

HYBRID E-MARKETING USING WEBPAGE MINING FOR WEBSITE MONETIZATION

A PROJECT REPORT

Submitted by

S. DINESH KUMAR 1305070

S. KEERTHIVASAN 1305083

A. MADHAN KUMAR 1305086

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



COIMBATORE INSTITUTE OF TECHNOLOGY

(Government Aided Autonomous Institution Affiliated to Anna University)

COIMBATORE-641 014

ANNA UNIVERSITY - CHENNAI 600 025

APRIL 2017

COIMBATORE INSTITUTE OF TECHNOLOGY
(A Govt. Aided Autonomous Institution Affiliated to Anna University)
COIMBATORE – 641 014

BONAFIDE CERTIFICATE

Certified that this project “**HYBRID E-MARKETING USING WEBPAGE MINING FOR WEBSITE MONETIZATION**” is the bonafide work of **S.DINESH KUMAR, S.KEERTHIVASAN** and **A. MADHAN KUMAR** under my supervision during the academic year 2015-2016.

Prof. K.S.Palanisamy, M.E.,
HEAD OF THE DEPARTMENT,
Department of CSE & IT,
Coimbatore Institute of Technology,
Coimbatore – 641 014.

Mrs. P.Kumudha, M.E.,
SUPERVISOR,
Department of CSE & IT,
Coimbatore Institute of Technology,
Coimbatore – 641 014.

Certified that the candidates were examined by us in the project work viva-voce examination held on

Internal Examiner

External Examiner

Place: Coimbatore

Date:

ACKNOWLEDGEMENT

We take this opportunity to express our sincere thanks to our Director, **Dr.S.R.K.Prasad**, for providing us the facilities and support that helped us to complete this project successfully.

We express our sincere thanks to our Secretary **Dr.R.Prabhakar** and our Principal **Dr.V.Selladurai** for providing us a greater opportunity to carry out our work. The following words are rather very meagre to express our gratitude to them. This work is the outcome of their inspiration and product of plethora of their knowledge and rich experience.

We record the deep sense of gratefulness to **Prof.K.S.Palanisamy, M.E.**, Head of the department of Computer Science and Engineering, for his encouragement during this tenure.

We equally tender our sincere thankfulness to **Mrs.P.Kumudha, M.E.**, Assistant Professor and our supervisor, Department of Computer Science and Engineering, for her phenomenal support and sustained guidance which helped in the timely completion of the project.

During the entire period of study, the entire staff members of the Department of Computer Science and Engineering have offered ungrudging help. We present our gratefulness to the Lab Assistants and other non-teaching staff for their timely support and assistance in the laboratory. It is also a great pleasure to acknowledge the unfailing help we have received from our friends.

It is a matter of great pleasure to thank our parents and family members for their constant support and co-operation in the pursuit of this endeavor. We express our sincere thanks to God Almighty for His cherished blessings.

ABSTRACT

In the present business world scenario everything is getting automated, thanks to the technology which blossomed in the last decade. In order for the business to run effectively gaining a competitive advantage over others tactical marketing skills are required. Since the entire human population is revolutionized by Internet it can be made as a medium of marketing. Websites traffic can be converted to revenue through Website Monetization. Google Ads uses this method in the form of cost per click or cost per impression basis. Google AdSense displays ads relevant to the web page content. Google AdChoice displays ads based on the preferences of individual users for online shopping. This system is a hybrid of both AdSense and AdChoice features. Keyword mining is from AdSense whereas Preference and Demography features are from AdChoice and these are used to target customers in both company side and public side. Advertisements from the companies are collected, stored in this system's database and are displayed in various websites. The audiences are targeted based on their preference and demography. By the process of data mining, webpages are classified into various categories based on their keywords extracted from them and are matched with the relevant ads in the database. Thus the companies can profit as they gain customers because of these ads and also the public is happy as they get their desired product.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ACKNOWLEDGEMENT	i
	ABSTRACT	ii
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	ix
1	INTRODUCTION	1
1.1	ADSENSE AND ADCHOICE	1
1.2	NAÏVE BAYES – DATA MINING	1
1.3	WEB ADVERTISING AND MONETIZATION	2
1.4	PURPOSE OF OUR PROJECT	2
1.5	SCOPE OF OUR PROJECT	3
2	LITERATURE SURVEY	4
2.1	TARGETED ADVERTISING FOR ONLINE SOCIAL NETWORKS	4
2.1.1	IMPLEMENTATION	4
2.1.2	LIMITATIONS	5
2.2	KEYWORD EXTRACTION FOR WEBPAGE CLUSTERS	5
2.2.1	IMPLEMETATION TECHNIQUE	5
2.2.2	LIMITATIONS	6

2.3	MINING TARGET USERS FOR ONLINE MARKETING BASED ON USER APP DATA	6
2.3.1	IMPLEMENTATION	7
2.3.2	LIMITATIONS	7
2.4	A PERSONALIZED PRODUCT SELECTION ASSISTANCE BASED ON E-COMMERCE MACHINE LEARNING	7
2.4.1	PERSONALIZED PRODUCT SELECTION DESIGN	7
2.4.2	LIMITATIONS	8
3	FUNCTIONAL REQUIREMENTS	9
3.1	ADVERTISER REGISTRATION	9
3.2	ADVERTISER LOGIN	9
3.3	PUBLISHER REGISTRATION	9
3.4	PUBLISHER LOGIN	10
3.5	ADVERTISEMENTS	10
3.6	WEBPAGE	10
3.7	DASHBOARD	10
4	HARDWARE AND SOFTWARE SPECIFICATIONS	11
4.1	HARDWARE SPECIFICATIONS	11
4.2	SOFTWARE SPECIFICATIONS	11

5	SYSTEM DESIGN	12
5.1	ARCHITECTURE	12
5.1.1	ENTITIES	12
5.1.2	DATA FLOW	13
5.2	PREDICTION USING NAIVE BAYES	14
5.2.1	DATA MINING	14
5.2.2	CLASSIFICATION	14
5.2.3	NAIVE BAYES	15
5.3	RANKING ALGORITHM	16
5.3.1	PARAMETERS	16
5.3.2	ALGORITHM	17
5.4	DATABASE MODELLING	17
6	IMPLEMENTATION	25
6.1	DATA COLLECTION AND PREPROCESSING	25
6.2	ADVERTISER AND PUBLISHER REGISTRATION AND LOGIN	27
6.3	NAÏVE BAYES PREDICTOR	33
6.4	RANKING ALGORITHM	34
6.5	TARGET USING ADVERTISEMENTS	36
7	CONCLUSION	38

8	FUTURE WORK	39
9	REFERENCES	40

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
5.1	Architecture Diagram	13
5.2	Posterior Probability Formula	15
5.3	Adcategory Table	18
5.4	Adcost Table	18
5.5	Adkeyword Table	19
5.6	Adpagemapping Table	20
5.7	Advertisement Table	20
5.8	Advertiser Table	21
5.9	Pagecategory Table	21
5.10	Pagekeyword Table	22
5.11	Publisher Table	22
5.12	Webpage table	23
5.13	Sequence Diagram	24
6.1	Advertiser Login	28
6.2	Advertiser Dashboard	28
6.3	Advertiser Registration	29
6.4	Add Advertisement	30
6.5	Publisher Login	31

6.6	Publisher Dashboard	31
6.7	Publisher Registration	32
6.8	Add Webpage	32
6.9	Webpage Model	33
6.10	Ad Model	34
6.11	Webpage with Ads	37

LIST OF ABBREVIATIONS

Acronym	Abbreviation
CPC	Cost Per Click
CPM	Cost Per Mille
CPI	Cost Per Impression
PPC	Pay Per Click
URL	Uniform Resource Locator
JSP	Java Server Pages
JVM	Java Virtual Machine
BLOB	Binary Large Object
AJAX	Asynchronous JavaScript and XML

CHAPTER 1

INTRODUCTION

1.1 ADSENSE AND ADCHOICE

Google uses its technology to serve advertisements based on website content, the user's geographical location, and other factors. Those wanting to advertise with Google's targeted advertisement system may enroll through Google AdWords. AdSense has become one of the popular programs that specializes in creating and placing banner advertisements on a website or blog, because the advertisements are less intrusive and the content of the advertisements is often relevant to the website. Many websites use AdSense to make revenue from their web content (website, online videos, online audio content, etc.), and it is the most popular advertising network. AdSense has been particularly important for delivering advertising revenue to small websites that do not have the resources for developing advertising sales programs and salespeople to seek out advertisers. To display contextually relevant advertisements on a website, webmasters place a brief Javascript code on the website's pages. Websites that are content-rich have been very successful with this advertising program, as noted in a number of publisher case studies on the AdSense website. Google has removed the policy of limiting AdSense ads to three ads per page. Now, AdSense publishers can place unlimited amount of AdSense ads on a page.

AdChoice is a subtype of AdSense where the advertisement is irrelevant of the webpage visited. The advertisements to display are controlled by the Advertiser itself and not by Google. This can lead to frustration as the user is unhappy to see the same advertisement again and again.

1.2 NAÏVE BAYES – DATA MINING

Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction. Naive Bayes classifiers are linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem, and the adjective naive comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption

is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives. Since it's based on probability it's very efficient while used for Text Classification. Naïve Bayes is the most sought out algorithm for Targeted Advertising due to its simplicity and speed.

1.3 WEB ADVERTISING AND MONETIZATION

Online advertising or Internet advertising or web advertising, is a form of marketing and advertising which uses the Internet to deliver promotional marketing messages to consumers. It includes email marketing, search engine marketing (SEM), social media marketing, many types of display advertising (including web banner advertising), and mobile advertising. Like other advertising media, online advertising frequently involves both a publisher, who integrates advertisements into its online content, and an advertiser, who provides the advertisements to be displayed on the publisher's content. Other potential participants include advertising agencies which help generate and place the ad copy, an ad server which technologically delivers the ad and tracks statistics, and advertising affiliates who do independent promotional work for the advertiser.

Website monetization is the process of converting existing traffic being sent to a particular website into revenue. The most popular ways of monetizing a website are by implementing Pay per click (PPC) and Cost per impression (CPI/CPM) advertising. Various ad networks facilitate a webmaster in placing advertisements on pages of the website to benefit from the traffic the site is experiencing. Hence AdSense and other E-Advertising companies use website monetization for their revenue.

1.4 PURPOSE OF OUR PROJECT

The system aims to do E-marketing by connecting the business advertisers with the public (customers) by displaying relevant and preferred online advertisements in Websites, resulting in website monetization. The people should not be frustrated by the banner ads. Instead their likes and preference are used to predict the advertisements and these are filtered the viewer's location.

1.5 SCOPE OF OUR PROJECT

The system can be used by both Advertiser who wants to advertise and promote their recent product and also a Publisher who wants to turn their webpage's traffic into good revenue. Advertiser can also keep track of how well their advertisement is performing. (i.e.) How many times the ad has been displayed and how many clicks have been made. The publisher can keep track of how many ads have been displayed in their webpage and how many advertisement clicks have been made.

CHAPTER 2

LITERATURE SURVEY

2.1 Pinaki Mitra, Kamal Baid “Targeted Advertising for Online Social Networks”, IEEE 2014

Generating targeted advertisements is a major problem with the monetizing activity in online social networks. This method applies this concept in social networks meeting important challenges. It performs an algorithm based on keyword clustering to generate ads and forums, marketplaces, groups on social networks are chosen as targets for these advertisements.

2.1.1 Implementation

In this system there are two main components. The first component is a crawler that crawls, cleans and ingests user posts from the social networking sites. The second component spots keywords in a post, compensates for misspellings and named entity variations and eliminates the off-topic content. The resulting sets of most relevant keywords in a post are provided to Ad programs for ad generation.

(i) Crawling user posts

This component is implemented using Java’s URL and regular expression packages, a fetcher gathers pages while a parser extracts the user posts, timestamps and category the post was crawled from. Crawled posts are ingested into text files stripping html tags and removing the images present or advertisements already found on the page.

(ii) Keywords for advertising

The main goal of this component is to provide relevant advertisements. It performs spotting keywords and eliminating off-topic noise. In this algorithm the first round uses only the Yahoo Term Extractor (YTE) to spot keywords. This algorithm also spots the variant of keyword and records the transliteration found in the first round.

2.1.2 Limitations

In this project of targeting advertisements for online social networks, they have come across the following draw backs:

- The keyword extraction method used here takes some time.
- The system doesnt target the user preference and location.

2.2 Vladimir Salin, Maria Slastihina, Ivan Ermilov, Ren´e Speck, Soren Auer, and Alexander Sytnik “Keyword Extraction for Webpage Clusters”

Processing the huge volume of unstructured information present in websites is important to distinguish different clusters of related webpages. This method uses keyword extraction, which utilizes two different clustering algorithms. The first algorithm is based on the analysis of the textual information, the second one is based on the statistical information that is obtained from the Google Analytics API.

2.2.1 Implementation technique

HTML markup with the textual content provides us with a compilation, which is easy to process with statistical data extraction methods. But information within webpages like figures, SVG documents, videos require additional processing before the statistical data extraction methods can be applied.

There are three modules in this method which are as follows

1. Data collection module

This module crawls the website from a user-defined URL. The output of this module is a directed weighted link graph.

2. Web clustering module

The website clustering module clusters the graph obtained by the data collection module. It implements the clustering approaches like Website link graph representation clustering and Statistical access log clustering

3. Data extraction module

This module evaluates the clusters of webpages from the second module. The keywords relevant to each webpage in a cluster are extracted.

- **Crawling**

The data collection module crawls a selected website and extracts information like webpage URL, links between webpages and webpage content. These information are extracted using the JSoup HTML parser.

- **Clustering**

The clustering is done either by graph clustering approach or statistical based clustering. The graph clustering method is based on the link graph model. The statistical based clustering group webpages using access log statistics. The keywords are extracted from the clusters of the grouped webpages.

2.2.2 Limitations

The limitations involved in this project are as follows:

- The percentage of webpages covered by this clustering module is insufficient
- The statistical clustering module works only for websites with access log information

2.3 Mi Xiuqiang He, Wenyuan Dai, Guoxiang Cao, Ruiming Tang, Mingxuan Yuan and Qiang Yang “Mining Target Users for Online Marketing based on App Store Data”, IEEE 2015

This project makes use of a technique avoiding traditional method of targeting advertisements based on search keywords and page visiting. Here the strategy is targeting the users based on their downloaded applications. This project implements an algorithm called xRank that efficiently lists top potential target users for the advertiser. It finds precise target group of users compared to current system of online marketing.

2.3.1 Implementation technique

This system makes use of the app-user relationship knowledge for online advertisement marketing. The system gets input from the advertisers a set of applications that the target users may download then the remaining process is carried out by xRank algorithm. xRank is designed based on the Random Walk mechanism to reduce the sparseness of exact match of the app list. The apps installed by users potentially reflect their interests and preferences. xRank improves the PageRank algorithm. Then the xRank algorithm is used to group the users and it lists the potential list of users who are used as targets then for the advertisements

2.3.2 Limitation

This system has a major limitation which is Advertisers have to find and seed particular apps that has more probability of being widely downloaded by the target users which involves again a huge manual work.

2.4 Hong-we1 Yang, Zhi-geng Pan, Xi-zhao Wang, Bing Xu “A Personalized Products Selection Assistance Based On E-commerce Machine Learning”, IEEE 2005.

This system makes use of the consumers experience information overload and look for help in selecting from an overwhelming array of products. In order to overcome such a problem one option is to develop a personalized online assistance to retrieve product information that is related to the customers. This system combines the genetic algorithm and k nearest neighbour technology to reason about the customer’s personal preferences and then provide the most appropriate products to meet their demand and preference.

2.4.1 Personalized Product Selection design

An user profile is constructed by tracking the browsing and purchasing behaviour of the customer. The appropriate products are selected for the customer, based on the similarity measure technology.

The three steps involved in this process are

i. User profile acquisition

The user profile is constructed in the personalized products selection in the first step. The user profile recommends system to analyse each customer's demand and preference by tracking each customer's browsing information in order to make preparation for product selection.

ii. Dialog design

Dialog-based system offers guidance by asking goal-directed questions and presenting product alternatives. Designing the dialog is the most useful since it helps finding the accurate products. One possible method is to use the expected information gain by interaction.

iii. Display selected products

The selected products are displayed for the customer and also the customers can modify it by changing the weights of each attribute so that the system finds the most favourable product.

2.4.2 Limitation

This system has a major limitation which is in learning consumers interest, it is observed that the weights obtained from the k-NN algorithm often show warp because of the high correlation of the properties in the products which is a serious issue.

CHAPTER 3

FUNCTIONAL REQUIREMENTS

3.1 ADVERTISER REGISTRATION

The advertiser who would like to advertise their product online, can do their registration in this module. After registration the advertiser can access other functionalities like add advertisement, delete advertisement and modify advertisement. While registration, advertiser can provide their company details including their membership type (i.e.) either normal or premium. Premium user's ads are given more preference than normal user's ads.

3.2 ADVERTISER LOGIN

Advertisers, after their successful registration they can log in by using their registered email id and password. After logging in they can do the following functions.

(i) ADD PRODUCT

The advertiser can add any number of products to the advertisement list. The product keywords, description, image to be displayed are all entered.

(ii) DELETE PRODUCT

The advertiser can delete their product from their advertisement list.

(iii) MODIFY PRODUCT

The advertiser can make the required modification to the previously added advertisement.

3.3 PUBLISHER REGISTRATION

The webmaster or the publisher who wishes to display advertisements in their webpage for generating revenue can do the registration in this module. Once they add webpages the Ad engine starts its prediction to display ads in the webpage. Publisher info include their website domain and age.

3.4 PUBLISHER LOGIN

Publisher, after their successful registration they can log in by using their registered email id and password. After logging in they can do the following functions.

(i) ADD WEBPAGE:

The publisher can add their webpage by providing its details. The webpage details should include URL and domain name.

(ii) DELETE WEBPAGE:

The publisher can delete their webpage from their webpage list.

(iii) MODIFY WEBPAGE:

The publisher can make the required modification to the previously added webpage list.

3.5 ADVERTISEMENTS

Every advertisement belongs to an advertiser and it comes under a category. The category will be predicted based on the keywords entered by the advertiser. The advertisement specifies CPC and CPM. The advertisement is ready to pay that much amount to the system once their advertisements are displayed and clicked.

3.6 WEBPAGE

Every webpage belongs to a Publisher and it falls under a predefined category. The category will be predicted by the system using a classifier model. Each webpage will have a limitation to display advertisements.

3.7 DASHBOARD

Dashboards are available for both advertiser and publisher to keep track of their products and webpages. The count of ads displayed and clicked and revenue generated are shown here.

CHAPTER 4

HARDWARE AND SOFTWARE SPECIFICATIONS

4.1 HARDWARE SPECIFICATIONS

An AdServer with multithreading capacity is required to process multiple requests from different webpages. The server should have a RAM memory not less than 16 GB. Based on the no of advertisements stored in the Database the secondary memory is chosen. Uninterrupted Internet connection with bandwidth greater than 40Mbps is required for the server to run efficiently.

4.2 SOFTWARE SPECIFICATIONS

JVM with JDK and JRE packages is required. Any Java IDE with Dynamic Webpage support is used to run the JSP. Apache Tomcat 7.0.56 is used as a server. MySQL database is used by the AdServer. Apart from these Weka, JSOUP, Apache POI libraries are required for the JAVA program.

CHAPTER 5

SYSTEM DESIGN

5.1 ARCHITECTURE

The system interacts with two entities. They are Advertiser and Publisher. The system is called the AdServer as it controls the display of ads based on certain parameters.

5.1.1 ENTITIES

Advertiser

The company or the Business People who wants to promote or advertise their product enrolls in our system beforehand. They pay money based on No of Clicks (CPC) and No of Impressions (CPM) for every advertisement. They are shown with the progress of how much their Ads have been displayed. All the ads are given equal chance.

Publisher

The website which wants to make revenue by displaying banner ads in their webpage enrolls in our system beforehand. Based on their traffic (i.e.) No of people visiting the webpage to view its content, will be targeted with advertisements. Money is given to the webpages depending upon how many viewers visit the page and how many click the ads.

Central AdServer

This is our system where advertisers logins and uploads the ads and publisher logins and uploads the webpages. This takes care of predicting the exact category of an ad or webpage during registration. It also performs ranking of ads to determine which ad to display at a particular time for a viewer.

Viewer (Public)

This is the person who visits the publisher's webpage. Hence we have to predict his behavior and filter ads based on his location.

Database

Contains ads along with their description, advertiser id, keywords and image to display. Also contains webpages along with their category, URL and no of ads to display in a particular webpage. For each ad CPC, CPM, Clicks and Impression counts are maintained by updating regularly.

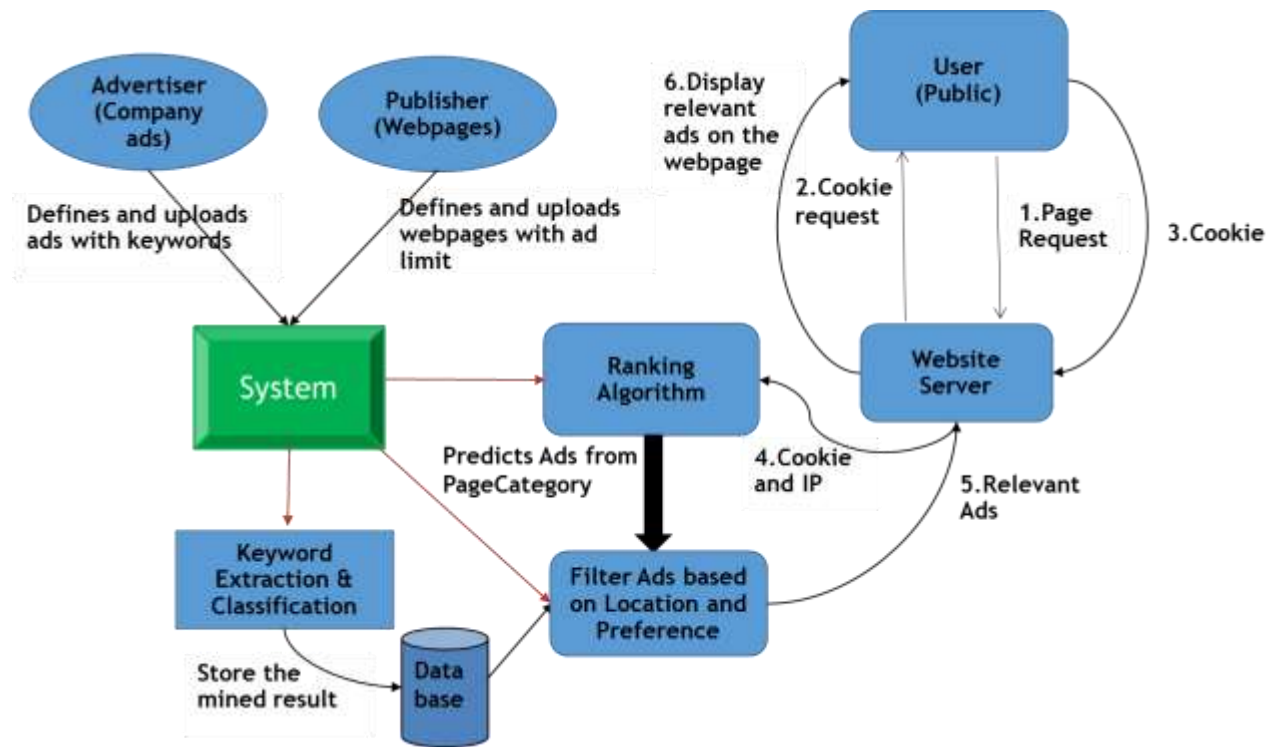


Figure 5.1 Architecture Diagram

5.1.2 DATA FLOW

The publisher and the advertiser uploads webpages and ads details to the system. The system classifies them based on Naïve Bayes model and stores them in the database along with the classified category.

1. The visitor wants to visit the webpage.
2. The webpage requests cookie from him.
3. The cookie is accessed by the website.
4. The webpage forwards its URL, IP address of the visitor and his cookie to the System.

5. The System predicts what ads to display and filters the ads based on the user's location obtained from his IP.
6. The filtered ads are displayed in the webpage.

5.2 PREDICTION USING NAÏVE BAYES

5.2.1 DATA MINING

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining involves six common classes of tasks:

1. Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
2. Association rule learning (Dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
4. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
5. Regression – attempts to find a function which models the data with the least error.

5.2.2 CLASSIFICATION

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given

patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

The Data Classification process includes two steps –

- Building the Classifier or Model
- Using Classifier for Classification

1. Building the Classifier or Model

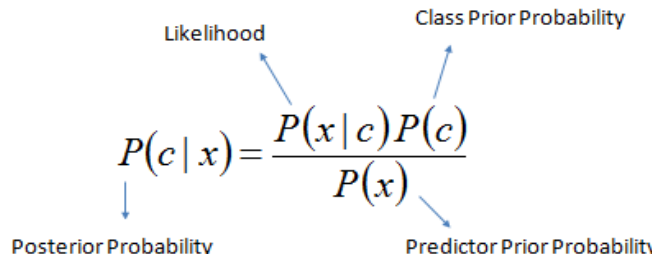
This step is the learning step or the learning phase. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

2. Using Classifier for Classification

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

5.2.3 NAÏVE BAYES

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.



$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$

Figure 5.2 Posterior Probability Formula

Bayes theorem provides a way of calculating the posterior probability, $P(c/x)$, from $P(c)$, $P(x)$, and $P(x/c)$.

Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

- $P(c/x)$ is the posterior probability of *class (target)* given *predictor (attribute)*.
- $P(c)$ is the prior probability of *class*.
- $P(x/c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

5.3 RANKING ALGORITHM

The amount of money paid by the advertiser for every click on the advertisement. It's a performance-based metric. This means the Publisher only gets paid when (and if) a user clicks on an ad, no matter how many impressions they serve trying to get the click. This favours the advertiser.

5.3.1 PARAMETERS

CPC (Cost Per Click)

The amount of money paid by the advertiser for every click on the advertisement. It's a performance-based metric. This means the Publisher only gets paid when (and if) a user clicks on an ad, no matter how many impressions they serve trying to get the click. This favours the advertiser.

CPM (Cost Per Mille)

It's when the price is based on 1,000 impressions. Almost all Publishers prefer to bill on impressions because it is an inventory based product, rather than a performance based product.

Premium Ad

The advertiser who uploads the ad can be a premium member or non-premium member. For premium membership yearly charges needs to be paid by the advertiser to our system.

5.3.2 ALGORITHM

The ranking takes place by,

$\text{Min (CPC/Clicks + CPM/Impressions) * Cost Factor}$

Where, Cost Factor = 1 for Non-premium Advertisers.

Cost Factor = 1.5 for Premium Advertisers.

1. Whenever a webpage request is made by the visitor the system runs a query to the local database and fetches the related ads that are capable to get displayed. These ads are filtered based on the visitor's location.
2. If the ads obtained is less than $2 * \text{No of ads supported by the webpage}$ then the global advertisements are also included to make the count greater than $2 * \text{No of ads}$.
3. These ads will be ranked by the above algorithm and the first $2 * \text{No of}$ are chosen by the system and are passed to the webpage for displaying.

5.4 DATABASE MODELLING

The database modelling deals with structure of the relational database. It shows the relationship between the table and columns in each field.

ADCATEGORY

Adcategory table contains three fields id, name and parent id. Id is the primary key. Name is the name of the category and parent id is the id of the category in which it comes under. A category will have n number of child categories.

Id	Name	Parent_Id
1	Arts & Entertainment	0
2	Autos & Vehicles	0
3	Beauty & Fitness	0
4	Books & Literature	0
5	Business & Industrial	0
6	Computers & Electronics	0
7	Finance	0
8	Food & Drink	0
9	Games	0
10	Hobbies & Leisure	0

Figure 5.3 Adcategory Table

ADCOST

Adcost table contains advertisement id which is the primary key, Clicks which refers to how many time the advertisement is clicked, Impressions used to quantify the display of an advertisement on a web page, CPM which refers to the *cost* an advertiser pays for one thousand views or impressions of an advertisement, CPC which an advertiser pays a publisher.

Advertisement_Id	Clicks	Impressions	CPM	CPC
1	100	1000	1.20	2.50
2	200	1200	1.50	3.10
3	300	1400	1.80	3.70
4	400	1600	2.10	4.30
5	500	1800	2.40	4.90
6	600	2000	2.70	5.50
7	700	2200	3.00	6.10
8	800	2400	3.30	6.70
9	900	2600	3.60	7.30
10	1000	2800	3.90	7.90

Figure 5.4 Adcost Table

ADKEYWORD

This table contains the splitted individual keywords from the various webpage along with its id and its category which is related by the adcategory table. This table contains 3 fields, id which is the primary key, Category to which category a particular advertisement fall and keyword.

Id	Category	Keyword
1	98	games
2	98	online
3	98	shooting
4	98	multiplayer
5	98	games
6	98	online
7	98	racing
8	98	games
9	98	online
10	98	racing

Figure 5.5 AdkeywordTable

ADPAGEMAPPING

The adpagemapping table relates the each advertisement with a webpage. Hence this decides the advertisement which should be displayed in a webpage. It contains id which is the primary key, AdCategory which refers to id in adcategory table, PageCategory which refers to the id in pagecategory table.

id	AdCategory	PageCategory
119	1	1
120	1	14
121	1	15
122	1	16
123	1	17
124	5	2
125	5	18

Figure 5.6 Adpagemapping Table

ADVERTISEMENT

The advertisement table contains Id which is a unique field, Advertiser_id which a foreign key to the advertiser table which shows who make the advertisement, Category which refers to which category to which it falls, Location which defines in which part of the globe the advertisement can be displayed, image which is BLOB file maintains the picture of the advertisement, imageurl which contains path location of an image.

Id	Advertiser_Id	Keywords	Category	Location	image	imageurl
1						
1	2	games+online+shooting+multiplayer	98	global	[BLOB - 11.7 KiB]	C:\Users\Madhan AMK\Desktop\CRM\image\game (1).jpg
2	2	games+online+racing	98	global	[BLOB - 10.5 KiB]	C:\Users\Madhan AMK\Desktop\CRM\image\game (2).jpg
3	2	games+online+racing+car	98	global	[BLOB - 22.3 KiB]	C:\Users\Madhan AMK\Desktop\CRM\image\game (3).jpg
4	2	games+online+card+rummy	98	global	[BLOB - 13.4 KiB]	C:\Users\Madhan AMK\Desktop\CRM\image\game (4).jpg
5	2	games+online+roleplay	98	global	[BLOB - 7 KiB]	C:\Users\Madhan AMK\Desktop\CRM\image\game (5).jpg

Figure 5.7 Advertisement Table

ADVERTISER

The advertiser table contains Id which is a unique field, Company name, CEO name, company age, company location, Subscriber mail id, Subscriber password, Premium membership which shows whether the advertiser is a normal or a premium member.

Id	Company_Nam	CEO_Nam	Company_Ag	Company_Loc	Subscriber_Mail_Id	Subscriber_P	Premium_Membership
1	OLA	Benjamin	10	global	ola@gmail.com	ola	1
2	Alienware	Franklin	6	canada	alien@gmail.com	alien	0
3	Annapoorna	John	20	australia	annapoorna@gmail	annapoorna	1
4	artmart	Vasan	12	india	artmart@gmail.com	artmartvasan	1
5	Entertainwala	Jennifer	20	usa	entertainwala@gma	entertainwala	0
6	Decathlon	mark henry	12	global	decathlon@gmail.co	decathlon	1
7	Amazon	panda	30	global	amazon@gmail.com	amazon	0
8	LinkedIn	jeffrey	6	india	linkedin@gmail.com	linkedin	1
9	GE	madhan	20	india	ge@gmail.com	ge	0
10	Builders	louis	50	global	build@gmail.com	build	1

Figure 5.8 Advertiser Table

PAGECATEGORY

Pagecategory table contains three fields id, name and parent id. Id is the primary key. Name is the name of the category and parent id is the id of the category in which it comes under. A category will have n number of child categories.

Id	Name	Parent_Id
1	Arts & Humanities	0
2	Business, Economy & Industry	0
3	Company Web Sites	0
4	Digital Society	0
5	Government, Law & Politics	0
6	Law and Legal System	0
7	Medicine & Health	0
8	Popular Science	0
9	Publishing, Printing and Bookselling	0
10	Science & Technology	0

Figure 5.9 Pagecategory Table

PAGEKEYWORD

Pagekeyword table contains Id which is a unique field, Primary category, secondary category which is a sub field to the Primary Category and Keyword of the pages.

Id	Primary_Category	Secondary_Category	Keyword
1	Arts & Humanities	Local History	uel
2	Sports and Recreation	British Countryside	domain
3	Government, Law & Politics	Central Government	politics
4	Government, Law & Politics	European Parliament Elections 2009	socialist
5	Science & Technology	Energy	nuclear
6	Sports and Recreation	British Countryside	ancient
7	Sports and Recreation	British Countryside	japanese
8	Sports and Recreation	British Countryside	festival
9	Sports and Recreation	British Countryside	onbashira
10	Science & Technology	Energy	laboratory

Figure 5.10 Pagekeyword Table

PUBLISHER

Publisher table contains Id which is a unique field, Domain i.e. name of their domain, Website age, admin_email_id, admin_password and their account_number.

Id	Domain	Website_Age	Admin_Email_Id	Admin_Password	Account_No
1	flipkart	20	flipkart@gmail.com	flipkart	1212312
2	uber	20	uber@gmail.com	uber	618923
3	games	40	games@gmail.com	games	1892378
4	dominos	29	dominos@gmail.com	dominos	1892371892
5	yahoo	50	yahoo@gmail.com	yahoo	2342342

Figure 5.11 Publisher Table

WEBPAGE

Webpage table contains Id which is a unique field, Publisher_id, Category to which that particular webpage falls, Page_url ie the page it should be redirected on matching the advertisement, No_of_Ads which represent the maximum number of advertisement that the page can display.

Id	Publisher_Domain	Category	Page_URL	No_Of_Ads
1	flipkart	6	http://localhost:8080/Advertising/req.html	3
2	uber	13	http://localhost:8080/Advertising/taxi.html	3
3	games	2	http://localhost:8080/Advertising/games.html	4
4	dominos	7	http://localhost:8080/Advertising/health.html	5
5	rank	98	http://localhost:8080/Advertising/web/req.html	3

Figure 5.12 Webpage Table

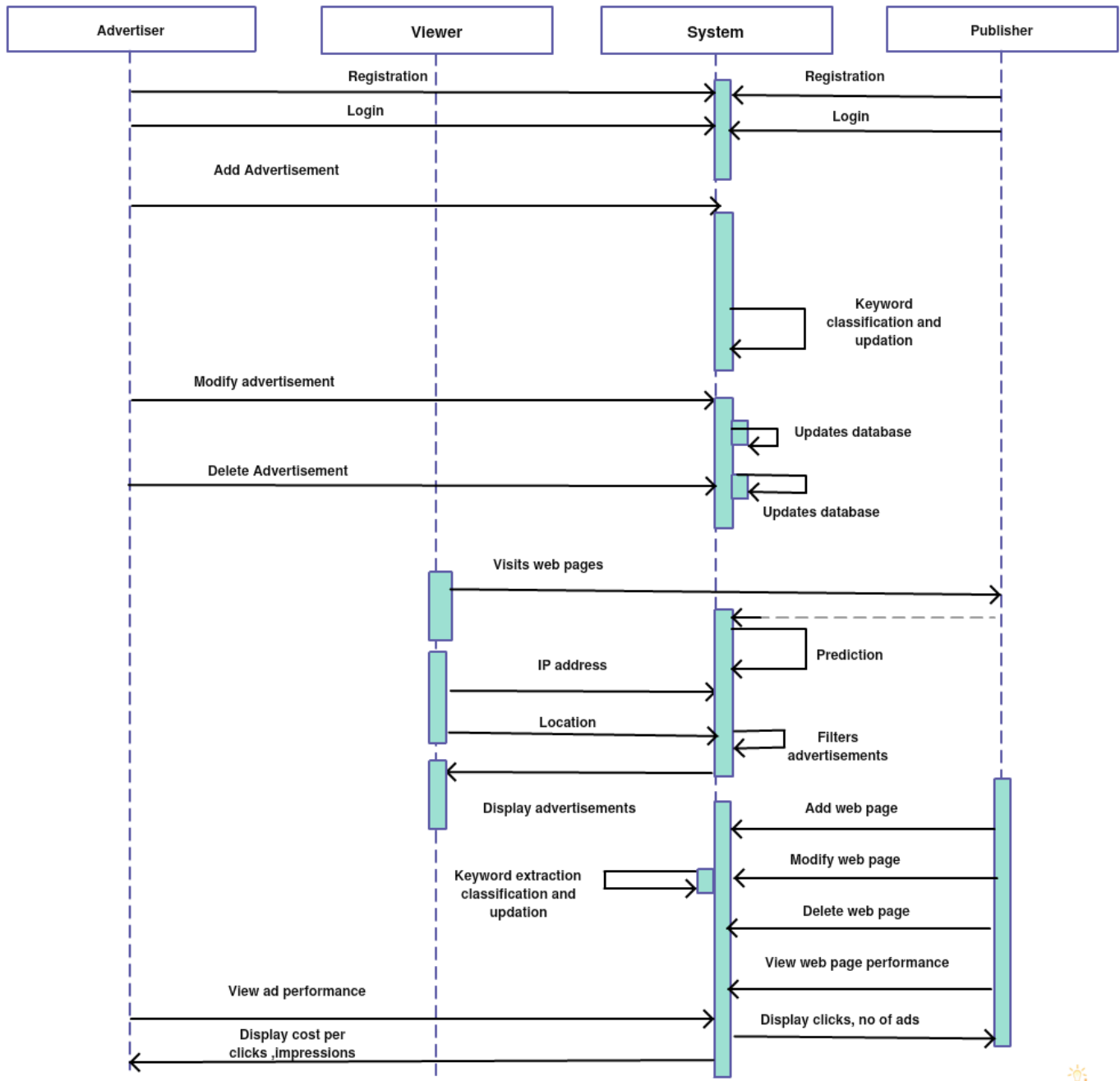


Figure 5.13 Sequence Diagram

CHAPTER 6

IMPLEMENTATION

6.1 DATA COLLECTION AND PREPROCESSING

Data collection through various resources have been carried out both manually and through crawler. Google has provided standard categories for advertisements so that even advertisements that will be added in future will fall under these categories. Thus with respect to the topics of ads provided by Google, categories are built in our project to personalize them. A category consists of subcategories and the advertisements are defined with the keywords under each category. In our project, to train our system a table is built in the database in the name of advertisement which consists of the combination of keywords specifying the product advertisement. It is constructed manually with the combination of keywords for that advertisement. Almost all the standard categories defined by Google are used in our project and filled with corresponding keywords. Duplicate entries are also removed and the collected data has undergone screening. In order to post advertisements that are relevant to the information the user is looking for in the webpage. The keywords of the webpage are obtained using crawlers and these collected keywords are updated in the page keyword table of the database. The crawling is done by the module Webpage Keyword Extractor. Data Pre-processing is an important activity in this module as the collected data might have noise, missing values, errors and outliers. So the collected data are subjected to data cleaning in pre-processing. Also the common stop-words present with the keywords are filtered to make the system perform efficiently. Stop-words are the common English words that may mislead the prediction. The modules that are involved in the data collection and pre-processing activities in the system are explained.

6.1.1 Webpage Keyword Extractor

This module makes use of the JSOUP parser to parse the html document of the webpages. Jsoup is a java html parser. It uses API to extract and manipulate data from URL or HTML file. This module also implements the infrastructure of parallelism and structuring tasks. Executor

service has been used here to manage and schedule several threads. Now the system manages and runs multiple threads in the same time in an asynchronous manner. The keywords are present in the keyword met tag in the html file of that webpage delimited by comma. The Jsoup parser initially parses the content present in the keyword Meta tag. This module also creates a work book instance holding reference to the xlsx file. The parsed keywords are separated using string tokenizer and are added in to the tree map. After subjecting them to screening and various data cleansing activities they are added to a new workbook. Then Java database connection is established and these keywords are updated in the database. As these processes run in multiple threads simultaneously the time of execution is reduced a lot.

6.1.2 Keyword Splitter

This module creates a workbook instance holding reference to the xlsx file. It constructs Tree map to maintain the details like id, primary category, and secondary category. The keywords are split using the string tokenizer and checked before adding to the tree map. Then this module writes the keywords that are split into a new workbook.

6.1.3 Inserting details from workbook to Database

This module initially creates a workbook instance holding reference to the xlsx file. Then it performs Java Database Connectivity. Iterators are used and the values from the cell values are obtained. These obtained cell values are set into corresponding data types and updated to the database tables using the sql queries.

6.1.4 Delimiter

This module creates a workbook instance holding reference to the xlsx file. It makes use of the iterator to iterate through each row and column. It takes the category value from the first cell. It obtains the set of keywords and splits them using string tokenizer delimited by space. The category and keyword values are stored in a tree map. Then the keys and values from the tree map are moved to another workbook.

6.1.5 Advertisement Delimiter

This module's main functionality is to collect all the keywords of a particular advertisement which are delimited by plus symbol. It creates a workbook instance holding reference to the xlsx file. Then it iterates through each row and column. It obtains category value and string containing all the keywords of that particular advertisement from the cell. The keywords from the cell are delimited by plus symbol. They are segregated using string tokenizer and stored in a tree map. The keywords and category of that advertisement are now stored in another workbook.

6.1.6 Stop-words filter:

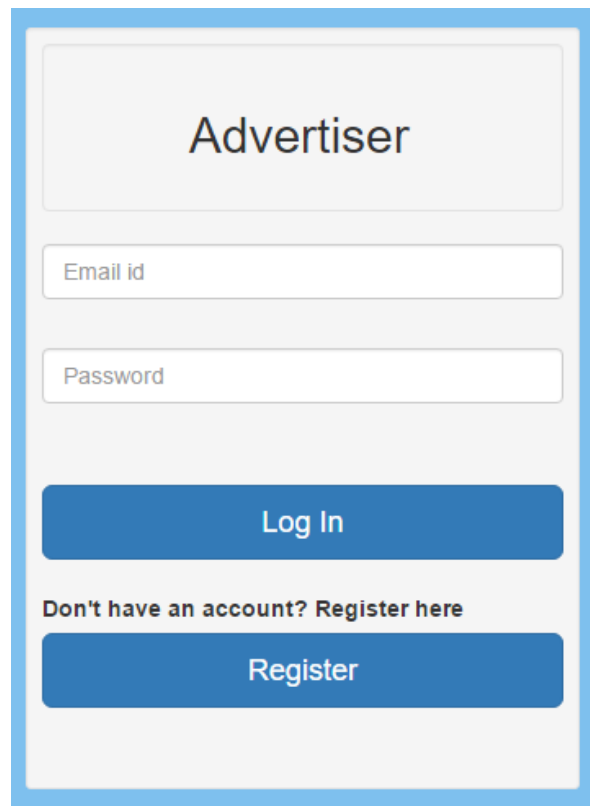
This module blacklists a set of certain common words that are present with the keywords that are extracted. In most situations the keywords extracted from the websites using crawler contain common stop-words that not useful. So the module uses iterator to iterate through each cells. Then the primary category, secondary category of the advertisement are collected. Initially the module creates an array of blacklisted stop-words that are to be filtered from the keywords. Then it checks the keywords if any blacklisted common word is present. Then the filtered keywords are added to new workbook.

6.2 ADVERTISER AND PUBLISHER – REGISTRATION AND LOGIN

Advertiser can register himself by providing the details of his company and he can opt for premium membership or not. After registering he should login and a dashboard is displayed for him. The dashboard contains sub-options like Add Ad, Delete Ad, Modify Ad, and View Ad Performance. The ad performance contains the details of no of clicks and no of impressions. Based on the no of clicks and impressions he has to pay money for our system.

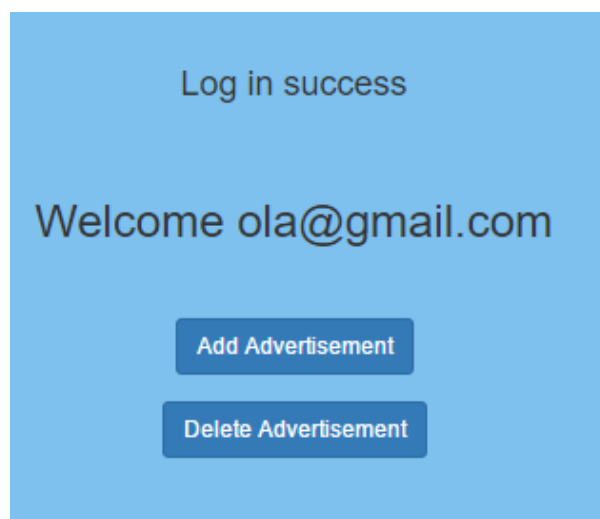
Publisher can register himself by providing the details of his website. After registering he should login and a dashboard is displayed for him. The dashboard contains sub-options like Add Webpage, Delete Webpage, Modify Webpage, and View Webpage Performance. The no of ads displayed in the webpage and clicks made are shown in the dashboard.

CSS have been used to create the front end. JSP pages are used to connect to the MySQL database to insert, delete and update records.



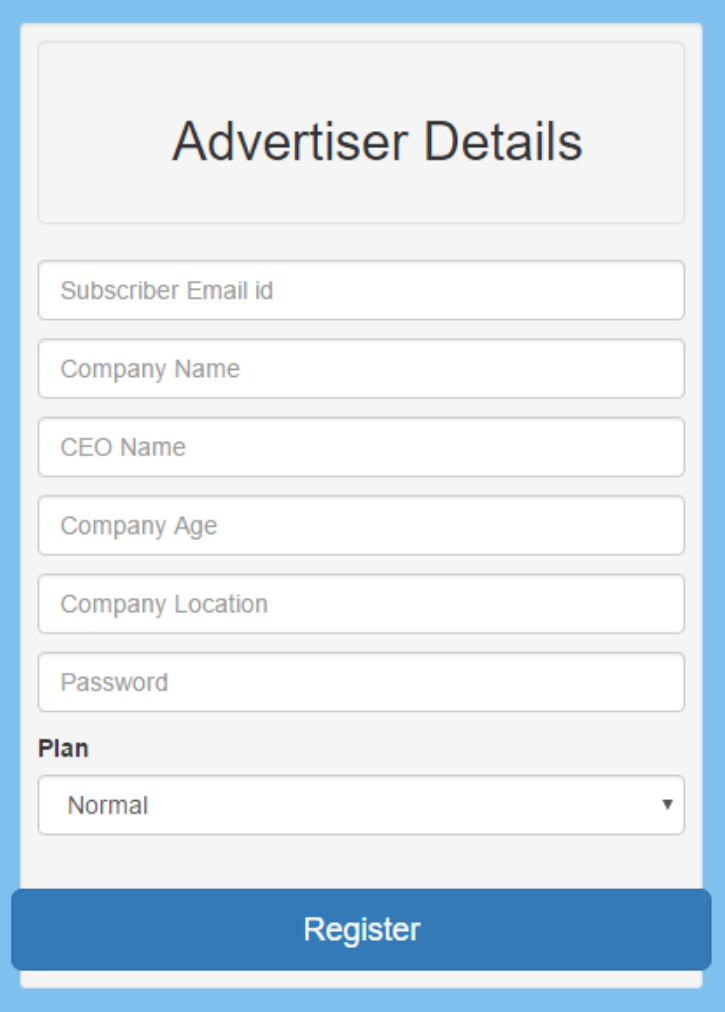
The image shows a login form for an advertiser. It is titled "Advertiser" in a large, bold, black font. Below the title, there are two input fields: "Email id" and "Password". Below these fields, there is a blue button labeled "Log In". Below the button, there is a link that says "Don't have an account? Register here". Below the link, there is another blue button labeled "Register". The entire form is enclosed in a light blue border.

Figure 6.1 Advertiser Login



The image shows a dashboard for an advertiser. It has a light blue background. At the top, it says "Log in success". Below that, it says "Welcome ola@gmail.com". At the bottom, there are two blue buttons: "Add Advertisement" and "Delete Advertisement".

Figure 6.2 Advertiser Dashboard



The image shows a registration form titled "Advertiser Details" enclosed in a light blue border. The form contains several input fields: "Subscriber Email id", "Company Name", "CEO Name", "Company Age", "Company Location", and "Password". Below these is a section labeled "Plan" with a dropdown menu currently showing "Normal". At the bottom of the form is a large blue button with the text "Register".

Advertiser Details

Subscriber Email id

Company Name

CEO Name

Company Age

Company Location

Password

Plan

Normal ▼

Register

Figure 6.3 Advertiser Registration

Add advertisement

Keywords :

Add keyword

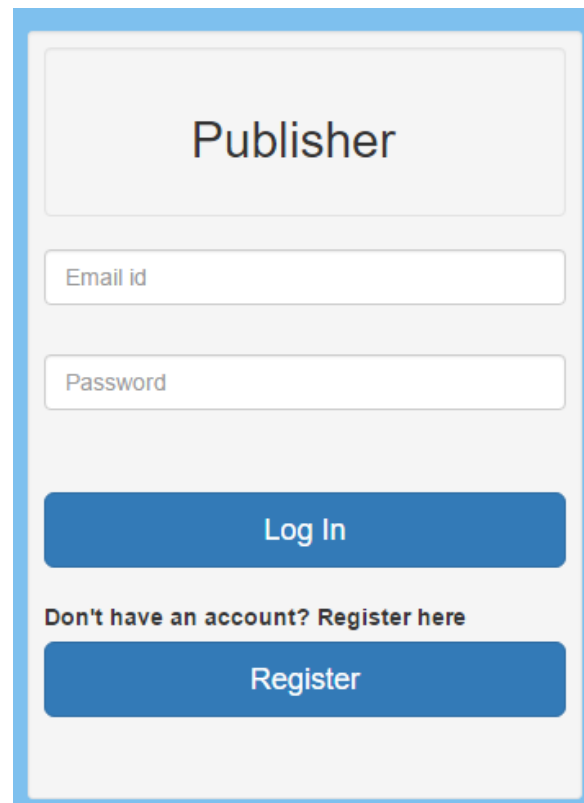
Primary Category

Arts & Entertainment ▼

Finalize keyword

Submit

Figure 6.4 Add Advertisement



The image shows a login form for a 'Publisher' role. It is enclosed in a light blue border. At the top, the word 'Publisher' is centered in a light gray box. Below this, there are two input fields: 'Email id' and 'Password'. Under the password field is a blue 'Log In' button. Below the button is the text 'Don't have an account? Register here' followed by a blue 'Register' button.

Publisher

Email id

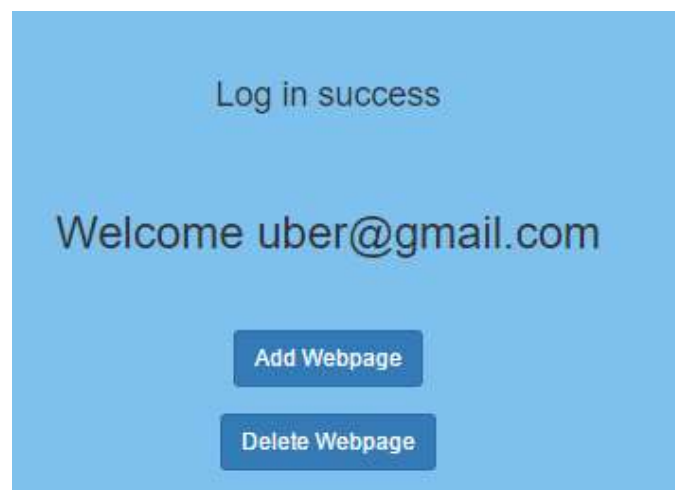
Password

Log In

Don't have an account? Register here

Register

Figure 6.5 Publisher Login



The image shows a dashboard for a 'Publisher' user. It has a solid blue background. At the top, it says 'Log in success'. Below that, it says 'Welcome uber@gmail.com'. At the bottom, there are two blue buttons: 'Add Webpage' and 'Delete Webpage'.

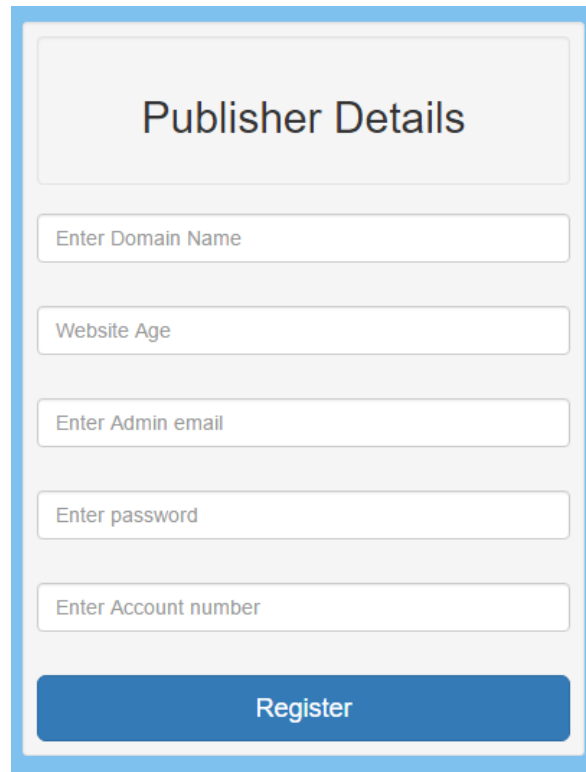
Log in success

Welcome uber@gmail.com

Add Webpage

Delete Webpage

Figure 6.6 Publisher Dashboard



A registration form titled "Publisher Details" with a light gray background and a blue border. It contains five text input fields: "Enter Domain Name", "Website Age", "Enter Admin email", "Enter password", and "Enter Account number". A blue "Register" button is at the bottom.

Publisher Details

Enter Domain Name

Website Age

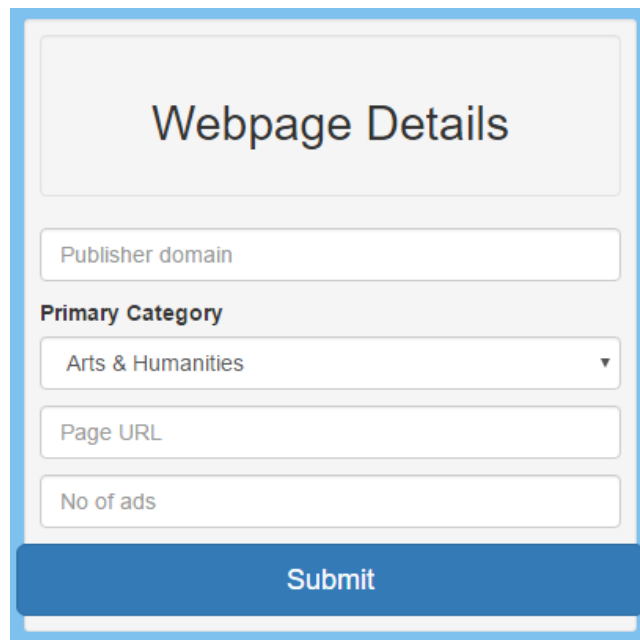
Enter Admin email

Enter password

Enter Account number

Register

Figure 6.7 Publisher Registration



A form titled "Webpage Details" with a light gray background and a blue border. It contains four input fields: "Publisher domain", a dropdown menu for "Primary Category" (currently showing "Arts & Humanities"), "Page URL", and "No of ads". A blue "Submit" button is at the bottom.

Webpage Details

Publisher domain

Primary Category

Arts & Humanities ▼

Page URL

No of ads

Submit

Figure 6.8 Add Webpage

6.3 NAÏVE BAYES PREDICTOR

Weka Library is used to perform classification of the advertisement and webpage category. While adding an ad the advertiser enters the keywords for an ad which is limited to 5. Also a drop-down list box is provided with the primary category items. Once he submits the details our system predicts the secondary category of the ad using two parameters- keywords and primary category. Since Naïve Bayes is a probabilistic classifier text classification is much accurate in it when compared to other classifiers.

Similarly when a webpage is added by the publisher only the primary category is obtained from him. Using JSOUP library the added webpage's html page is parsed to extract the keywords of the webpage. These keywords provide means of contextual mining. Weka library is used to predict secondary category from the primary category and the extracted keywords. The predicted result is stored in the database along with the other details of the webpage.

Using the ad keywords and webpage keywords present already in the database, two naïve bayes models are created. AdKeyword model and PageKeyword model. The class index is set as Secondary category for both the models as it is the one to be predicted. Then Naïve Bayes classifier is used to classify the new data using the already created model. The result of the classifier is stored in the database under the Category field.

Classifier Engine

```
query = new InstanceQuery();
query.setUsername("root");
query.setPassword("");
query.setQuery("select Primary_Category,Secondary_category,Keyword from pagekeyword");
PageKeywordRecords= query.retrieveInstances();
PageKeywordRecords.setClassIndex(PageKeywordRecords.numAttributes()-2);
PageKeywordModel= new NaiveBayes();
PageKeywordModel.buildClassifier(PageKeywordRecords);
```

Figure 6.9 Webpage Model

```

private static void CreateAdModel()
{
    InstanceQuery query;
    try{
        query = new InstanceQuery();
        query.setUsername("root");
        query.setPassword("");
        query.setQuery("select c.Name as PC,b.Name as SC ,Keyword from adkeyword a,
        adcategory b,adcategory c where a.Category=b.Id and b.Parent_Id=c.Id and c.Parent_Id=0");
        AdKeywordRecords= query.retrieveInstances();
        AdKeywordRecords.setClassIndex(AdKeywordRecords.numAttributes()-2);
        AdKeywordModel= new NaiveBayes();
        AdKeywordModel.buildClassifier(AdKeywordRecords);
    }
    catch(Exception e)
    {
        e.printStackTrace();
    }
}
private static void EvaluateModel(NaiveBayes model,Instances data)
{
    try{
        Evaluation eval = new Evaluation(data);
        eval.evaluateModel(model, data);
        System.out.println(eval.toSummaryString());
    }
    catch(Exception e)
    {
        e.printStackTrace();
    }
}
}

```

Figure 6.10 Ad Model

6.4 RANKING ALGORITHM

Adcost table contains the details of how much an advertiser will spend for an ad. CPC and CPM are the two cost measures which denotes the cost paid for clicks and impressions respectively.

Webpage table contains a column called No of ads. It denotes the max no of ads that can be displayed in a webpage and its chosen by the publisher. The publisher's revenue is highly affected by this parameter.

A javascript snippet is given to the webpage at the time of his registration. This code will display the ad images in his site which are given by our system.

When a visitor visits a webpage that has been registered with us already, the JS in the webpage fetches his IP address and sends an AJAX request to our rank.jsp page. This page is responsible for ranking and filtering the ads by ensuring equal opportunities to all advertisements.

The system (rank.jsp) first checks whether the incoming request page has been signed up with us or not. This ensures security as hackers who are trying to access our advertisements are prevented. After this based on the no of ads displayable on the webpage the ads in the database are ranked and filtered.

The requesting webpage's category is found from the database. Its already predicted by our system. Then the corresponding ad category which are mapped to this webpage's category are obtained from the adpagemapping table.

The ads which have the categories same as that of the obtained adcategories from the adpagemapping table are taken out from the database for ranking.

These ads are ranked based on 5 parameters- CPC, CPM, No of clicks, No of impressions and Premium Ad.

The ranking algorithm,

$\text{Min}(\text{CPC/Clicks} + \text{CPM/Impressions}) * \text{Cost Factor}$

Where, Cost Factor = 1 for Non-premium Advertisers.

Cost Factor = 1.5 for Premium Advertisers.

1. Whenever a webpage request is made by the visitor the system runs a query to the local database and fetches the related ads that are capable to get displayed. These ads are filtered based on the visitor's location. The query is

“Select distinct(Id) from Advertisement where Category in (select AdCategory from adpagemapping where PageCategory in (select Category from webpage where Id = page_id)) or Category in (select Id from adcategory where Parent_Id in (select AdCategory from

adpagemapping where PageCategory in (select Category from webpage where Id = page_id)))and location like location”

2. If the ads obtained is less than $2 * \text{No of ads supported by the webpage}$ then the global advertisements are also included to make the count greater than $2 * \text{No of ads}$. The query is
“Select distinct(Id) from Advertisement where Category in (select AdCategory from adpagemapping where PageCategory in (select Category from webpage where Id = page_id)) or Category in (select Id from adcategory where Parent_Id in (select AdCategory from adpagemapping where PageCategory in (select Category from webpage where Id = page_id)))and location in (location,global)”

3. These ads will be ranked by the above algorithm and the first $2 * \text{No of}$ are chosen by the system and are passed to the webpage for displaying. The query is
“Select Advertisement_Id,a.CPC,a.CPM,a.Clicks,a.Impressions,c.Premium_Membership from adcost a, advertisement b,advertiser c where a.Advertisement_Id=b.Id and b.Advertiser_Id = c.Id and b.Id in adlist”

4. Whenever the ad is displayed its Impressions count is updated in the database. Similarly whenever its clicked the Clicks count is updated.

6.5 TARGET USING ADVERTISEMENTS

Once the viewer opens a webpage which has been previously registered by its webmaster (publisher) in the system, the system begins to predict ads that will suit the user. Webpage category is taken as the user’s preference and the user’s IP address is used to get his location. The ads are displayed along with the webpage at the sides. The displayed ads location or content cannot be accessed in html page. All the data of ads comes from the JSP page in a response object. That object cannot be manipulated and hence security side of the system is taken care of. The ads change once in 5 secs inorder to tempt the user. If the user clicks the ad image the company’s website is opened along with the product’s specification and the click count is updated in the database by the system.

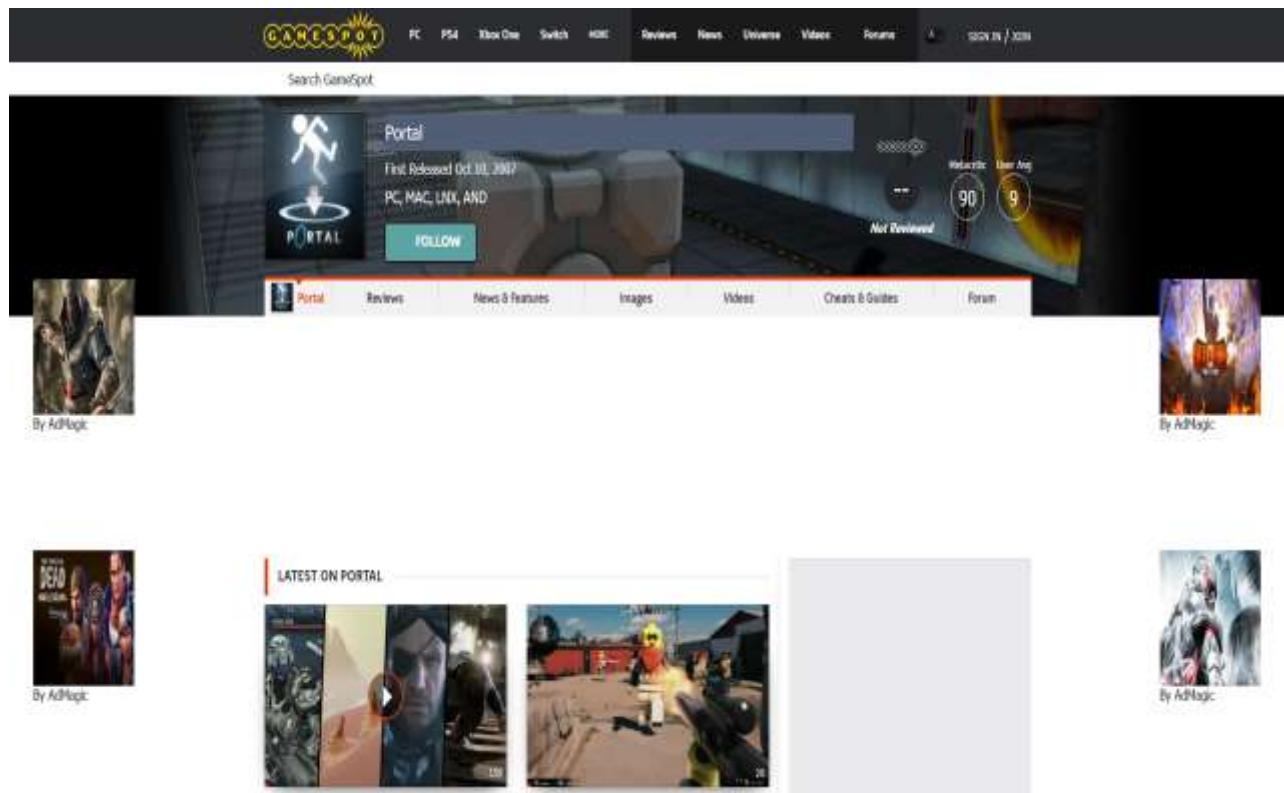


Figure 6.11 Webpage with Ads

The webpage which has been already registered with us using its URL now displays ads related to its content. Here a games website contains ads related to gaming only. They are located at the sides of the webpage. The no of ads to be displayed are determined by the webpage at the time of registration. The ads change once in 5 secs due to the javascript code at the back. The ad's URL is not exposed in the source code of the webpage. Instead it's embedded in the response of JSP page. Hence the security feature is high in the system.

CHAPTER 7

CONCLUSION

- We were able to build a system that merges features of both AdSense (contextual ads) and AdChoice (location and preference).
- It can be used by both Advertiser who wants to advertise and promote their recent product and also a Publisher who wants to turn their webpage's traffic into good revenue.
- Our main motive was to keep the targeting simple and not money based greedy targeting.
- This was fulfilled by the system and the ads were displayed perfectly.
- In the future targeted advertising will be an important part in marketing and hence new powerful systems like these are need to tackle large data.
- The advertising should be done by the system itself and not by the advertiser.
- Hence this system provides a way to make an impact in advertising.

CHAPTER 8

FUTURE WORK

- In the advancement of this project, our future work lies in the field of obtaining the preference of the user by extracting his previous browsing history and collecting data from cookies.
- The data from cookies needs to be mined and finally the filtered ads could be displayed to the viewers.
- This part involves some legal issues due to the extraction of private cookies of users.
- Privacy of the user should not be misused at the same time only useful information must be analyzed from cookies.
- Prediction criteria of advertisements will include past behavior also now.
- The keyword extraction may not be very accurate if the webpage owner doesn't choose appropriate keywords for his page. Hence text mining can be done to determine the profile of the content.

CHAPTER 9

REFERENCES

- [1] Pinaki Mitra, Kamal Baid “Targeted Advertising for Online Social Networks”, IEEE 2014.
- [2] Vladimir Salin, Maria Slastihina, Ivan Ermilov, Ren´e Speck, Soren Auer, and Alexander Sytnik “Keyword Extraction for Webpage Clusters”
- [3] Xiuqiang He, Wenyuan Dai, Guoxiang Cao, Ruiming Tang, Mingxuan Yuan and Qiang Yang “Mining Target Users for Online Marketing based on App Store Data”, IEEE 2015.
- [4] Hong-we1 Yang, Zhi-geng Pan, Xi-zhao Wang, Bing Xu “A Personalized Products Selection Assistance Based On E-commerce Machine Learning”,IEEE 2005.
- [5] D. M. Boyd and N. B. Ellison, “Social network sites: Definition, history, and scholarship,” *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, November 2009.
- [6] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, “How much can behavioral targeting help online advertising?” in *WWW*, 2009, pp. 261–270.
- [7] K. Devika and S. Surendran. “An overview of web data extraction techniques”, *International Journal of Scientific Engineering and Technology*, 2013.
- [8] N. Erbs, P. B. Santos, I. Gurevych, and T. Zesch. Dkpro “Keyphrases: Flexible and Reusable Keyphrase extraction experiments”, *ACL* 2014, page 31, 2014.