various types of feasibilities available that depend on different factors like technical, operational. A feasibility study could be used to test a proposal for new system.

2.3.1 Economical Feasibility

This includes an evaluation of all incremental costs and benefits expected if proposed system is implemented.Costs-benefit analysis which is to be done during economical feasibility delineates costs for project development and weighs them against system benefits. The system adds information of colleges and companies for which colleges and companies pays as it provides their information as well as company jobs. So developing this system is economically feasible.

2.3.2 Operational Feasibility

Operational feasibility determines whether the proposed system satisfied the user objectives and can be fitted in to current system operation. The system The Prepaid Rationing System can be justified as operationally feasible based on the following Operational feasibility determines if the proposed system satisfied the user objectives and can be fitted in to current system operation. The system Prepaid Rationing System can be justified as operationally feasible based on the following:

1. The methods of processing and presentation are completely acceptable by the users because they meet all their requirements.

2. The users have been involved during the preparation of requirement analysis and design process.

3. The system will certainly satisfy the user objectives and it will also enhance their capability.

4. The system will certainly satisfy the user objectives and it will also enhance their effectively.

5. The system will certainly satisfy the user objectives with the help of agile development.

2.3.3 Technical Feasibility

It is concerned with hardware and software feasibility. In this study, one has to test whether the proposed system can be developed using existing technology or not. As per client requirements the system to be developed should have speed response because of fast exchange of information, reliability, security, scalability, integration and availability. To meet these requirements.We as a developer found Jsp specifications as a right choice because of its features platform independence and reusability [7] .

2.4 Project Scheduling

This section specifies the project scheduling of the project. Software project scheduling is an activity that distributes estimated efforts across the planed project duration by allocating the effort to specific software engineering task. In this phase we are identifying all major software engineering activity and the product function to which they are applying. As we have selected the linear sequential model for developing our project we divide the work according to the phases of this model. As we are four partners working on this project and having nine months we have scheduling the project as shown in figure. If the project has been developed according to the schedule, the project schedule define the task and milestones that must be tracked and controlled as the project proceed. A gantt chart helps in scheduling the activities of project, but it does not helping identifying them during the scheduling activity and also during implementation of the project, new activity may be identified that were not envisioned during the initial planning. The leader must then go back and revise the breakdown structure and the schedule to deal with these activities.

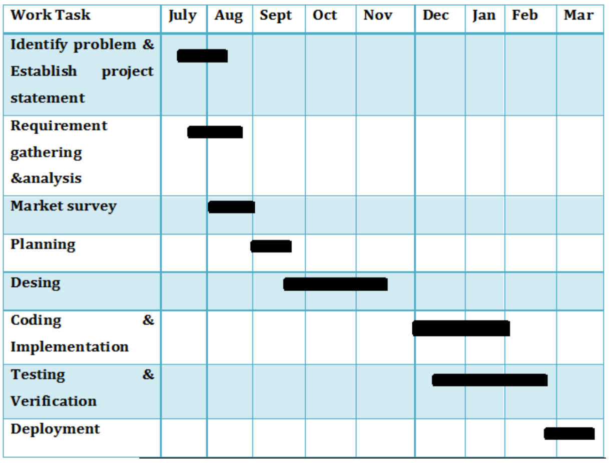


Figure 2.1: Timeline Chart.

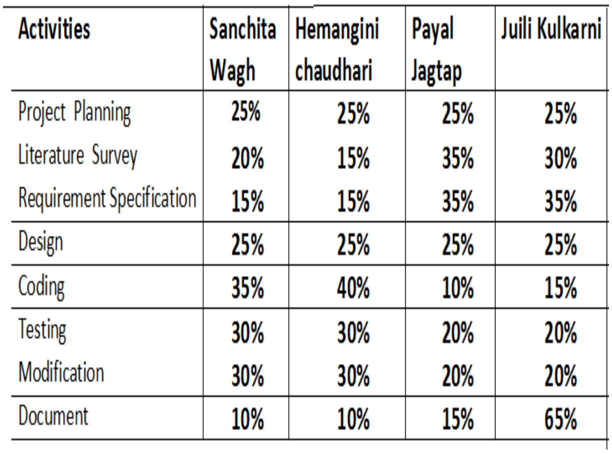
2.5 Effort Allocation

A recommended distribution of effort across the denition and development phases is often referred to as the 40-20-40 rule. Forty percent of all effort is allocated to front-end analysis and design.A similar percentage is applied to back-end testing.Developer can correctly infer that coding (20 percent of effort) is de-emphasized. This effort distribution should be used as a guide line only.The characteristics of each project must dictate the distribution of effort. Work expended on project planning rarely accounts for more than 23 percent of efforts,unless the plane commits an organization to large expenditures with high risk.Requirment analysis may comprise 10-25 percent of project effort.

Effort expended on analysis or prototyping should increase indirect proportion with project size and complexity. A range of 20 to 25 percent of eorts is normally applied to software design.Time expended for design review and subsequence iteration must also be consider. Because of the effort applied to software design, code should follow with relatively little difficulty. A range of 15-20 percent of overall effort can be achieved. Testing and subsequent debugging can account for 30-40percent of software development effort. Team size 4 percent project work should be done in given time. Two working days/weeks and two to four hours/day. Four people have contributed for these work as a team player. Contribution of each member to the project is given in table.

2.6 Summary

This chapter describes the overall phases of project development and also covers the analysis of the project such as it's Feasibility study, Literature Survey and also an overview of proposed system. The next chapter describes the Algorithms implemented in this project.



Activities	Sanchita Wagh	Hemangini chaudhari	Payal Jagtap	Juili Kulkarni
Project Planning	25%	25%	25%	25%
Literature Survey	20%	15%	35%	30%
Requirement Specification	15%	15%	35%	35%
Design	25%	25%	25%	25%
Coding	35%	40%	10%	15%
Testing	30%	30%	20%	20%
Modification	30%	30%	20%	20%
Document	10%	10%	15%	65%

Figure 2.2: Effort Allocation Table.

Chapter 3

Algorithms

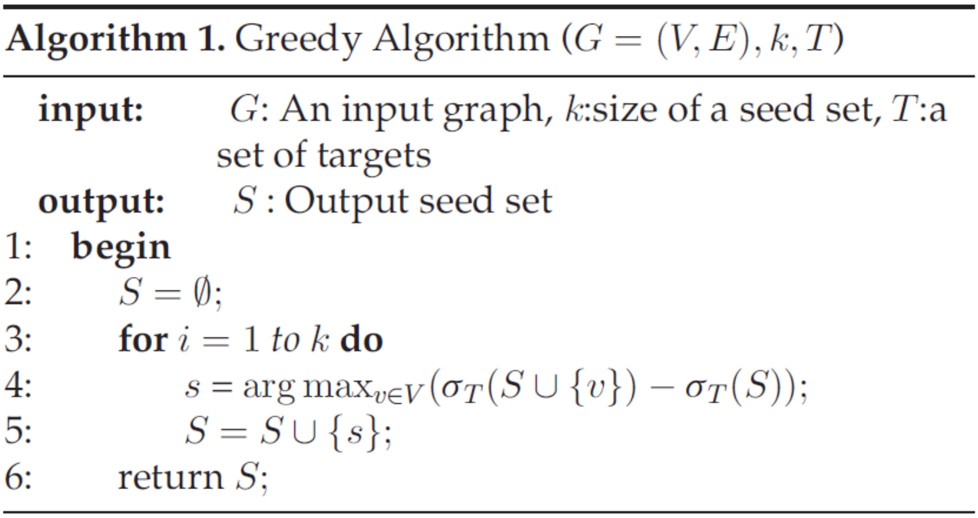
This chapter is elaborated in following sections: Section 3.1 describes the Algorithms of Analyzer that is greedy algorithm and influence maximization algorithm. Proposed System Algorithm is discussed in Section 3.2 that is Artificial Neural Networks C4.5 Algorithm. Section 3.3 describes the backpropagation Algorithm used for verification in system study. Finally, section 3.4 contains Summary.

3.1 Analyzer

Existing System also known as Analyzer model which is a social network is represented as a directed graph G = (V,E) where V is a set of nodes that represents users and E is a set of directed edges that represent relationships between users.For every edge (u,v) belongs to E, (u,v) has a weight, denoted as p(u,v), that is the probability that u influences v directly.

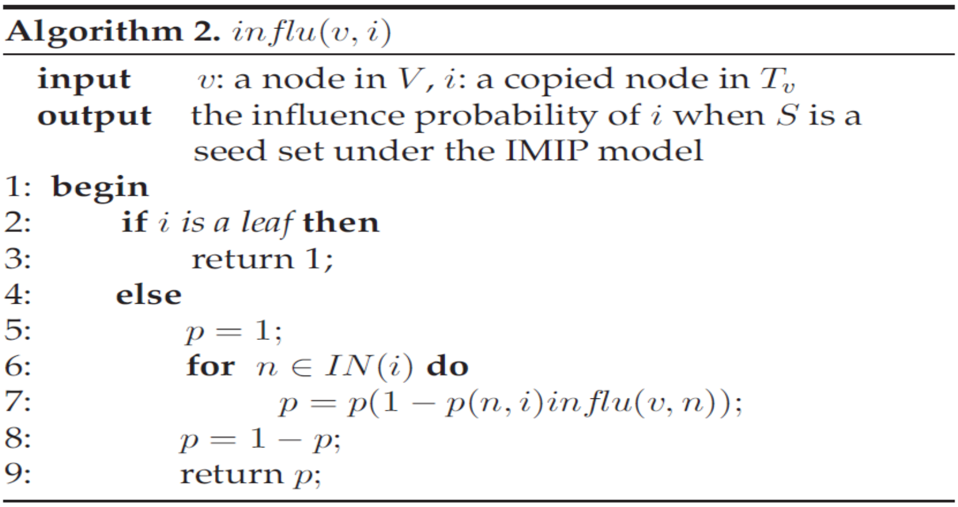
Given a certain item to be promoted, three kinds of users can exist in target aware viral marketing. First, there are target users who have an interest in the item. Second, there are non-target users who can be influenced for the item to introduce it to their friends. Finally, there are non-target users who are immune to being influenced for the item, because they do not want to introduce it to their friends.

It is easy to see that the IC model can handle the first case and the second case. However, the IC model cannot handle the third case, because it does not distinguish such immune nodes from the others. Nevertheless, we can easily modify the IC model to support the third case by adding one condition to it. The modified IC model says that user u has onetime chance to influence an uninfluenced neighbor v, which is not immune, at time t , when u is influenced at time t. Fortunately, this modification only marginally affects the proposed method, since immune nodes can be handled like seed nodes except that they do not influence other nodes and are not counted for influence spread. This is because seed nodes cannot be either influenced by another node like immune nodes. Thus, for simplicity, we stick to the original IC model to explain the proposed method.Thus, the greedy algorithm to find



```
Algorithm 1. Greedy Algorithm ($(G = (V, E), k, T)$)
input:   $G$: An input graph, $k$:size of a seed set, $T$:a
         set of targets
output:  $S$ : Output seed set
begin
    $S = \emptyset$;
    for $i = 1$ to $k$ do
        $s = \arg\max_{v \in V}(\sigma_T(S \cup \{v\}) - \sigma_T(S))$;
        $S = S \cup \{s\}$;
    return $S$;
```

Figure 3.1: Greedy Algorithm.



```
Algorithm 2. $influ(v, i)$
input    $v$: a node in $V$, $i$: a copied node in $T_v$
output   the influence probability of $i$ when $S$ is a
         seed set under the IMIP model
begin
    if $i$ is a leaf then
        return 1;
    else
        $p = 1$;
        for $n \in IN(i)$ do
            $p = p(1 - p(n, i)influ(v, n))$;
        $p = 1 - p$;
        return $p$;
```

Figure 3.2: Influ(v,i).

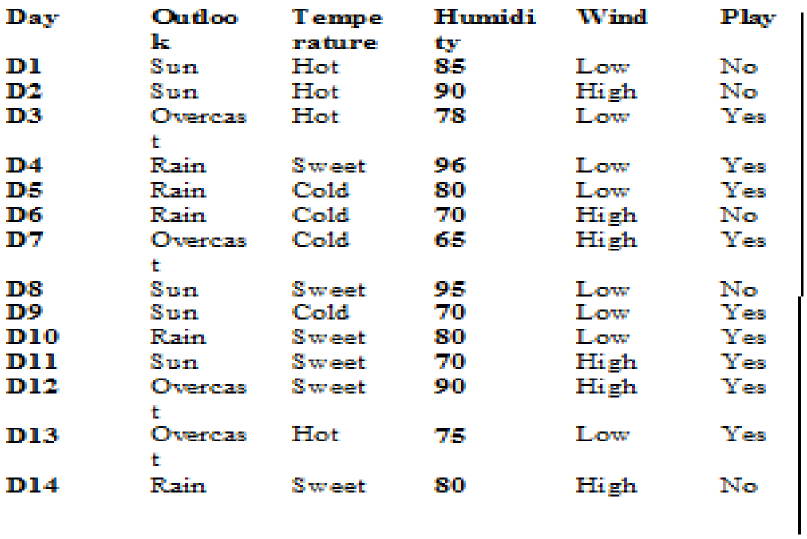
out the k-seeds is shown in Figure 3.1 For processing an IMAX query efficiently with high accuracy, we need a novel preprocessed structure requiring a reasonable space based on a concrete and effective expectation model for influence spread.

Since calculating the influence spread of a seed set is P-hard,existing studies usually use the Monte-Carlo simulations to approximate the influence spread. However, the simulations are still very expensive, so we need a new expectation model to approximate the influence spread. The hardness of calculating the influence spread lies in that a node can influence another node through various paths and the paths are complicatedly entangled. Thus, the new expectation model starts from simplifying the paths with an important property, called the independence between paths. For every two paths that share the destination and may share the source, if they do not share any node except the destination and the source, the two paths are defined to be independent. There is an interesting observation in the independence between paths. Suppose two paths P, Q are independent and they do not share the source. If the source of P is a seed, the other nodes in P can be influenced by the seed but nodes in Q cannot be influenced by the seed. This observation leads to the second Algorithm. Let the influence probability of node v belongs to V be the probability that v is influenced by a node of a given seed set S and is denoted as p(S,v).For every node v belongs to V, if all paths which start from seed and have v as the destination are independent of each other, by the IC model.

The intuition of the IMIP model is as follows. Consider the situation that a node becomes a new seed in the greedy algorithm. As we mentioned, when the new seed is on the maximum influence path from node v to another node u, the PMIA heuristics in find the alternative maximum influence path from v to u, since the seed blocks v on the maximum influence path. However, it is quite expensive to compute it in the query processing time. In the IMIP model, even if the new seed is on one of IMIPs from v to u, we can efficiently estimate the influence from v to u using the independence among the other IMIPs. That is the intuition of the IMIP model. Now we can compute the influence probability of node v given seed set S under the IMIP model. As shown in Figure 3.2 the algorithm of computing Influence probability. In lines 2-3 of Algorithm 2, if i is a leaf, then the algorithm returns 1, since a leaf corresponds to seed. Otherwise, in lines 5-9, the algorithm computes the influence probability of i according to the IC model, then returns it. Thus, influ(v,root(v)) returns the influence probability of v.

3.2 Artificial Neural Network- C4.5 Algorithm

Artificial Neural Networks (ANNs) are loosely modeled after the brain. ANNs are composed of units(also called nodes) that are modeled after neurons, with weighted links interconnecting the units together.The main difference between ANNs and other learning mechanisms is that it is composed of these simple units and they work together in a highly parallel manner. Classification is most common method used for finding the mine rule from the large database. Decision tree method generally used for the Classification, because it is the simple hierarchical structure for the user understanding and decision making. The algorithm which we implemented is the C4.5 classification Algorithm, which is a natural extension of ID3 Algorithm. C4.5 algorithm was proposed by Ross Quinlan , to overcome the limitations of ID3 algorithm. One limitation of ID3 is that it is overly sensitive to features with large numbers of values. To overcome this problem, C4.5 uses Information gain. This computation does not, in itself, produce anything new. However, it allows to measure a gain ratio.



Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sun	Hot	85	Low	No
D2	Sun	Hot	90	High	No
D3	Overcast	Hot	78	Low	Yes
D4	Rain	Sweet	96	Low	Yes
D5	Rain	Cold	80	Low	Yes
D6	Rain	Cold	70	High	No
D7	Overcast	Cold	65	High	Yes
D8	Sun	Sweet	95	Low	No
D9	Sun	Cold	70	Low	Yes
D10	Rain	Sweet	80	Low	Yes
D11	Sun	Sweet	70	High	Yes
D12	Overcast	Sweet	90	High	Yes
D13	Overcast	Hot	75	Low	Yes
D14	Rain	Sweet	80	High	No

Figure 3.3: Sample Dataset.

$$GainRatio(p,T) = \frac{Gain(p,T)}{SplitInfo(p,T)}$$

(a) Gain Ratio

$$SplitInfo(p,test) = -\sum_{j=1}^{n} P'\left(\frac{j}{p}\right) \times \log\left(P'\left(\frac{j}{p}\right)\right)$$

(b) Split Info.

Figure 3.4: Gain Ratio and Split-Info.

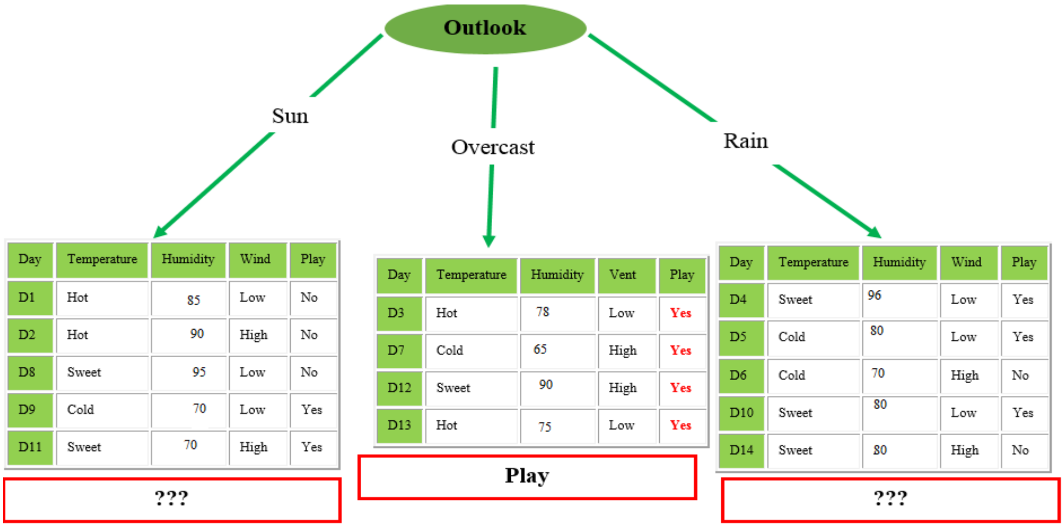
3.3 Backpropagation Algorithm

The backpropagation algorithm is a multi-layer network using a weight adjustment based on the sigmoid function, like the delta rule. The backpropagation method, as well as all the methods previously mentioned are examples of supervised learning, where the target of the function is known. The Figure 3.8 is an example of the backpropagation algorithm working on a small Artificial Neural Network. The Network has a single hidden layer of size two and input and output nodes of size 3. Initialize the weighted links. Typically the weights are initialized to a small random number. Then, for each training example in the testing set is shown in Figures 3.9, Figure 3.10, Figure 3.11 respectively.

Input the training data to the input nodes, then calculate O(k), which is the output of node k. This is done for each node in the hidden layer(s) and output layer.

3.4 Summary

In this Chapter we discussed the various algorithms which we studied and then implemented.The next Chapter describes the various software and hardware system requirement specifications with the functional and non-functional requirements.



Day	Temperature	Humidity	Wind	Play
D1	Hot	85	Low	No
D2	Hot	90	High	No
D8	Sweet	95	Low	No
D9	Cold	70	Low	Yes
D11	Sweet	70	High	Yes

???

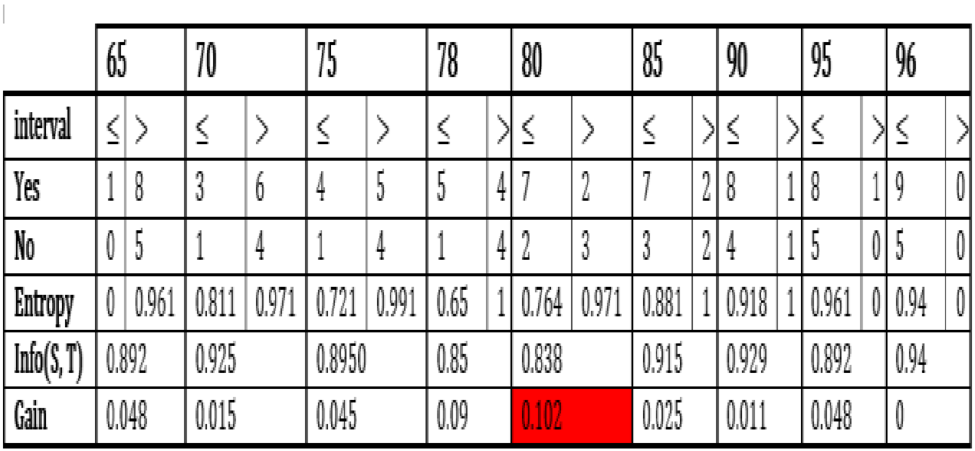
Day	Temperature	Humidity	Vent	Play
D3	Hot	78	Low	Yes
D7	Cold	65	High	Yes
D12	Sweet	90	High	Yes
D13	Hot	75	Low	Yes

Play

Day	Temperature	Humidity	Wind	Play
D4	Sweet	96	Low	Yes
D5	Cold	80	Low	Yes
D6	Cold	70	High	No
D10	Sweet	80	Low	Yes
D14	Sweet	80	High	No

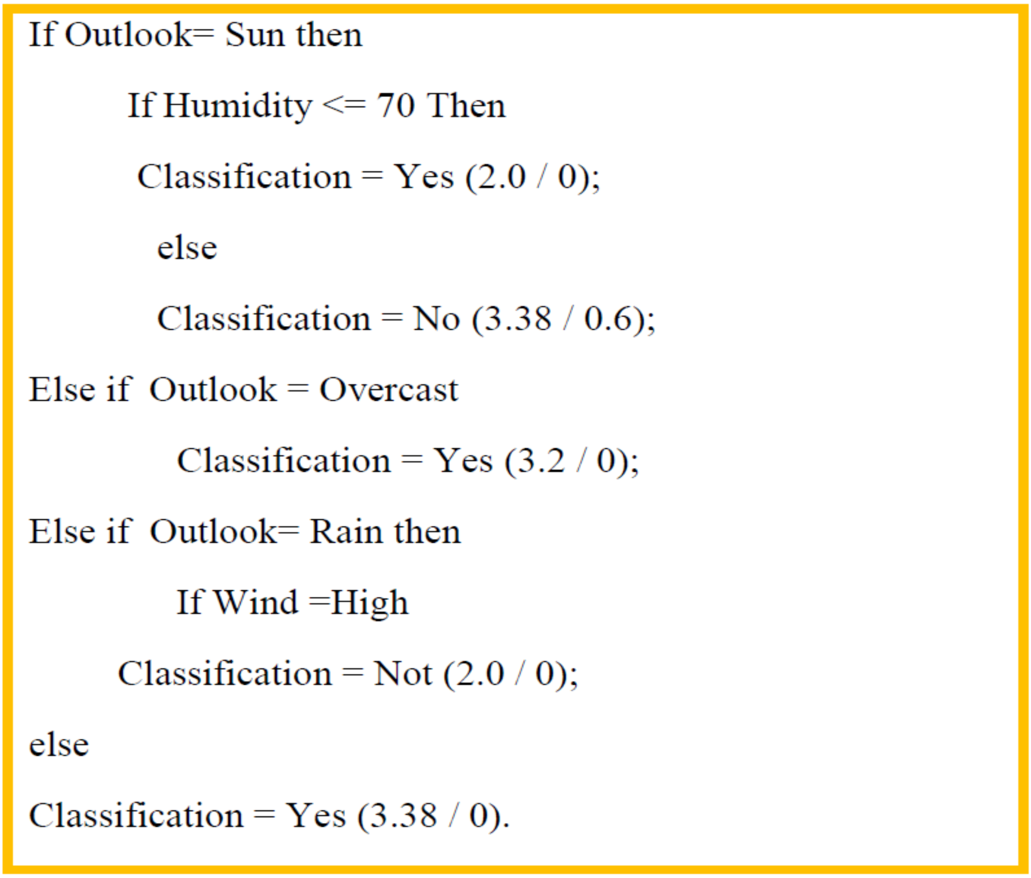
???

Figure 3.5: Root Node of C4.5 Decision Tree.



	65		70		75		78		80		85		90		95		96	
interval	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	1	8	3	6	4	5	5	4	7	2	7	2	8	1	8	1	9	0
No	0	5	1	4	1	4	1	4	2	3	3	2	4	1	5	0	5	0
Entropy	0	0.961	0.811	0.971	0.721	0.991	0.65	1	0.764	0.971	0.881	1	0.918	1	0.961	0	0.94	0
Info(S, T)	0.892		0.925		0.8950		0.85		0.838		0.915		0.929		0.892		0.94	
Gain	0.048		0.015		0.045		0.09		0.102		0.025		0.011		0.048		0	

Figure 3.6: Gain Calculation For The Attribute Continuous Humidity using C4.5 Algorithm.



```
If Outlook= Sun then
          If Humidity <= 70 Then
           Classification = Yes (2.0 / 0);
             else
             Classification = No (3.38 / 0.6);
Else if  Outlook = Overcast
               Classification = Yes (3.2 / 0);
Else if  Outlook= Rain then
               If Wind =High
         Classification = Not (2.0 / 0);
else
Classification = Yes (3.38 / 0).
```

Figure 3.7: Decision Rules.

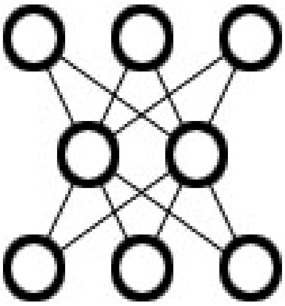
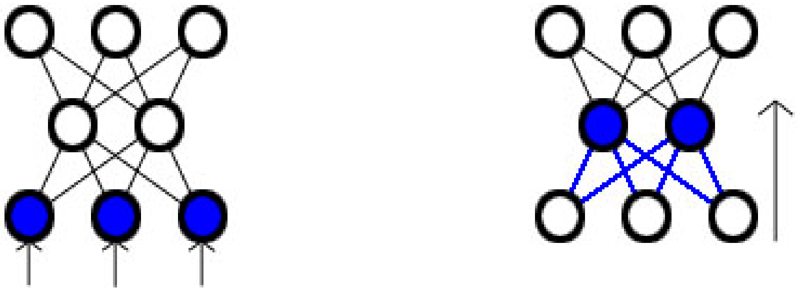
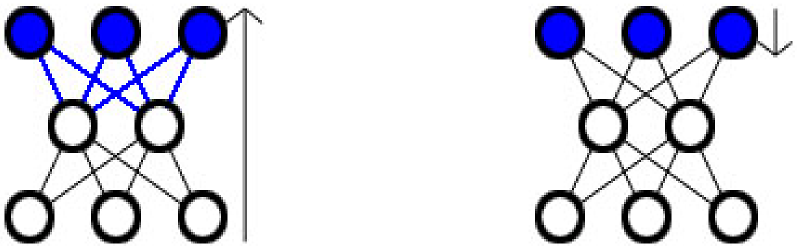


Figure 3.8: Network with Hidden Layer Size of 2.



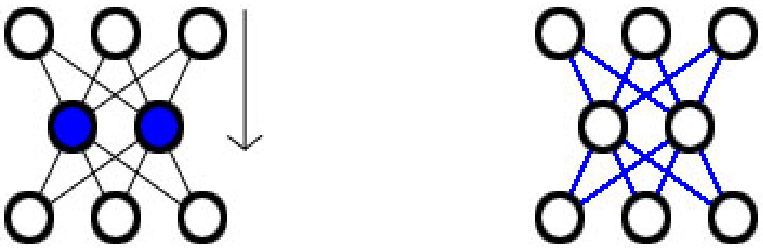
(a) (b)

Figure 3.9: Input the Training Data to Input Nodes.



(a) (b)

Figure 3.10: Calculate the Output of Node K.



(a) (b)

Figure 3.11: Calculated Target Node.

Chapter 4

System Requirement Specification

Software Requirement Specification is the official statement of what is required to the system developers. It should include both user requirements and a detailed specification of the system requirements. Requirement analysis is done in order to understand the problem the software system is to solve. This chapter is elaborated in following sections: Section 4.1 describes the Hardware Requirements. Software Requirements are explained in Section 4.2. Section 4.3 explains the Functional Requirements. Non Functional Requirements are described in Section 4.4. Section. Finally, section 4.5 contains Summary.

4.1 Hardware Requirements

The hardware requirements includes:

1. Processor : Pentium III
2. Speed : 1.1 Ghz
3. RAM : 256 MB(min)
4. Hard disk : 20 GB
5. Floppy drive : 1.44 MB
6. Key board : Standard Windows keyboard.
7. Mouse : Two or Three button mouse.
8. Monitor : SVGA.

4.2 Software Requirements

The software requirements includes:

1. Operating System : Windows95/98/2000/XP
2. Application Server : IIS Server.
3. Database : MySql Server 2005.
4. Front End : Java,HTML,Jsp.
5. Scripts : JavaScript.
6. Server Side Script : ASP.net.

4.3 Functional Requirements

The System does the following functional task:

1. Allows only the authorized user to interact with the website
2. Information of developer and instruction for accessing website
3. Easy enhancement

4.4 Non-Functional Requirements

Non functional requirement include various requirement but the most prominent ones. Effective software project estimation is one of the challenging and important activities in software development. Proper project planning and control is not possible without sound and reliable estimate.

4.5 Summary

In this chapter, Hardware Requirements, Software Requirements, Functional and Non-Functional Requirements are explained. In next chapter, the System Design is described through various UML diagram.

Chapter 5

System Design

System design provides the understanding and procedural details necessary for implementing the system.System design is a meaningful engineering representation of something that is to be built. It can be traced to a customers requirements and at the same time assessed for quality against a set of pretend criteria for good design. In the software engineering context, design focuses on three major areas for concern: architecture, interfaces and components. This chapter is elaborated in following sections: Section 5.1 describes the Module to module interaction diagram. UML Diagrams are described in Section 5.2. Finally, section 5.3 contains Summary.

5.1 System Architecture

The system architecture provide details of how the components or modules are integrated and is described with the help of Unified Modelling Diagram.

5.1.1 Data Flow Diagram

The Data Flow Diagram is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system. The Figure 4.1 shows the Data flow Diagram for Admin.It shows that how modules are connected to each other during Admin workflow, i.e how many activities are performed by it.

The Figure 4.2 shows the Data flow diagram for User.It shows that how modules are connected to each other during User workflow, i.e how many activities are performed by it.

5.2 UML Diagrams

UML is a method for describing the system architecture in detail using the blueprint. UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. UML is a very important part of developing objects oriented software and the software development process. UML uses mostly graphical notations to express the design of software projects. Using the UML helps project teams communicate, explore potential designs, and validate the architectural design of the software.

5.2.1 Use Case Diagram

A use case is a set of scenarios that describing an interaction between a user and a system. A use case diagram displays the relationship among actors and use cases. The two main components of a use case diagram are use cases and actors.An actor is represents a user or another system that will interact with the system you are modeling. A use case is an external view of the system that represents some action the user might perform in order to complete a task.It consists of actors like User and Admin.

The use cases in Figure are Add Webpage, Add Web content, Find Mini sessions .Actor admin are connected to use cases according to their role in system. The Figure 4.3 shows the Use case Diagram for Admin. The use cases in Figure are Search Webpage, Select navigation, view MP Model. Actor User are connected to use cases according to their role in system. The Figure 4.4 shows the Use case Diagram for User.

5.2.2 Component Diagram

A component diagram shows the organization and dependencies among set of components. These diagrams are used to model static view of the system. The Figure 4.5, shows the component diagram for the proposed system for Admin. The components used in Figure 4.6 are Login,Student and admin. The task specification of each of the component is represented in the Figure 4.6. The task specifications of User is shown with various components connected to it.Search page, Web Personalization, Web transformation, Mp model are its various components.

5.2.3 Sequence Diagram

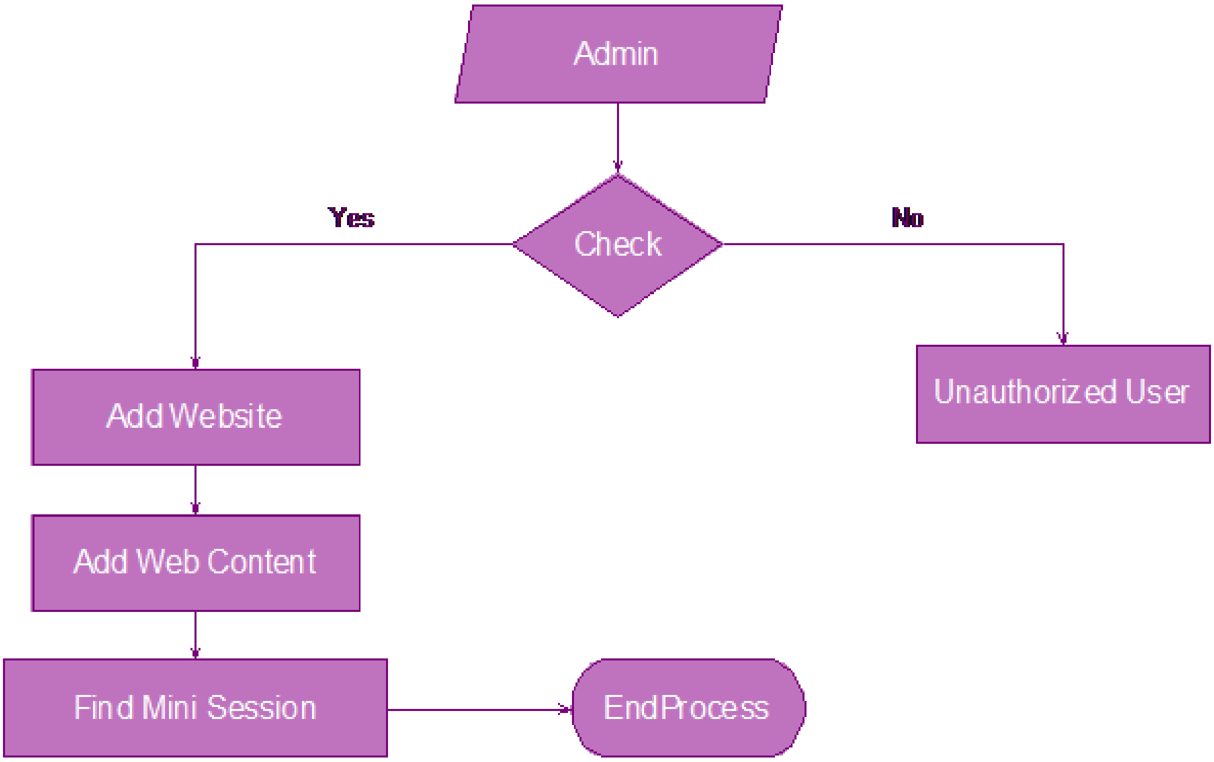
The Figure 4.7 shows sequence diagram of Admin workflow system. The Figure 4.8 shows the sequence diagram of User workflow system.

5.2.4 Activity Diagram

The Figure 4.8 shows the Activity diagram for the proposed system.

5.3 Summary

This chapter includes various UML diagrams which shows the overall design of project. The next chapter describes the Implementation of the system.



(a)

Figure 5.1: DFD for Admin.

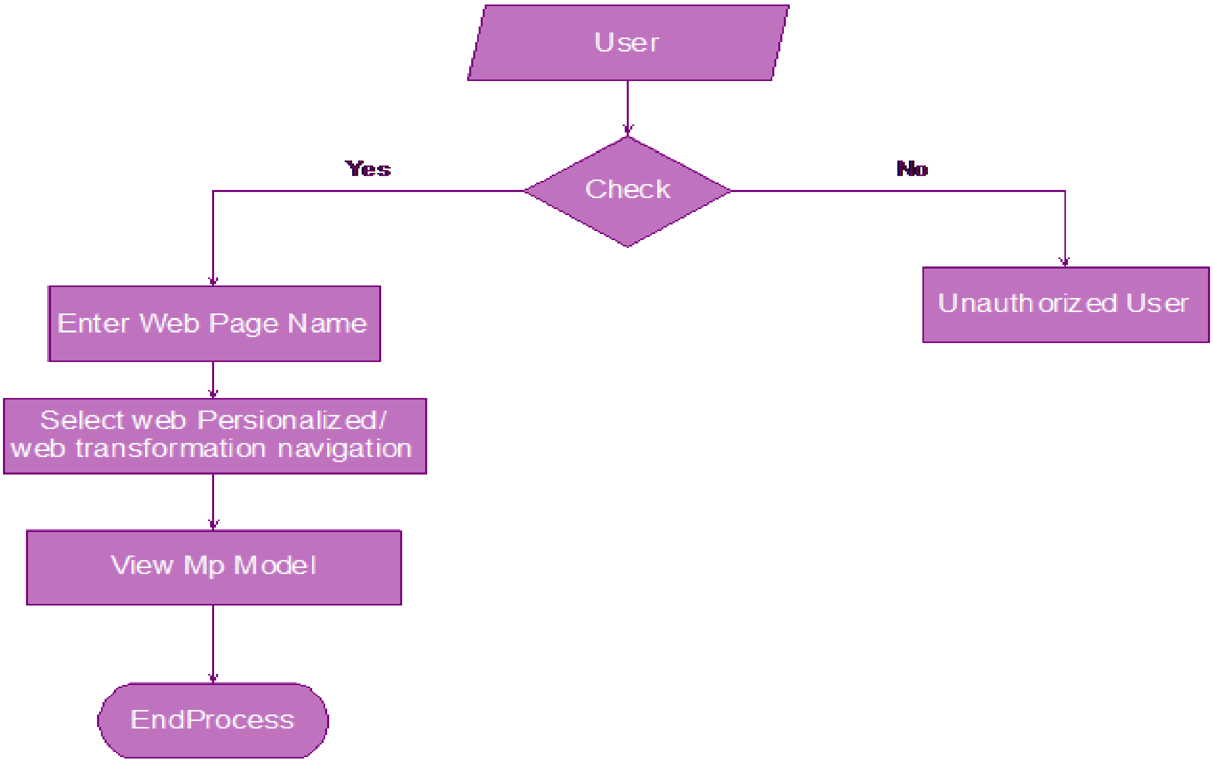


Figure 5.2: DFD for User.

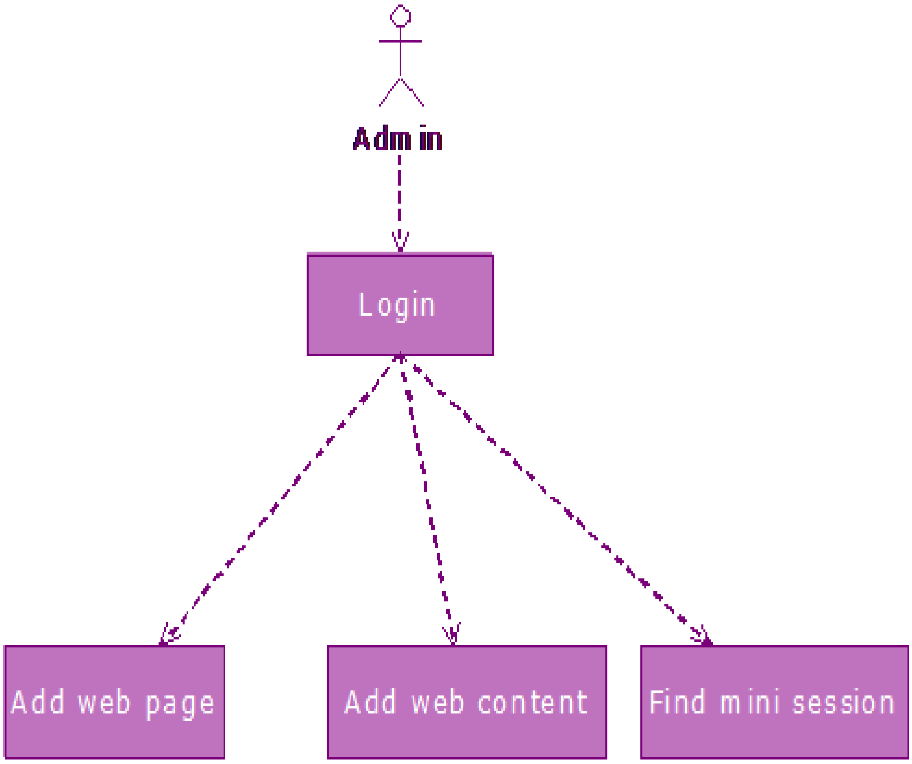


Figure 5.3: Use Case Diagram for Admin.

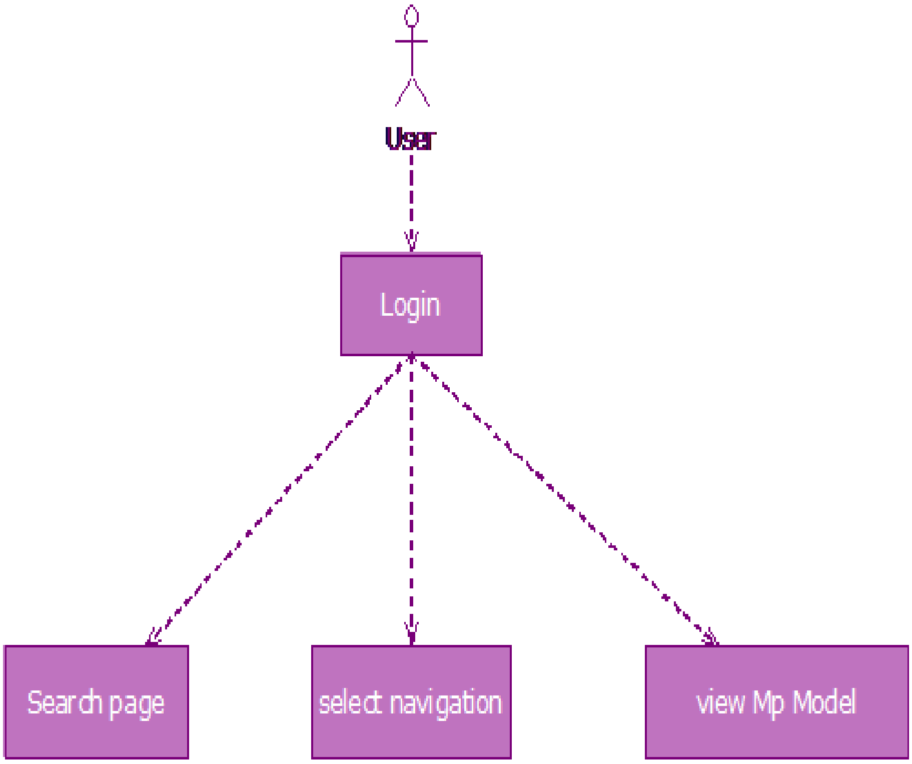


Figure 5.4: Use Case Diagram for User

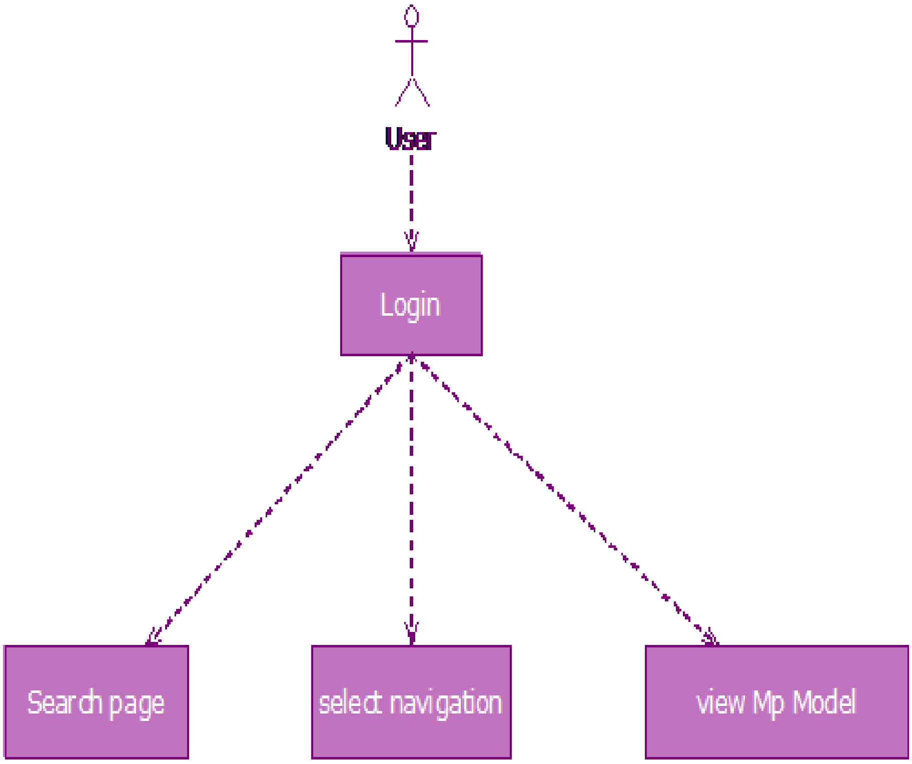


Figure 5.5: Component Diagram for Admin.

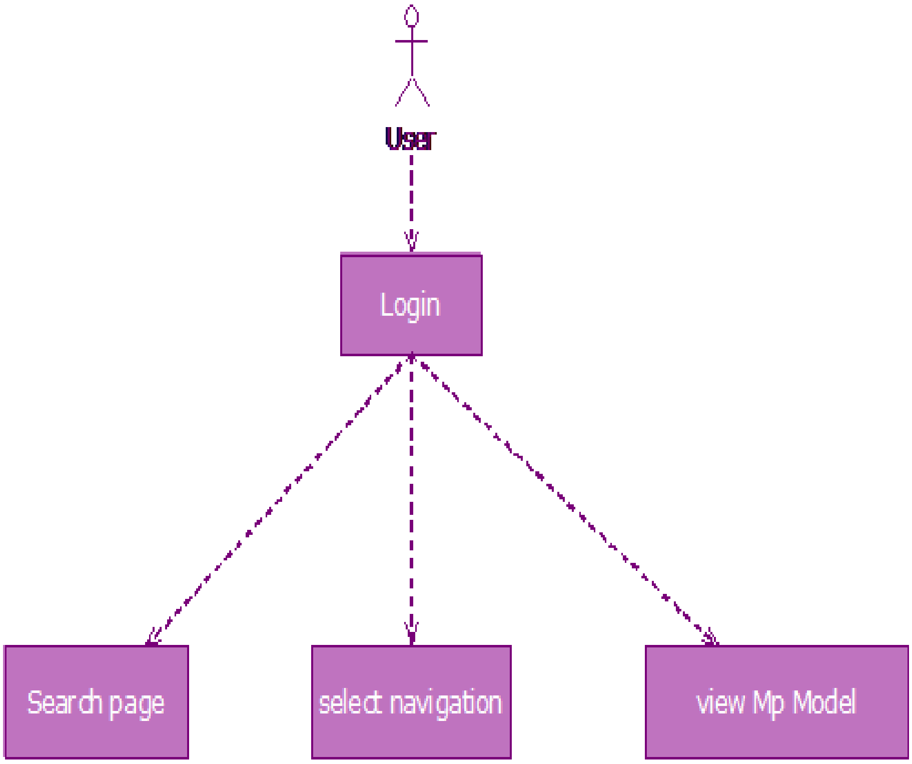
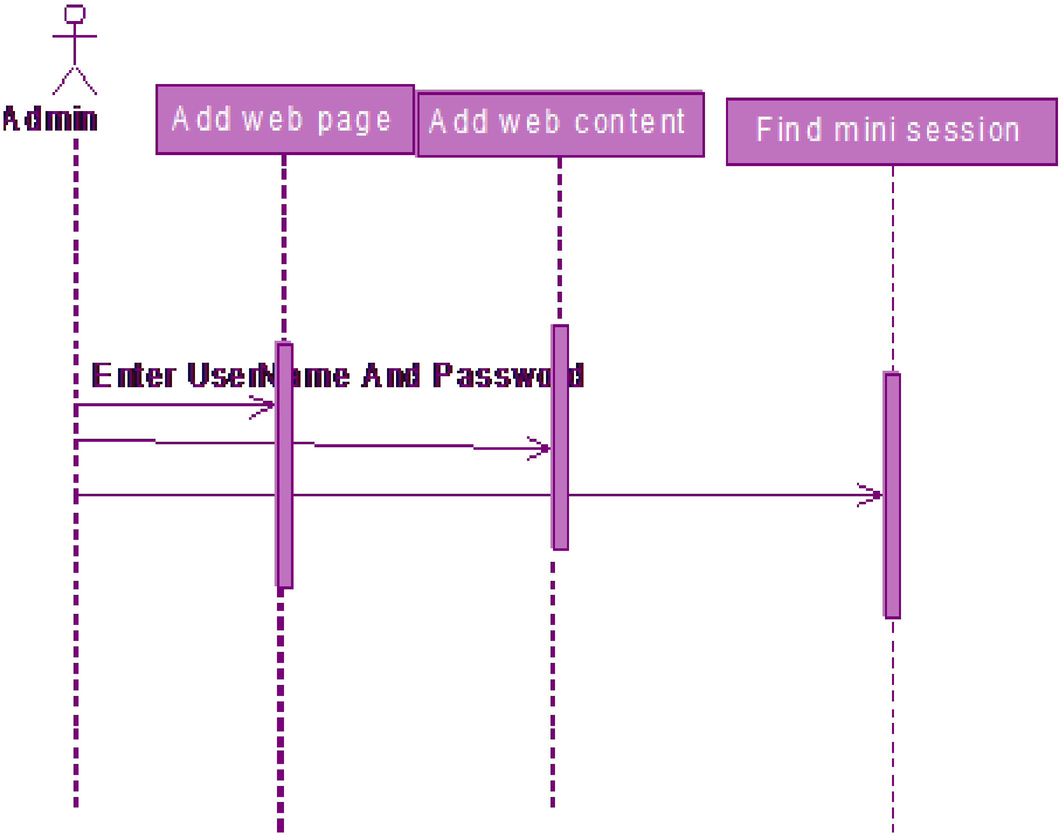
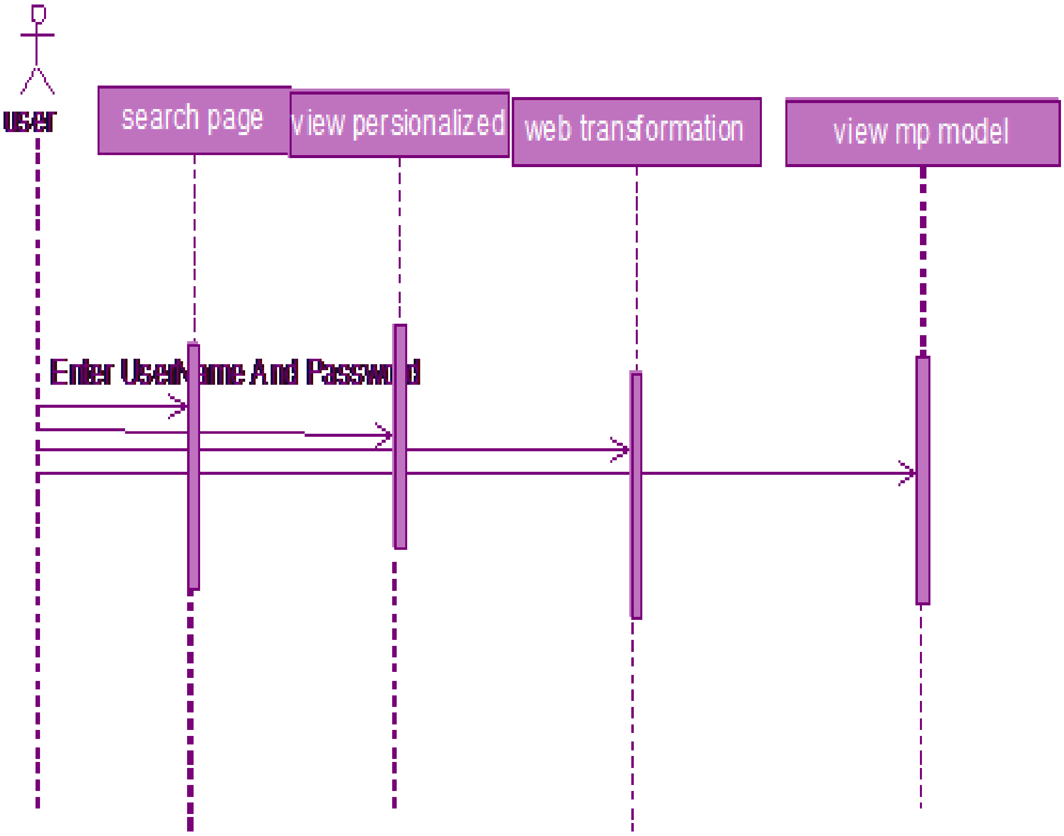


Figure 5.6: Component Diagram for User.



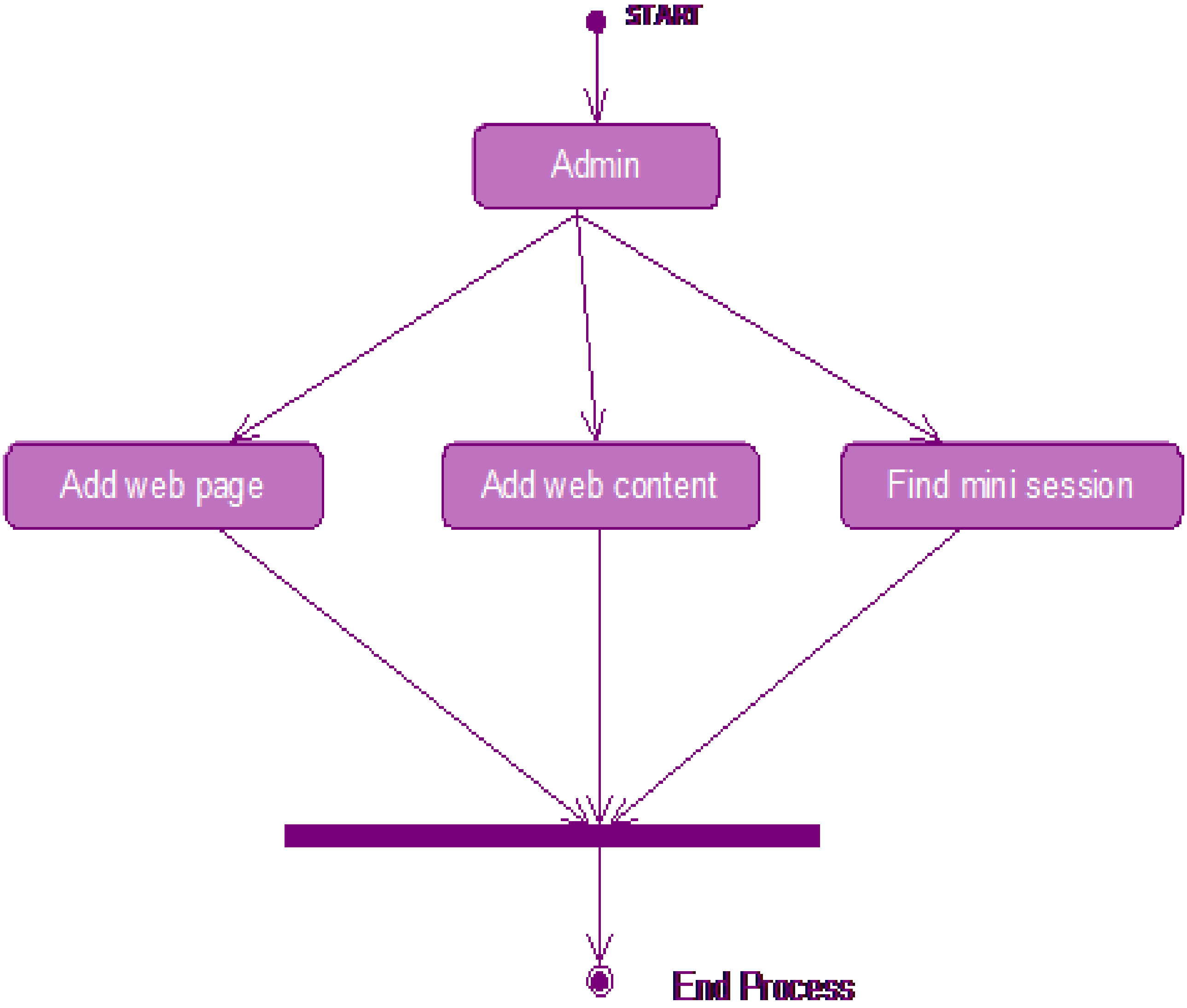
(a)

Figure 5.7: Sequence Diagram for Admin.



(a)

Figure 5.8: Sequence Diagram for User.



(a)

Figure 5.9: Activity Diagram.

Chapter 6

Implementation

Important phase in system development is the successful implementation of the new system design. Implementation includes all those activities that take place to capture information then report it i.e store it and then show it as per the requirement of the user. This chapter is elaborated in following sections: Section 6.1 describes the Web Personalization. Section 6.2 describes Web Transformation. Section 6.3 tells about Maximal Forward Reference. Section 6.4 describes Mini Sessions. Section 6.5 tells Out-Degree Threshold. Section 6.6 explains about the Flow of System in Development. Finally, section 6.7 contains Summary.

6.1 Web Personalization

Web personalization is the process of tailoring webpages to the needs of specific users using the information of the users navigational behavior and profile data. Perkowitz and Etzioni describe an approach that automatically synthesizes index pages which contain links to pages pertaining to particular topics based on the co-occurrence frequency of pages in user traversals, to facilitate user navigation. The methods proposed by Mobasher et al. and Yan et al. create clusters of users profiles from weblogs and then dynamically generate links for users who are classified into different categories based on their access patterns.

6.2 Web Transformation

Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large set of users instead of personalizing pages for individual users. Fu et al. describe an approach to reorganize web pages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al. propose a heuristic method based on simulated annealing to relink web pages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices.

6.3 Maximal Forward Reference

We use backtracks to identify the paths that a user has traversed, where a backtrack is defined as a users revisit to a previously browsed page. The intuition is that users will backtrack if they do not find the page where they expect it . Thus, a path is defined as a sequence of pages visited by a user without backtracking, a concept that is similar to the maximal forward reference defined in Chen et al. Essentially, each backtracking point is the end of a path. Hence, the more paths a user has traversed to reach the target, the more discrepant the site structure is from the users expectation.

6.4 Mini Sessions

Recall that a mini session is relevant only if its length is larger than the corresponding path threshold. Consequently, only relevant mini sessions need to be considered for improvement and this leads to a large number of irrelevant mini sessions (denoted as TI) being eliminated from consideration in our MP model.

6.5 Out-Degree Threshold

Web pages can be generally classified into two categories : index pages and content pages. An index page is designed to help users better navigate and could include many links, while a content page contains information users are interested in and should not have many links. Thus, the out-degree threshold for a page is highly dependent on the purpose of the page and the website. Typically, the out degree threshold for index pages should be larger than that for content pages.

6.6 Flow of System in Development

There are times when the requirements for a problem are well understoodwhen work flows from communication through deployment in a reasonably linear fashion. This situation is sometimes encountered when well-defined adaptations or enhancements to an existing system must be made (e.g., an adaptation to accounting software that has been mandated because of changes to government regulations). It may also occur in a limited number of new development efforts, but only when requirements are well defined and reasonably stable.

The waterfall model, sometimes called the classic life cycle, suggests a systematic, sequential approach to software development that begins with customer specification of requirements and progresses through planning, modeling, construction, and deployment, culminating in ongoing support of the completed software [4] .

In an interesting analysis of actual projects, Bradac found that the linear nature of the classic life cycle leads to blocking states in which some project team members must wait for other members of the team to complete dependent tasks. In fact, the time spent waiting can exceed the time spent on productive work! The blocking states tend to be more prevalent at the beginning and end of a linear sequential process. Today, software work is fast-paced and subject to a never-ending stream of changes (to features, functions, and information content). The waterfall model is often inappropriate for such work. However, it can serve as a useful process model in situations where requirements are fixed and work is to proceed to completion in a linear manner.

6.7 Summary

This chapter explains Implementation details of the project like Web personalization, Web Transformation, Maximal forward Reference, Mini Sessions, Out-Degree Threshold and Flow of System. The next chapter describes the System Testing.

Chapter 7

System Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

All the testing of our project is done by tester is explained in this chapter. Section 7.1 describes How to implement unit testing. Integration testing is Described in Section 7.2. Section 7.3 describes Acceptance testing. Section 7.4 describes Functional Test. Section 7.5 System test. Section 7.6 explains Black Box Testing. Finally, Section 7.7 tells the summary of the Chapter. Tests are the individual tests specified in a test plan document. Each test is typically described by 1.An initial system state,2.A set of actions to be performed.,3.The expected results of the test. Testing As a Continuous Process: All testing follows a pre-planned process, which isagreed to. All tests consider not only a nominal system condition but also address anomalous and recovery aspects of the system. The system is tested in a stressed environment, nominally in excess of 150 percent of its rated capacities. All test products (test cases, data, tools, configuration, and criteria) are documented in a software description document. Every test shall be described in traceable procedures and have pass-fail criteria included.

7.1 Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

1. Test strategy and approach
   (a) Field testing will be performed manually and functional tests will be written in detail.
2. Test objectives
   (a) All field entries must work properly.
   (b) Pages must be activated from the identified link.
   (c) The entry screen, messages and responses must not be delayed.
3. Features to be tested
   (a) Verify that the entries are of the correct format
   (b) No duplicate entries should be allowed
   (c) All links should take the user to the correct page.

7.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or one step up software applications at the company level interact without error.

1. Test Results

    (a) All the test cases mentioned above passed successfully. No defects encountered.

7.3 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

1. Test Results

    (a) All the test cases mentioned above passed successfully. No defects encountered.

7.4 Functional Test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

1. Valid Input : identified classes of valid input must be accepted.
2. Invalid Input : identified classes of invalid input must be rejected
3. Functions : identified functions must be exercised.
4. Output : identified classes of application outputs must be exercised.
5. Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.5 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.6 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot see into it. The test provides inputs and responds to outputs without considering how the software works [6] ..

7.7 Summary

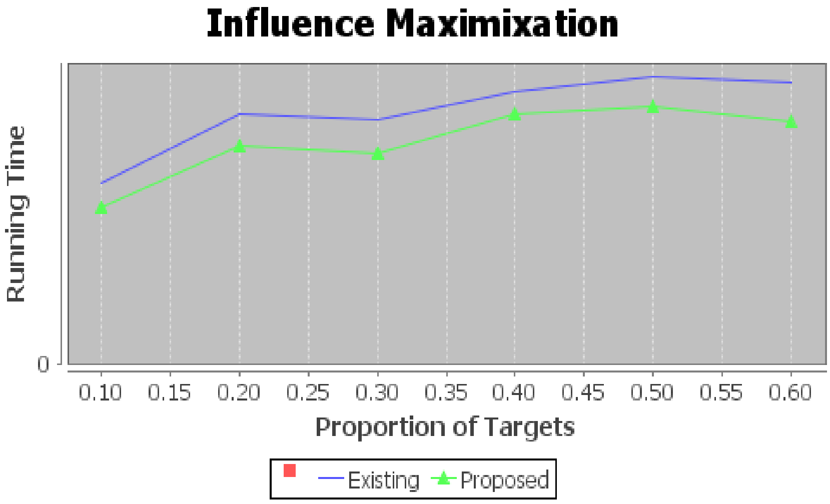
This chapter describes How to implement various types of testing and test case and test results of the project. Next chapter describes the Results of Influence Maximization in Viral Marketing of Social Networks.

Chapter 8

Result and Discussion

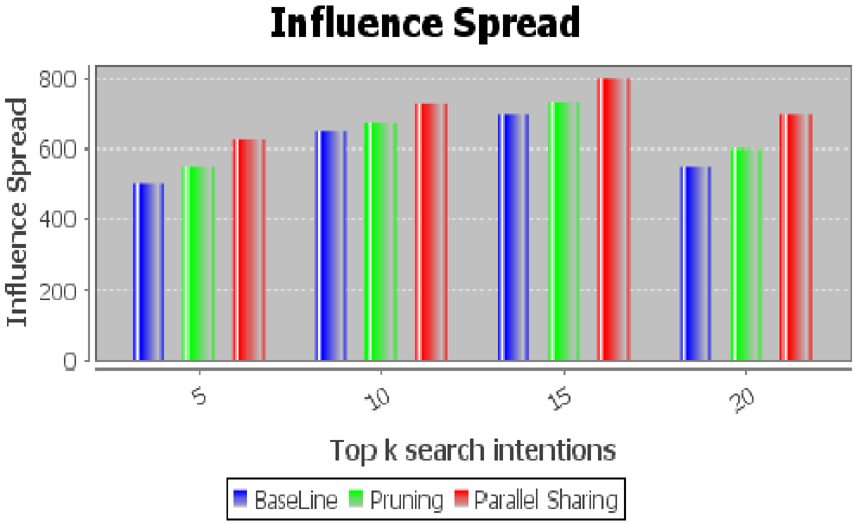
We have proposed a mathematical programming model to improve the navigation effectiveness of a website while minimizing changes to its current structure, a critical issue that has not been examined in the literature. Our model is particularly appropriate for informational websites whose contents are relatively stable over time. It improves a website rather than reorganizes it and hence is suitable for website maintenance on a progressive basis. The tests on a real website showed that our model could provide significant improvements to user navigation by adding only few new links. Optimal solutions were quickly obtained, suggesting that the model is very effective to real world websites. In addition, we have tested the MP model with a number of synthetic data sets that are much larger than the largest data set considered in related studies as well as the real data set. The MP model was observed to scale up very well, optimally solving large-sized problems in a few seconds in most cases on a desktop PC.

To validate the performance of our model, we have defined two metrics and used them to evaluate the improved website using simulations. Our results confirmed that the improved structures indeed greatly facilitated user navigation. In addition, we found an appealing result that heavily disoriented users, i.e., those with a higher probability to abandon the website, are more likely to benefit from the improved structure than the less disoriented users. Experiment results also revealed that while using small path thresholds could result in better outcomes, it would also add significantly more new links. Thus, Webmasters need to carefully balance the tradeoff between desired improvements to the user navigation and the number of new links needed to accomplish the task when selecting appropriate path thresholds. Since no prior study has examined the same objective as ours, we compared our model with a heuristic instead. The comparison showed that our model could achieve comparable or better improvements than the heuristic with considerably fewer new links.



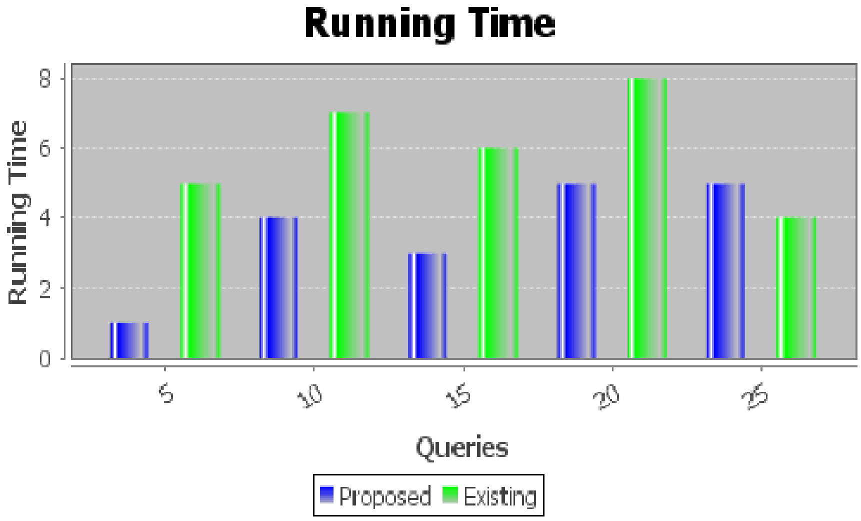
(a)

Figure 8.1: Execution Time Graph.



(a)

Figure 8.2: Influence Generation Graph.



(a)

Figure 8.3: Runtime Query Graph.

Chapter 9

Conclusion and Future Scope

To expand the impact on particular clients in informal organizations we preparing detail IMAX question.Since ascertaining its target capacity is P-hard and IMAX inquiry handling is NP-hard, we concentrate on the most proficient method to inexact ideal seeds effectively. To estimate the target's estimation capacity, the IMIP model is proposed in view of freedom between ways. To prepare an IMAX question proficiently,the quick ravenous based estimate utilizing the IMIP model and removing possibility for ideal seeds is proposed.

Bibliography

[1] Pingdom, Internet 2009 in Numbers, http://royal.pingdom. com/2010/01/22/internet-2009-in-numbers/, 2010.

[2] Tealeaf, The Two Waves of Online Abandonment: The 2007 Harris Interactive Survey of Online Customer Behavior, http://www.tealeaf.com/downloads/tealeaf-executivebriefHarris2007. pdf, 2007.

[3] D. Dhyani, W.K. Ng, and S.S. Bhowmick, A Survey of Web Metrics, ACM Computing Surveys, vol. 34, no. 4, pp. 469-503, 2002.

[4] H. Kao, J. Ho, and M. Chen, WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model, IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614-627, May 2005.

[5] H. Kao, S. Lin, J. Ho, and M. Chen, Mining Web Informative Structures and Contents Based on Entropy Analysis, IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1, pp. 41-55, Jan. 2004.

[6] C. Kim and K. Shim, TEXT: Automatic Template Extraction from Heterogeneous Web Pages, IEEE Trans. Knowledge and Data Eng., vol. 23, no. 4, pp. 612-626, Apr. 2011.

[7] http://www.sourcefordgde.com