Automation and Personalization of E-Commerce Field

Suharsha Wickramasinghe[1*], Hasitha Asurasinghe[2], Dilini Subhani[3], Thanuka Gunathilaka[4], Rohan Samarasinghe[5]
Department of Information Technology, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka.
suharshapw@gmail.com[1], hasitha1991@gmail.com[2], dilisubhani@gmail.com[3], vthanuka@gmail.com[4], rohan.s@sliit.lk[5]

Abstract – This research article will be focusing on how e-commerce websites can be enhanced by providing automated advertisements extraction, advertisements categorization and target right customers for the products and services by implementing efficient and accurate personalization. Furthermore the article will be discussing the ways of using algorithms such as Clustering algorithms and Naive Bayes algorithms in its best practice as a tool to achieve this task. The process flow used in the design shows clearly the methodology used in advertisement extraction, categorization, and personalization and publishing them on user's computer screen. The information such as user's navigation data and social media data are brought together and analyze them to make e-commerce process more convenient for users and administrators alike.

Keywords – K-Means Clustering, Naïve Bayes Theorem, Data Extraction, Classification, Jenkins, E-Commerce, Personalization, User Profiling, Social Media Data Analysis, Automation, Natural Language Processing

I. INTRODUCTION

Commerce is integral to human life for thousands of years and today, it is an important yet time consuming activity done by everyone on a daily basis. For example, buying daily needs such as groceries, clothing, house hold items etc. are part of commerce activity carried out by human beings every day. Today, with the development of technology and wide use of computers and portable electronic devices has created new ways of engaging in commerce activities. Some may describe this as intervention of technology in traditional commerce process of shopping activities. For a number of years, E-Commerce has played a major role in improving shopping experience of customers by introducing new technology and processes. In simple terms, E-Commerce could be described as websites where sellers and buyers interact to sell or purchase goods and services without physically visiting a shop or shopping mall. Sellers can advertise their items on internet sites and the interested buyers can visit them and purchase them online. In most cases in the modern world, these websites deliver goods to customer's doorstep from various parts of the world using air freight, sea freight services, courier services and normal postal services. This method of E-Commerce has made consumers life more comfortable and easier by providing online shopping with the convenience of your own home without visiting a shop.

But the main question still remains unanswered whether the goods and services advertised on these websites are being displayed to the right potential consumer? Can a regular internet user log on and see what he/she is interested in, rather than getting bombarded with numerous advertisements that have no relevance to customer's requirements? Today's busy life style of web users, demand convenience and less time consumption when purchasing goods and services through internet. It appears that design features of most websites have overlooked this basic customer requirement. Any user who logs on to a website gets bombarded with too many advertisements then that user generally, get frustrated and end up finding it difficult to navigate through to the advertisements that the customer was looking for. Hence, these types of advertisements will lose their marketing value and turn away many potential viewers.

The solution to this problem is to design a system that uses data extraction technologies to extract advertisements from a given set of websites at given time. This process involves extraction of advertisements and categorization in an order using categorization engine. After that identify the advertisement category that is best suited for the

customer's particular need and the original advertisement that the customer was looking for.

The system also implements personalization service by analyzing user data. These data may include information from Facebook, user's navigation data such as click data etc. All these information is processed combining new technologies to build the system which has an automated backend that implements user personalization.

II. LITERATURE REVIEW

The World Wide Web is not what it used to be few years back. With the exponential advancement of technology, thus making it available to majority of the population via computers, laptops, mobile phones and many new hand held devices, the internet is accessible to any individual despite of their actual knowledge in the field. It has been calculated that 40% of the world population has access to the internet now.

Therefore it is very clear that the population is moving towards the better convenience of life through easy access to their daily activities at their most convenient places. With the emergence of social media websites such as Facebook.com the usage of internet has taken a different path towards personalized internet where the websites users access, they prefer them to be personalized for their needs.

Personalization has been discussed in the context of numerous areas such as recommender systems, adaptive hypermedia, information filtering and user modeling. The basic definition of personalization is providing an individual with what they want and need without explicitly asking from them. Personalization provides the ability to provide content and services tailored to individuals on the basis of knowledge about their preferences and behavior.

With the aforementioned increment in the amount of users connected to the internet, advertisers focus their advertising efforts more on the online aspect of it. Advertising is still the single most important revenue for many companies on the web. There is seemingly no end in sight, according to marketing research companies, revenue figures from online ad sales continue to grow at a rate of 200% a year with a revenue of over 2 billion dollars in 1998 alone. Advertising without any targeting mechanisms has advantages such as no privacy concerns and it is easy to set up. However the effectiveness takes a very low value with a low click through rate since most of the ads will be unrelated to the viewer's interest. The traditionally rather homogeneous group of single, young males dominating the early Web has long since given way to a very diversified user base, including 8 almost 25% of users that are 50 years and older, as well as nearly 40% women which makes it difficult to display a single advertisement that appeals to all its viewers.

Personalized advertising allows advertisers to control who see their advertisements. However systems like DoubleClick's DART use unique identifiers to group the collected data by user. In either case advertisers have to constantly monitor and manually revise their targeting parameters in order to achieve maximum effectiveness.

Personalization and automation act as major anchor points when it comes to E-Commerce. Because automation ensures accurate and up to date content while personalization ensures targeting a product or service to the right customer. As one category in e-commerce, the advertising of discounted products has been emerging in the internet for a long time. Just like every other aspect of it, the trend is customers search online for their favorite product discounts and come across numerous web platforms which provide the information they are looking for. As one issue of this, the customers point out that when they search for their interest, they also get bombarded with many other products which don't interest them at all. Therefore there is a high chance that the marketers lose their potential customers who were just a simple click away from claiming their product online.

In the context of Sri Lankan e-commerce, there are many identified discounted product advertising platforms such as 'Wow.lk', 'Discount365.com', 'Mydeal.lk'. But the lack of personalization has kept them from achieving their expected results.

III. RESEARCH METHODOLOGY

The design and development of the system was started after a thorough analysis of the available systems and their personalization processes.

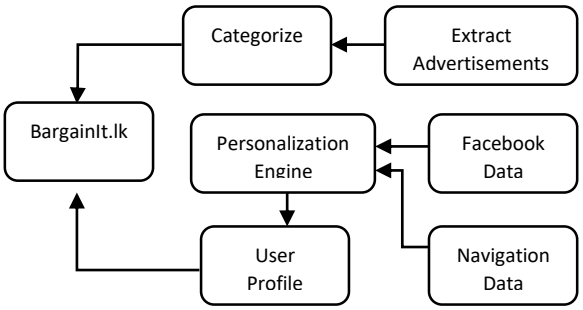

Fig. 1. System Overview

The first and most important step of automation is the extraction of advertisements. This component is the most significant part in this project as all other components depend on the discounted product information that was extracted as part of the automation process. Initially, it is necessary to define a set of online advertising sites and these data sources are chosen manually. Generally, at this stage consideration is given for extraction of text for this process. After defining the set of online advertising websites, it should further analyze the data to identify the products that are offered discounts on sale price and other products that do not have any discounts. Data mining algorithms are used to analyze the specific data in order to sort data to identify only the products with discounts. Once this step has been completed, it is possible to provide those discounted product advertisements to auto categorization purpose.

- Web Data Sources

The advertisement sources will be manually fed to the system such as advertising sites, shopping sites etc. For example, these sites may include wow.lk, mydeal.lk etc. After that these websites will be analyzed to identify the format of each website and the way in which the data is organized in order to retrieve the necessary details.

- Advertisements Selector

Once the advertisement sources have been defined, the next step is to select only the advertisements which are having discounted offers. This is done by analyzing format of each website by using "discount" or "offers" as keywords.

- Advertisements Extractor

After selecting the advertisements with discounted prices, the next step is to extract the details of those advertisements. Hibernate technology that is an Object-Relational Mapping (ORM) solution for JAVA is used as an open source persistent framework to maintain the data extracted into a table. It is a powerful, high performance Object-Relational Persistence and Query service for any Java application that extract information such as the title of the advertisement, ad image, discounted rate, market price and selling price etc.

- Scheduler

The scheduling phase is the final step of the advertisement extraction process. As most of the web sites change their content daily there should be a mechanism to extract the latest details of the advertisement. Therefore, it is necessary to develop a process to access the data sources according to a given predefined time and retrieve the latest details automatically. This is achieved by implementing the process known as Jenkins server. Jenkin Server is an open-source continuous integration software tool that is written in Java.

- Advertisements Categorization

Once the advertisement extraction process is completed it then leads to publishing phase of those extracted advertisements. In order to provide a better user experience, it is important to identify the category of each advertisement before publishing them. The system uses Naïve Bayes classification to achieve this task.

The Naïve Bayes (1) classification used in the system is based on Python language. The predefined sets of categories must be trained using a training data set. For each category that the research team has identified, there has to be a minimum of five hundred train data keywords. As the number of train data keywords increases, with it, the accuracy of the model will increase. Every time new data was added to training data set, as a confirmation of the training accuracy, the model run through a testing data set to confirm the accuracy. After the extraction of advertisements, these advertisements will go through the categorization engine which will analyze each advertisement title using Natural Language. When the processing is completed, it will suggest three most suitable categories for them with their matching percentages. After that the process will assign the most matching category to the given advertisement. This process will repeat as a loop through all the advertisements.

- Naïve Bayes Algorithm

Naïve Bayes' (1) is a learning algorithm that is frequently employed to iron out text classification problems. It is computationally very efficient and easy to implement. For this experiment, project team has used large data sets, which are frequently available in the text classification literature.

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable "y" and a dependent feature vector "x_1" through "x_n", Bayes' theorem states the following relationship.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)} \quad (1)$$

- Personalization

The primary objective and the key component of this research were to achieve Personalization of E-

Commerce. The registered users were analyzed using two types of data to come up with a user model. They were internal data such as search attempts, navigation pattern etc. which were obtained by analyzing user's behavior when navigating through the website. The other type of data is external data which were obtained by user's information outside the system. For example, external data may be obtained from media profile data from sites such as Facebook page. These data were matched with the relevant advertisements using targeting model to make sure only the relevant advertisements are getting displayed and at the same time studying and collecting information about user behavior patterns to further enhance the accuracy of the information contained on advertisements.

- Facebook Data Analysis

The initial step was to identify the most usable content from service providers from Facebook, Twitter etc. to suit the Sri Lankan context. After conducting surveys and asking questions personally, we concluded that Facebook is more popular among local internet users than Twitter and other similar social media websites. There are some constraints with the usage of Facebook data by outside parties. We were able to overcome this by identifying the required attributes and requesting permission from the users prior to obtaining those data.

- Login

During the sign up process, the user will be given the choice of signing up using either email or Facebook account. When signing up using Facebook, the user's permission is sought to collect user information. The system then collect these data from Facebook using the Graph API provided by the developers of Facebook and store them in a data store for processing.

- Text Processing

The system performs preliminary text mining techniques on the obtained data removing stop words, stemming and extracting the most frequent words and phrases. After that the processed text will go through a NLP tool in order to identify their category. This step is repeated for every page data that the user has liking to, in order to gain an insight into the user's personal type.

Natural Language Processing analyze each title and description of the pages that the user preferred accessing, to determine the matching category of that particular page. After analyzing each and every page in this manner, the system comes up with a list of categories of pages that the user liked accessing to determine user's preference. After that using an algorithm, the most frequent categories will be selected. In this way, the system is able to provide output similar to user's preferred categories on Facebook and other similar webpages.

This is a repetitive process where data is analyzed every time the user logs in using Facebook account. Every time a user logs in, user's profile gets updated with the latest results obtained from analyzing Facebook data.

- Navigation Data Analysis

The second method of analyzing user's preferences is using navigation data throughout the system. After researching user navigation data from other websites, we were able to conclude that navigation data could reflect great insight into user's preferences.

During the development process, it was identified that there was one deficiency with the suggested system. That was accessibility for a shopping cart to use the bargains and discounts. Since the system only displays advertisements about bargains and discounts, it did not provide the opportunity of claiming them through itself. In order to address this, it was decided to use other types of user navigation data.

After a thorough research in this field, we concluded that the best option is to use user's clicks on advertisements to come up with a user profile. Along with that we have identified the most important attributes which could affect user's decisions such as click time of the day, day of the week, price and discount etc. This information was stored along with the click data. In this way, user's personal information, clicked time, clicked day of the week, clicked advertisement's category, price on advertisements and the discount amount etc. were all collected. These wide ranging data were able to give a better understanding about user's preference.

After researching many methodologies available in the current e-commerce context for determining implementation process, it was realized that the amount of such data collected every day, does not provide efficient method by calculating the distance of each row for implementation. Therefore, instead of calculating method, it was decided to implement clustering mechanisms to cluster similar types of users into groups and identify common advertisements preferred by those groups.

Clustering is the process of putting together classes, or conceptually meaningful groups of objects that share common characteristics to identify patterns.

The Fig. 2 shows an example of clustering where a group of people are clustered depending on their

income, education and age. Clustering is the process of grouping similar types of people together.

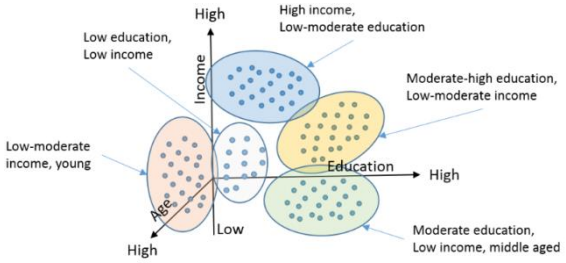

Fig. 2. Clustering Example

In the proposed system, clustering was used to cluster similar types of users with the help of their personal data and clicks data. In the beginning we were able to identify the most important attributes which would give the clustering algorithm the distinct idea about the user. Information such as age, profession and clicks are the main data that determined clustering.

Next step was to choose a suitable clustering algorithm based on the performance and expected accuracy. This could be either K-Mean or Expectation-Maximization. These algorithms provide all the required functionalities and have different representation of end result and accept different kinds of attributes. As an example, K-Means requires numbering of clusters to be predefined whereas Expectation-Maximization (EM) determines the number of clusters by itself, using an iterative process. Finally, the selected algorithms were customized to suit selected attributes and the required final output.

- K-Means Algorithm

The Fig. 3 shows the K-Means algorithm and the meaning of the symbols. In the K-Means algorithm, at first choose K initial centroids. Where K is a user-specified parameter and it is the number of desired clusters. Each point is then assigned to the closest centroid, and each collection of points assigned in a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster and by repeating the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same.

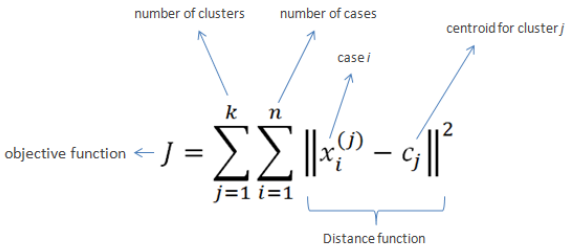

Fig. 3. K-Means Algorithm

Once, the decision of choosing programming language was made, focus was on learning about already available frameworks which provided pre-built algorithms. After analyzing the existing solutions, K-Means and EM algorithms, it was decided to use K-Means as the algorithm of clustering.

IV. RESULTS

The outcome at the end of the life cycle of backend would be a list of advertisements extracted from a given set of websites which are processed and categorized accordingly.

The personalization engine will analyze user data and keep the user profiles updated every time user interacts with the system. The updated user profile information then suggests advertisements that are compatible with the user interests.

V. RESEARCH FINDINGS

The research team was able to come up with a fully automated backend for any e-commerce solution that has the capability of automated categorization as well. This has made the operation easier and less time consuming at the same time delivering latest discounts to the users.

This paper also proposes a solution to an already widely known problem that the sellers are not receiving expected amount of sales. During the research, it was identified that several major e-commerce websites which are used by majority of local internet users as well as few popular local websites such as wow.lk, mydeal.lk experienced lower sale targets than expected. Even though there were frequent visitors to these websites, it appears that often, they were not able to find what they were looking for quickly and accurately.

The research revealed that navigation patterns tend to show user's preferences and personality types. With the help of personal details such as age, the users were clustered to form a model to target advertisements that were within user's interests. As a result, the team was able to find interesting

relationships about regular shoppers. One such relationship is how age affects the buying patterns.

VI. CONCLUSION AND FUTURE WORK

This research provided methodology for building automated e-commerce websites with efficient recommended systems for those websites that do not have a shopping cart or any explicit method to collect information about users. This information is valuable for the purpose of evaluation of their personality, using data mining and machine learning techniques and algorithms. The uniqueness of this system makes it stand out from other similar systems. It provides easy administration and convenient user navigation experience throughout the website.

This solution has many other applications for systems of any field to come up with solutions for users depending on their navigation patterns and Facebook data.

- Further Improvements

a. Developing a mobile application which stays in synch with the web platform.
b. Provide location based discounts and bargains.
c. Integrate more social media platforms for fine tuning personalization.
d. Personalize further to identify spending capabilities.
e. Integrate banking card transaction tracking to further personalize.
f. Dynamically adjust to new ads extraction websites.
g. Store advertisement images locally.
h. Offer admin dashboards to chosen sellers to update their content.

ACKNOWLEDGMENT

The above study has been taken under the Sri Lanka Institute of Information Technology, Sri Lanka. With deepest gratitude and thanks to our supervisor Dr. Rohan Samarasinghe and our external supervisor Mr. Hansa Perera for the enormous supervision and support given throughout the entire research project. Lecturer-In-Charge of the course Comprehensive Design and Analysis of Projects, Mr. Jayantha Amararachchi for the assistance and guidance offered to us in order to prepare for this task as a winner. Further extending gratitude to all academic staff, non-academic staff, colleagues and friends for their generous support. Finally deepest gratitude and thanks to parents and family members for the support and the encouragement given to us throughout the entire project.

REFERENCES

[1] Sviatoslav Braynov , “Personalization and Customization Technologies”, Department of Computer Science and Engineering, State University of New York at Buffalo, NY 14260.

[2] Atefeh Danesh Moghdam, Ali Jandaghi, Seyed Omid Safavi, “The Probability of Predicting E-Customer’s Buying Pattern Based on Personality Type”, Iran.

[3] Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, Kiran Sagoo, “Real-World Behavior Analysis through a Social Media Lens”, Computer Science and Engineering, Arizona State University.

[4] Toby Segaran, “Programming Collective Intelligence”.

[5] “Cluster Analysis: Basic Concepts and Algorithms”.