International Journal of Information Science and Management
Vol. 20, No. 3, 2022, 141-163
https://dorl.net/dor/20.1001.1.20088302.2022.20.3.9.6

Original Research

Predicting Customers' Behavior Using Web Content Mining and Web Usage Mining

Bahareh Sheykh Abbasi
M. A., Information Technology Engineering, Faculty of Electrical, Computer and IT Engineering, Qazvin Islamic Azad University, Qazvin, Iran.
b.abbasi87@yahoo.com
ORCID ID: https://orcid.org/0000-0003-4419-7397

Neda Abdolvand
Associate Prof., Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran.
Corresponding Author: n.abdolvand@alzahra.ac.ir
ORCID ID: https://orcid.org/0000-0003-3623-1284

Saeedeh Rajaee Harandi
M. A., Information Technology Management, Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran.
sa.rajaeeharandi@gmail.com
ORCID ID: https://orcid.org/0000-0001-9710-3644

Received: 10 May 2021
Accepted: 28 November 2021

Abstract

Today, e-commerce has become a competitive space for online retailers. Therefore, personalization has become a vital part of e-commerce websites' success, challenging marketers and researchers. This study aims to provide a model for web personalization and mining user interests using a hybrid web usage and web content mining approach. The navigational patterns of web users and the interests of each user on web pages of a Persian website were extracted through web usage mining and topic modeling. Users were then clustered using the dependency distribution algorithm, and 25 categories were extracted. To better understand the behavioral patterns of web users they were categorized using the Support Vector Machine algorithm based on the users' interests and navigational behaviors. The most important result of the proposed system is that the patterns of users' navigation are understandable, and the subsequent analyses will be much simpler.

Keywords: E-commerce, Web Personalization, Web Mining, Web Content Mining, Web Usage Mining.

Introduction

Today, with increased competition between internet-based businesses, a website's success lies in its ability to turn random visitors into potential customers; thus web personalization has become an essential part of e-commerce helping businesses retain their existing customers and acquire new ones (Dai, 2005). Web personalization systems provide recommendations based on users' interests without their explicit request (Jalali, Mustapha, Sulaiman & Mamat, 2010). Such systems can learn from the customers' behavior and recommend personalized products to customers (Nkwo, Orji, Nwokeji & Ndulue, 2018). In addition, personalized systems can lead to increased digital sales and conversion rates, accurate product recommendations, and improved customer loyalty (Salonen & Karjaluoto, 2016).

Besides, retailers can utilize web personalization to gather customers' data, provide personalized recommendations, and do purposeful promotions (Pappas, Kourouthanassis, Giannakos & Lekakos, 2017).

Conventional personalization methods include two user-based techniques and item-based or content-based techniques. The item-based technique focuses on recommending the most relevant items to individual users. User-based techniques focus on the similarity of users and provide recommendations based on the most relevant users (Eirinaki & Vazirgiannis, 2003). Integrating these techniques will increase the accuracy and quality of users' behavior predictions. In addition, using content mining will make it easier to understand the patterns obtained from usage mining and provide website owners with deeper insights into user navigation (El Aissaoui, El Madani, Oughdir & El Allioui, 2019; Senkul & Salin, 2012).

In some studies conducted in this field, to achieve better results, semantic information of web page content has been integrated into the results of web page mining. Of course, text-analysis-based methods also have their challenges, and researchers are trying to address them (Modi & Jagtap, 2018; Senkul & Salin, 2012). Despite the advantages of integrating web content and usage mining in personalization, search engines still have low accuracy and efficiency in responding to users. Therefore, this question arises: how can we integrate the navigational information of users with the knowledge of users' interests obtained from the content of web pages to achieve a more effective personalization of a Persian website? This study aims to provide an innovative model for a retail website to obtain more accurate information about the users' navigational behavior and their interests in visiting the websites. Integrating content and usage mining will lead to identifying different groups of customers to predict the future behavior of new customers and their interests. To extract users' interests, topic modeling of pages and web usage mining techniques has been used, which has not been observed in a Persian language website. Since the extracted labels in this study are the subject of web pages and are a multi-class classification issue, the support vector machine (SVM) algorithm is used for classification. This is one of the supervised learning methods. It has better performance in data classification with non-binary tags (Habib, 2021; Shotorbani, Ameri, Kulvatunyou & Ivezic, 2016; Wu et al., 2008) and has not been used in studies in this field so far.

The rest of the research is organized as follows. The research method and the proposed model are discussed after reviewing the literature, followed by the conclusion and recommendation for future studies.

Literature Review

Web personalization

Personalization, often referred to as "one-to-one marketing," is a marketing strategy in which businesses use the science of data analysis and digital technology to predict users' needs to improve efficiency and retain loyal users and website customers (Salonen & Karjaluoto, 2016). Web personalization is the most common method used by marketers. It employs knowledge obtained from analyzing users' navigational behavior and personal interests, along with the structure and content of the website (Kalaignanam, Kushwaha & Rajavi, 2019; Sharma & Rana, 2020). There are numerous advantages to personalized marketing for both marketers and customers. For example, personalized products and services attract customers' attention and increase loyalty (Sunikka & Bragge, 2012).

Conventional personalization methods, including collaborative filtering, rule-based filtering, and content-based filtering systems, often have problems such as lack of scalability, reliance on user ratings, and inability to identify richer sets of meaningful relationships between objects (in content-based systems). Therefore, recent studies advocate the use of the web mining approach in web personalization (Modi & Jagtap, 2018; Taherizadeh & Moghadam, 2009; Carmona Ramírez-Gallego, Torres, Bernal, Del Jesús & García, 2012).

Web Mining

Using data mining techniques to discover the knowledge of web documents and services is known as web mining (Modi & Jagtap, 2018; Taherizadeh & Moghadam, 2009; Carmona et al., 2012). This process is crucial for understanding e-commerce users, turning information into competitive advantages, and enabling organizations to make data-driven decisions and improve their decision-making strategies. Web mining helps organizations to gain new customers, maintain existing ones, and increase customer satisfaction (Shanthi, 2017). The main applications of web mining in e-commerce include personalization, system improvement, site improvement, business intelligence, and application descriptions (Das & Turkoglu, 2009). Web mining methods based on the usage of data collected from the website include three categories (Modi & Jagtap, 2018; Taherizadeh & Moghadam, 2009; Carmona et al., 2012; Kumar & Thakur, 2018): web structure mining, which is used to identify the relationship between user hyperlinks and web pages (Prathyusha, Kumari, Sumathi & Rangaswamy, 2019; El Aissaoui et al., 2019). Web content mining, which involves identifying web resources, categorizing documents, and extracting and clustering information on the web pages (Van Aartsen, El-Gayar & Noteboom, 2020), and web usage mining, which focuses on extracting users' navigational behaviors by analyzing web server logs (Yadav, Feeroz & Yadav, 2012; Bin & Zhijing, 2003).

Web Content Mining

Web content mining analyses user profiles to suggest items or pages (Dai, 2005; Arinaki & Wazirjanis, 2003). This method attempts to discover beneficial information from the web page's content and prepare structured reports from this data (Grace, Maheswari & Nagamalai, 2011). Web page content usually includes text, images, videos, audio, or built-in records such as lists and tables. In this context, text exploration has been studied more than any other related field (Salman, Zaki & Shiltag, 2020; Liu & Chen-Chuan-Chang, 2004). Application of web content mining includes detecting the topic of a web document, classifying web documents, finding the same web pages on different servers, facilitating web user queries, and filtering information for users (Grace et al., 2011). Usage mining aims to predict the behavior of users (Yadav et al., 2012).

Web Usage Mining

Web usage mining automatically discovers users' navigational behavior from the weblogs (Kumar & Kumar, 2021; Vijaiprabhu & Meenakshisundaram, 2020). It aims to predict the users' behavior through their online interactions (Vijaiprabhu & Meenakshisundaram, 2020; Salman et al., 2020; Velásquez, 2013).

Web usage mining seeks to determine what the user is searching for on the Internet. Through web usage mining, helpful information such as user navigation patterns is extracted and analyzed using web log data. Both web designers and users benefit from this process. On the one hand, web designers can determine the navigational behavior of web users by analyzing the navigational patterns in the web server log, thereby determining the site's most popular

pages and which are most likely to be visited together. On the other hand, web users can use this information to access the web more effectively (Sun & Zhang, 2004). organizations can use the discovered information, to calculate the value of customers' lifetime, design appropriate strategies for marketing, evaluate the effectiveness of advertising campaigns, and deliver personalized content to visitors (Carmona et al., 2012; El Aissaoui et al., 2019; Rathi & Singh, 2019; Prathyusha et al., 2019; Salman et al., 2020).

Integrating Web Usage Mining into Web Content Mining

Although each individual's navigation is unique, patterns emerge in distinct group behaviors when many people access a particular resource. If behavioral groups can be easily identified, it is likely to suggest web pages more carefully (Senkul & Salin, 2012). Integrating the results of web usage mining into web content mining will improve the accuracy and quality of predictions (Yilmaz & Senkul, 2010). Profiles of users' interests, obtained from integrating content information and usage, introduce links tailored to users' interests. Enriching usage information with content makes it easier to provide better recommendations (Taherizadeh & Moghadam, 2009; Mobasher, Dai, Luo, Sun & Zhu, 2000; Sinkul & Salin, 2012). In addition, web content mining facilitates understanding patterns derived from usage mining and provides website owners with deeper insights into user navigation (El Aissaoui et al., 2019; Dai, 2005).

Researchers have tried to increase the accuracy and quality of web page predictions and achieve more personalization that is effective by integrating the results of web usage mining into web content mining. For example, Bharti and Raval (2019) combined content and usage mining approaches to develop strategies to improve web page access predictions. In this approach, new sessions are replaced with old sessions. The study by Herwanto (2016) used a combination of web usage and content mining to obtain a navigational pattern of overall visitors. Using Latent Dirichlet Allocation (LDA) for content mining and Self-Organizing Map, they indicated that their proposed system can make a navigational profile that matches the overall navigational profile for new users. In another study, Khatri and Gupta (2015) improved web personalization by combining web use and content mining. This model considers the user's web access activities to extract his/her navigational behavior to create a knowledge base. They then used the knowledge base with the extracted features to predict user-specific content. This method provides an effective periodic recommendation and helps the user to access more information that is efficient. Arbelaitz, Gurrutxaga, Lojo, Muguerza, Pérez and Perona (2013) also used a combination of web usage and content mining in tourism. This study proposes a system that uses usage information and content without prior knowledge of the web content structure to predict links and provide information on users' navigation. Taherizadeh and Moghaddam (2009) proposed a system for finding effective association rules by combining web content analysis with Web usage mining. The resulting association rules were related to web content and provided information for creating user profiles, improving, and evaluating web pages.

According to studies conducted in this field, previous research has tried to add content features to the results of usage mining to increase the accuracy and quality of patterns and facilitate their understanding. However, researchers have generally combined the results of usage mining with the web anthology, the main limitation of which is the unavailability of information about the structure and web anthology. Despite the researchers' efforts to increase the accuracy and quality of the proposed systems and achieve more effective personalization, search engines suffer from a lack of accuracy and efficiency in responding to users. This paper aims to provide a practical model of users and their interactions with the website that directly

impact the accuracy of the recommendations. Previous studies on integrating web usage mining into web content mining are summarized in Table 1.

Table 1

Previous Studies on Integrating Web Usage Mining Into Web Content Mining

Resource	Objective	Data	Method	Results
(Bharti & Raval, 2019)	To combine content and usage mining approaches to develop strategies to improve Web page access predictions.	The web access log data is collected using Google Cloud's Log Viewer	Modified Jaccard Set Similarity method; Alignment algorithm; CLUTO's k-way clustering algorithm	The proposed system had better results than previous ones and can provide more effective periodic recommendations.
(Herwanto, 2016)	To combine content and usage mining approaches to obtain a navigational pattern of overall visitors	World Wide Web	Latent Dirichlet Allocation (LDA); Self-Organizing Map	The evaluation indicated that the proposed system can make a navigational profile that matches the overall navigational profile for new users
(Arbelaitz et al., 2013)	Modeling of tourist website users using a combination approach of web usage and content mining in tourism.	Bidasoa Turismo (BTw) website, www.bidasoaturismo.com, is based on usage data collected over 10 months and the corresponding content data.	PAM algorithm; SPADE algorithm; K-means algorithm	The system was successful and obtained profiles that in more than 60 cases matched the actual user navigation sequence and in more than 90 cases matched the user's interests.
(Guerbas et al., 2013)	Using a combination approach of web usage and content mining in predict the survey patterns of users	Two different real datasets including e NASA space center web server log and the dataset provided to them by other researchers (ie., Lei and Ghorbani, 2004)	DBSCAN algorithm, OPTICS algorithm, and scalable KNN algorithm	The performed experiments show the applicability and effectiveness of the proposed approach.
(Senkul & Salin, 2012)	Using a combination approach of web usage and content mining in predict the survey patterns of users	Log data of live Web sites	SPADE algorithm and pattern evaluation framework with window counting parameter	According to the results, more recommendations that are accurate can be obtained by including semantic information in creating a navigation pattern, which indicates an increase in the quality of the patterns.
(Carmona et al., 2012)	Using a combination approach of web usage and content mining to analyze the E-Commerce Website	User history databases associated with an e-commerce website	K-means algorithm, Apriori algorithm, and NMEEF-SD algorithm	The results provide some guidelines for improving the usability of the website and user satisfaction. According to the results, the webmaster team needs to seriously improve the keywords related to Iberian products because there is no information about user access through these keywords.
(Taherizadeh & Moghaddam,	To propose a system for finding effective	An IIS server log access file from the	Apriori algorithm, K-means algorithm,	The resulting association rules were related to Web

Resource	Objective	Data	Method	Results
2009)	association rules by combining Web content analysis with Web usage mining.	Web server of a software provider company	and CWF text viewer	content and provided information for creating user profiles, improving, and evaluating Web pages.

Materials and Methods

This study aims to personalize a retail website by integrating web usage and content mining. This study is based on CRISP-DM methodology, one of the most used processes for data mining projects (Beheshtian-Ardakani, Fathian & Gholamian, 2018). The research framework is shown in Figure 1.

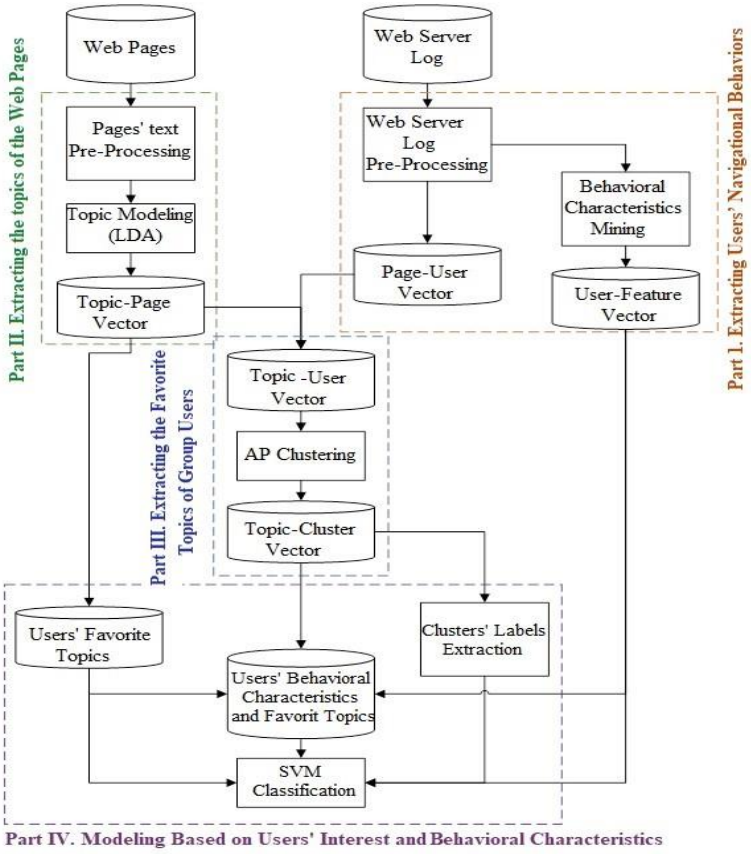

Figure 1: Research Framework

This research consists of four parts: part I. involves discovering the navigational patterns of users through web usage mining. In part II, the content of web pages was extracted using the Latent Dirichlet Allocation (LDA) algorithm (Xing, Lee & Shibani, 2020). LDA is a topic modeling algorithm, which is a powerful method that allows identifying topics within the documents and mapping documents to those topics (Abdi Ghavidel et al., 2015; Du, Yi, Li, Chen, Fan & Su, 2020; Yang & Zhang, 2018; Herwanto, 2016; Shotorbani et al., 2016; Akanfe ,Valecha & Rao, 2020; Blei, 2012). Most text feature extraction methods are based on a set of statistical features, which lack optimal performance for semantic clustering. Thus, in recent years, the use of semantic methods especially the Latent Dirichlet Allocation (LDA) algorithm for clustering has become more common. Topic modeling with LDA is a computational

content-analysis technique that can be used to examine the hidden thematic structure of a set of texts (Maier et al., 2018; Blei, 2012). Using the LDA algorithm, certain topics can be discovered that are not easy to identify with existing clustering methods. It has been proved a powerful approach for quickly identifying key thematic clusters in large text bodies and modeling topics as latent structures embedded in a text set. Compared to simple co-occurrence analysis, a topic model can reveal a hidden semantic connection between words, even if they have never occurred together in the document. Besides, the advantage of LDA compared to other topic clustering methods is its mixed membership approach, meaning that a document can contain several topics, which is a useful assumption (Maier et al., 2018). In part III, by combining users' navigational behavior patterns in scrolling web pages and topics related to each page, users' favorite topics and the extent of this interest were determined. Then, users were classified based on their favorite topics using the dependency distribution algorithm, and the clusters were evaluated using Davis-Boldin and ANOVA tests. In part IV, the support vector machine (SVM) algorithm, one of the most effective and accurate classification methods (Habib, 2021; Shotorbani et al., 2016; Wu et al., 2008) was used to classify users based on their interests and navigational behaviors. The analysis was performed using Python and R programming languages.

Part 1. Extracting Users' Navigational Behaviors (Usage Mining)

In this part, the important features of users who visited the website were extracted based on the pages each user had seen and the frequency of visiting those pages. The system proposed in the study was built for the XYZ online shopping site, based on real usage data collected from July to October 2019 and the corresponding content data. For this purpose, 553,957 original log files recorded by web servers were collected. The XYZ website offers a wide range of products, including electrical appliances, digital accessories, home appliances, health and care products, tools, automotive equipment, and sports products. At the time of this study, the website had 374 pages, and the contents of each page were extracted and reviewed separately. The fields were then selected for analysis. Selected fields include the logical address (IP address), submitted request, date and time, request status, the page from which it was referred, and the software agent.

These records contained raw data that needed to be cleared and preprocessed before further analysis. To this end, records containing unsuccessful requests and requests that included images, videos, and scripts were removed. Since user clicks lead to many indirect requests being sent to the servers recorded in the log files, such records were also removed. The network robots are typically used in search engines to crawl pages for information (Sujatha, 2012). These robots create many records in log files, which divert the pattern discovery process from the main path; these records were also removed. Besides, requests related to the front page of the website were removed. After clearing data, the number of records dropped to 24,737.

In the next step, users and their related sessions were identified. The user identification was made based on an experimental method using the IP address, type, and model of the operating system or browser configuration. In other words, users were separated based on the type and model of browser or operating system at the same time as the first request was sent to the webserver and received a unique user ID in the database tables. Through this method, 2,921 unique users were identified.

Then, sessions related to each user were identified. Because sessions encode users'

navigational behavior, they are important for discovering patterns (Pierrakos, Paliouras, Papatheodorou & Spyropoulos, 2003). Each user session includes a set of pages, that a user has viewed during his/her visit to a particular site (Chitraa & Thanamani, 2011).

A time-based exploration method was used in which a threshold of 30 minutes was used to detect the end of a session to identify the session. According to this method, if 30 minutes have elapsed since the last request sent by the user to the webserver, the user's session ID will be invalidated, and he will receive a new session ID (Fong, Zhou, Hui, Hong & Do, 2011; Jalali et al., 2010). After identifying each user's access path in the next step, five common parameters were determined based on user behavior in their visits to the website, including the duration of visiting each page, the total number of days the user visits the website, the number of daily sessions of the user in visiting the website, the average number of pages that each user viewed in a session, and the last time (day) that user visited the website. By calculating these parameters for each user, a user-feature vector was created. Each vector is related to a user and his navigational characteristics during website visits. Given the different scales of the parameters, the Z-score was used to normalize the parameters. The Z-Score method shows how many standard deviations a certain value is far from the mean of the whole data set and is calculated through equation 1 (Al Shalabi & Shaaban, 2006):

$$Z = \frac{X - \mu}{\delta} \quad (1)$$

Where σ is the standard deviation and μ is the mean, Z is a z-score for X. The value of the z-score is between 3 and -3. If the z-score is 0, it is on average. A positive or negative z score indicates a higher or lower raw score than the mean, respectively (McLeod, 2019). The pseudo-code of extracting users' navigational behaviors is as figure 2:

First part – Extract user navigation behavior
Input: *webserver log*
Output: *User-Page vector, User-Attribute vector*

For log in webserver log:
Step 1. Remove logs with invalid status code
Step 2. Remove logs of images, videos, CSS and scripts request
Step 3. Remove logs of robots and crawler request
Step 4. Define each user with IP address and respective agent
For user in webserver log:
Step 5. Assign each 30 minutes of remaining on webpages as a session
$User\text{-}Page = \{ p_i^1 \; p_i^2 \ldots \; p_i^j \}$
Return User-Page
For user in webserver log:

$$total\_page\_duration_{u_i} \leftarrow \sum_{j=1}^{n} page\_duration_{u_i}^{p_j}$$

$$total\_page\_count_{u_i} \leftarrow \sum_{j=1}^{n} page\_count_{u_i}^{p_j}$$

$$total\_session\_count_{u_i} \leftarrow \sum_{j=1}^{n} session\_count_{u_i}^{p_j}$$

$$number\_of\_days_{u_i} \leftarrow \sum_{j=1}^{n} days_{u_i}^{j}$$

$$avg\_page\_duration_{u_i} \leftarrow total\_page\_duration_{u_i} / total\_page\_count_{u_i}$$
$$avg\_session\_per\_day_{u_i} \leftarrow total\_session\_count_{u_i} / number\_of\_days_{u_i}$$
$$avg\_page\_per\_session_{u_i} \leftarrow total\_page\_count_{u_i} / total\_session\_count_{u_i}$$
$$last\_reffer_{u_i} \leftarrow current\_date \; - \; last\_session\_reffer_{u_i}$$
$$User\text{-}Attribute = \{avg\_page\_duration_{u_i}, number\_of\_days_{u_i},$$
$$avg\_session\_per\_day_{u_i}, avg\_page\_per\_session_{u_i}, last\_reffer_{u_i}\}$$

Return User-Attribute

Figure 2: The pseudo-code of extracting users' navigational behaviors

Part II. Extracting Webpages Topics (Content Mining)

- **Preprocessing and Preparation of Contents**

At this part, the concepts contained in the content of each web page were extracted. At the time of this study, the website had 374 pages, and the contents of each page were extracted and reviewed separately. First, the text on each page was extracted using the OutWit Hub software. This software identifies and extracts links, images, documents, phrases, and more, and converts the extracted data into tables (Briediene, Kilpys & Krilavičius, 2019). Then, the extracted text entered the PyCharm programming environment of Python, and the cleaning and preprocessing operations were applied to them.

First, to unify the data, forms, letters, and additional and irrelevant words, all the same letters with different spelling and coding were replaced with an equivalent letter. Then, all the abbreviations and shapes in the text were removed, and the tokens were identified by calculating the distance between the words left over from the original text. In this step, the textual content is divided into smaller units called tokens. Finally, the stop words were removed. Stop words include conjunctions, prepositions, verbs, adverbs, adjectives, and the like that are repeated repeatedly in the text but do not have a specific meaning and will not play a role in the processing. Since many English words have been used in the text of the reviewed website, the English-language Stop words have also been removed.

- **Extracting the Topics of Webpages**

In this part, topics related to each page were extracted using the Latent Dirichlet Allocation (LDA) algorithm. This algorithm is a three-level hierarchical Bayesian model in which each element is a combination of a set of topics and each topic is modeled as a combination of the main set of probabilities (Blei, 2012). In this way, the text of all the web pages was given as an input to this algorithm, and a set of topics with a set of keywords was received as the output. Then, the dependence coefficient of each word on the subject was determined, which indicates the degree to which the word describes the subject. For each topic, words were selected with a dependence coefficient of at least 0.03. Then, the dependency coefficient for five to 30 topics on the page was calculated to determine the optimal number of topics on each page (Table 2). The pseudo-code of extracting the topics of web pages is as figure 3:

Second part – Extract topics of webpages
Input: *webpages text*
Output: *Page-Topic vector*

For text in webpages text:
Step 1. Replace letters that have more than one format and coding with specified format
Step 2. Replace signs, abbreviations and figures with an empty space
Step 3. Extract tokens based on spaces
Step 4. Remove stop words
Step 5. Topic modeling with LDA()
$Page\text{-}Topic = \{t_T^1\ t_T^2 \dots t_T^s\}$
Return Page-Topic

Figure 3: The pseudo-code of extracting the topics of web pages

According to Table 2, the dependence coefficient of five topics on each page is higher than the other values, but because this number does not cover many of the topics on the website, 25 topics per page were selected.

Table 2
Dependency Coefficient for 5 To 30 Topics

Topic No.	Degree of Dependency	Topic No.	Degree of Dependency
5	0.509	19	0.442
7	0.439	21	0.474
9	0.453	23	0.470
11	0.436	25	0.476
13	0.474	27	0.422
15	0.400	29	0.457
17	0.445	-	-

Table 3 shows the keywords related to each topic. For each topic, words with a dependence coefficient of at least 0.03 were selected. According to the extracted topics and their keywords, it can be said that in most cases, the keywords were properly distributed in the topics, but some topics included irrelevant keywords. The reason for this is the low volume of content on the pages of the studied website.

In the table the words, “Alcatel”, “Sam”, “Zirowatt”, “Sanam”, “Edifier”, “Zirowatt”, “Sanam”, “Huawei”, “Ronix”, “Lenovo”, “Sony”, “Sahand”, "Speedo", "Banio", "Kennedy", "Burger", "Matin", "Sedir", "Princeley", "Zova", "Snooze", "Beach", "Diar", "Nami", "Simin", "Barez", "Hitech", "Adidas","Javid","Benjamin","Datis","Paliz", and "Jahan" are the brands of the products.

Table 3
Keywords Specifying Each Topic

Topics	Keywords	Topics	Keywords
Topic 1	Battery, Black, Telephone, Alcatel, Internal, Clock, Wire	Topic 13	Sweeper, Kennedy, Red, Silver, Vacuum, Cleaner, Desktop
Topic 2	Refrigerator, Freezer, White, Sam, Side, Dishwasher, Steel	Topic 14	Grading, Cutting, Burgers, Scissors, Hoses, Sinks, Saws
Topic 3	Machine, Washing Machine, Underwater, White, Silver, Sanam, Consumption	Topic 15	Control, Sound, Base, Ruler, Wire, Soundbar, Player
Topic 4	Phones, Mobile Phones, Hedphones, Edifiers, Cables, Huawei, Audio	Topic 16	TV, Desk, Smart, Matte, Cedar, Curved, Tablet
Topic 5	Ronix, Screw, Carrying, Four-Way, Industrial.	Topic 17	Otto, Princely, Zova, Eau De Cologne, State, Titanium, Snooze
Topic 6	Laptop, Speed, Screen, Quality, Display, Charger, Lenovo	Topic 18	Car, Steering Wheel, Engine, Pillow, Game, Nnao, Console
Topic 7	Camera, Film, Sony, Soundbar, Frame, Digital, Display.	Topic 19	Ball, Volleyball, Shaving, Car, Multi, Beach, Princely
Topic 8	Swimming, Sahand, Football, Hat, Soccer, Water, Speed	Topic 20	Homeland, Business, Name, Simin, Travel, Blanket, Finishing
Topic 9	White, Wall, LED, Bath, Toilet, Smoky, Desktop	Topic 21	Fireplace, Heater, Datis, Hood, Second, Turquoise, Benjamin
Topic 10	Speaker, Flash, Home, Memory, Soft, Audio, Player	Topic 22	Capability, Simple, Small, Cool, Cooler, Users, Gas
Topic 11	Tires, Prominent, Car, Freezer, Rio, Pride, Peugeot	Topic 23	Lcteric, Glass, Hair Dryer, Kettle, Saw, Flame
Topic 12	High-Tech, Digital, Yellow, Thermos, Pink, Adidas, Push	Topic 24	Dessini, Car Wash, Car Iron, Pot, Grantine, World, Paliz
Topic 25	Box, Javid, Badminton, Features, Rocket, Pair, Necklace		

To extract the topics and their keywords, the distribution of each topic on the web pages was also determined. Finally, the page-subject vector was obtained for each page through equation 2, which indicates the degree to which the page r is dependent on subject s:

$$Page\_Topic_r^s = \{t_r^1\ t_r^2\ \dots t_r^s\} \quad (2)$$

Part III. Extracting interests of users

In this part, users’ favorite topics and the amount of this interest were determined. Topic modeling of pages and the techniques of web usage mining has been used to extract users’ interests, which has not been done before on a Persian language website. For this purpose, the dependence coefficient of each subject-page (page-subject/page-title vector, obtained from part II) was multiplied by the number of times that page was observed by the user (user -page vector obtained from part I) by matrix (equation 3 and 4):

$$User\_Topic_i^s = User\_Page_i^j \times Page\_Topic_r^s \quad (3)$$

$$User\_Topic_i^s = \{ut_i^1\ ut_i^2\ \dots ut_i^s\} \quad (4)$$

The resulting user-subject vector demonstrates the interest of user i in subject s.

In the next step, users with similar interests were clustered, and finally, the interest of each cluster was obtained. The dependency distribution algorithm as one of the new techniques for user clustering was used for clustering. There is no need to specify the number of clusters (He, Chen, Wang, Xu, Bai & Meng, 2010). The APCluster () function from a library with the same name was used to implement this algorithm. This library is written in R. The input of this algorithm was a similarity matrix, which was obtained through the negDistMat () function. By setting the parameter r in this function, the similarity between the two data points is calculated based on the negative Euclidean distance. The pseudo-code of extracting the interests of users is as figure 4:

Third part – Extract interests of user
Input: User-Page vector, Page-Topic vector
Output: Cluster-Topic vector

Step 1. $User\_Topic_i^s \leftarrow User\_Page_i^j \times Page\_Topic_r^s$
$User\_Topic_i^s = \{ut_i^1\ ut_i^2 \dots ut_i^s\}$
Step 2. Cluster users with common interest in topics with APCLUSTER()
Step 3. $Cluster\_Topic_C \leftarrow \sum_{ic=1}^{i} \sum_{sc=1}^{s} User\_Topic_{ic}^{sc}$

Figure 4: The pseudo-code of extracting the interests of users

Using this clustering algorithm, twenty-five clusters were obtained. Then, the Davis-Boldin index was used to evaluate clustering, which was 0.613 for 25 clusters.

In the next step, to determine the label of each group of users, a topic that the group was most interested in was used. Then, the interest rate of each cluster in each topic was determined by aggregating the interest rate of all users in each topic (equation 5). In this way, the tropics that each group of users was most interested in were determined, and that topic was considered the cluster tag.

$$Cluster\_Topic_C = \sum_{ic=1}^{i} \sum_{sc=1}^{s} User\_Topic_{ic}^{sc} \quad (5)$$

Finally, to determine the significant difference between the clusters, the ANOVA test was used at the significant level of P<0.05. The results of this test for each topic in each cluster are given in Table 4. It is well noted that the sequence of topics for each cluster shows the users' interest rates from high to low. According to the table, the P-value for all variables is lower than 0.05, which shows the heterogeneity of the population and different means of clusters.

Table 4
The ANOVA Test Results

Cluster	Favorite Topics of Each Cluster	ANOVA Test		
		Sum Sq	Mean Sq	F-Value Pr (>F)
1	Photographic and Video Equipment, Digital/Mobile Devices; Home Appliances/Electrical Appliances; Kitchen Appliances/Refrigerator	193	8.04	108<2e-16***
2	Home Appliances/Home & Kitchen Appliances; Digital/Mobile Devices; Kitchen Appliances/Refrigerator; Cosmetics; Home Appliances/Cooling Equipment	254	10.60	120<2e-16***

Cluster	Favorite Topics of Each Cluster	ANOVA Test		
		Sum Sq	Mean Sq	F-Value Pr (>F)
3	Jewelry & Sport Equipment; Kitchen/Refrigerator Appliances; Home/Kitchen Appliances; Home Appliances/Customer Goods; Digital/Mobile Devices	115	4.81	48.8<2e-16***
4	Home/Consumables/Digital/Mobile Devices; Home Appliances/Cooling Equipment; Home/Kitchen Appliances	263	10.96	63.6<2e-16***
5	Car/Consumable; Car/Accessories; Digital/Mobile Devices; Kitchen Appliances/Washing Machines; Home/Kitchen Appliances	7.9	0.331	5.98<2e-16***
6	Car/Accessories; Digital/Mobile Devices; Kitchen Appliances/Washing Machine; Home/Kitchen Appliances	12.9	0.536	9.46<2e-16***
7	Jewelry & Sport Facilities; Photographic & Video Equipment; Home/Kitchen Appliances; Car/Consumer Devices; Digital/Mobile Devices	555	23.11	270<2e-16***
8	Photographic & Video Equipment; Automotive/Peripheral; Digital/Mobile Devices; Home Appliances/Cooling Equipment; Home & Kitchen Appliances	124	.17	123<2e-16***
9	Car/Accessories; Kitchen/Fridge Electrical Appliances; Digital/Mobile Devices; Home & Kitchen Appliances; Kitchen Appliances/Washing Machine	63.1	2.63	44<2e-16***
10	Kitchen Appliances/Washing Machine; Digital/Mobile Devices; Home Appliances/Cooling Equipment; Home & Kitchen Appliances; Home Appliances	2.8	0.1172	8.1<2e-16***
11	Digital/Mobile Devices; Home & Kitchen Appliances; Home Appliances/ Video & Audio Devices; Home Appliances/Cooling Equipment; Gardening Tools	842842	35.1	275<2e-16***
12	Kitchen Appliances/Refrigerator; Car/Accessories; Home Appliances/Video & Audio Devices; Gardening Tools; Home And Kitchen Appliances	225	9.37	140<2e-16***
13	Car/Accessories; Car/Consumable; Home & Kitchen Appliances; Digital/Mobile Devices; Jewelry & Sport Facilities	240	9.99	159<2e-16***
14	Home/Consumable; Jewelry & Sports Facilities; Digital/Mobile Devices; Car/Consumable Equipment; Photographic & Video Equipment	72.5	3.023	35.3<2e-16***
15	Home Appliances/Cooling Equipment; Digital/Mobile Devices; Home Appliances; Home Electrical Appliances; Car/Accessories	6.2	0.240	3.24<2e-16***
16	Digital/Mobile Devices; Car/Accessories; Jewelry & Sports Facilities; Home Appliances/Cooling Equipment; Tourism & Travel Consumable	18	7.77	99<2e-16***
17	Car/Consumer Devices; Car/Accessories; Digital/Mobile Devices; Kitchen Appliances/ Washing Machine; Home Appliances	15	0.624	11.5<2e-16***
18	Home Appliances/Audio Devices; Jewelry & Sports Facilities; Self/Accessories; Home & Kitchen Appliances; Gardening Tools	425	17.71	139<2e-16***
19	Home Appliances/Cooling Equipment; Jewelry &	12.7	0.528	13.1<2e-16***

Cluster	Favorite Topics of Each Cluster	ANOVA Test		
		Sum Sq	Mean Sq	F-Value Pr (>F)
	Sports Facilities; Digital/Mobile Devices; Photographic & Video Equipment; Home Electrical Appliances			
20	Home Appliances & Consumable; Car/Accessories; Digital/Mobile Accessories; Kitchen Appliances/Washing Machine; Jewelry & Sport Facilities	4.1	0.1717	5.18<3.8e-16***
21	Home/ Electrical Appliances; Digital/Mobile Devices; Home Appliances/Cooling Equipment; Home Appliances/Home & Kitchen Appliances; Kitchen Appliances/Washing Machine	421	17.53	138<2e-16***
22	Sport Facilities; Digital/Mobile Devices; Home Appliances/Home Electrical Appliances; Cooling Equipment; Home & Kitchen Appliances	470	19.6	188<2e-16***
23	Sports Facilities; Gardening Tools; Car/Accessories; Sports/Consumable; Home & Kitchen Appliances	317	13.22	146<2e-16***
24	Sanitary Ware; Jewelry & Sport Facilities; Car/Accessories; Gardening Tools; Home Appliances/ Video & Audio Devices	8	0.333	6.01<2e-16***
25	Home Appliances & Consumable; Car/Accessories; Gardening Tools; Digital/Mobile Devices; Home Appliances/Audio & Video Devices	517	21.55	196<2e-16***
Sig: 0.001'***'; 0.01 '**'; 0.05'*'				

Part IV. Modeling Based on the Users' Interests and Behavioral Characteristics

In this part, the behavioral parameters of each user were combined with the user-interests vector for modeling. Then, the label for each user was identified according to the cluster to which it belonged and the label for that cluster. In this way, 16 classes of users were identified. Some of these categories, along with the features of the users in each category, are described in Table 5. These features were considered for each group of users, and classification was based on it to build a model.

Table 5

Some Information Extracted For Each Class of Users

Label	Subject Viewed By Users	Behavioral Characteristics of Users
Home Appliances & Consumable	Home Appliances/Consumable; Car/Consumable; Digital/Mobile Devices; Home Appliances/Home & Kitchen Appliances; Kitchen Appliances/Washing Machine; Sanitary Ware	Users who visited the website within one to three days had one session per day; 490 days have passed since their last visit; they viewed on average 5 pages per session, and the average time taken to visit each page is 365 seconds.
Kitchen Appliances/Refrigerator	Car/Accessories; Kitchen Appliances/Refrigerators; Sanitary Ware; Home Appliances/Audio & Video Devices; Gardening Tools	Users who visited the website within one to three days had one session per day; 690 days have passed since their last visit; they viewed on average 4 pages per session, and the average time taken to visit each page is 334 seconds.

Digital/Mobile Devices	Home Appliances/Home & Kitchen Appliances; Car/Accessories; Home Appliances/Audio & Video Devices; Digital/Mobile Devices; Home Appliances/Cooling Equipment	Users who visited the website within one to three days had one session per day; 464 days have passed since their last visit; they viewed on average 4 pages per session, and the average time taken to visit each page is 328 seconds.
Sport Facilities	Car/Consumer Devices; Car/Accessories; Sports Facilities; Home Appliances & Consumables; Home Electrical Appliances	Users who visited the website within one to three days had one session per day; 141 days have passed since their last visit; they viewed on average 3 pages per session, and the average time taken to visit each page is 310 seconds.
Car/Accessories	Car/Accessories; Digital/Mobile Devices; Digital Accessories/Computer Accessories; Home Electrical Appliances	Users who visited the website within one to three days had one session per day; 770 days have passed since their last visit; they viewed on average 5 pages per session, and the average time taken to visit each page is 273 seconds.

Since the extracted labels in this study are the subject of web pages and are a multi-class classification issue, the SVM algorithm was used for classification. The inputs of this algorithm were the behavioral characteristics and interest vectors, and the algorithm's outputs were the user's tags. In modeling, 80% of the data (2344 users) was used as training data, and 20% (577 users) was used as test data. The CARET library was used for modeling. The pseudo-code of modeling based on the users' interests and behaviors are as figure 5:

Fourth part – Construct model based on user interests and behaviors
Input: User-Topic vector, Cluster-Topic vector, User-Attribute
Output: forecast model

Step1. Choose second, third and fourth highest topics according to cohesion value as interests of user
Step 2. Choose first topic as label of cluster
Step 3. Place together topics of interest, attributes of user and label it with label of cluster that user belong
Step 4. Construct model and classify users with SVM()

Figure 5: The pseudo-code of modeling based on the users' interests and behaviors

After modeling, the model was evaluated using the F-measure and confusion matrix. With the implementation of the proposed model, the accuracy was 0.376, and the F-measure was 0.452. The reason for the low accuracy of the model is the low relation and content of the web pages.

Discussion

This research aimed to obtain more accurate information on users' navigational behavior

and their interests in visiting the websites Table 6 indicates the top five topics of interest in each cluster and the degree of interest of users to that topic. The topic with the highest degree of interest was selected as the cluster label, slightly different from subsequent topics. Therefore, choosing these values as a cluster label has been an excellent way to categorize users based on it. As can be seen in the table, in most cases, the fifth topic is further away from the second, third, and fourth topics. From the fifth issue onwards, the values are further reduced and therefore not considered in the cluster evaluation. Marketers can use the information in table 6 to predict the users' following behavior and improve customer satisfaction by providing personalized items. For example, in cluster 1, users who have seen the Items of topic 7 (i.e., Jewelry and Sports Facilities, Photographic and Video Equipment, Home and Kitchen Appliances, Car/Consumer Goods or Digital/Mobile Devices) in the first step will see the items of topic 4 in the second step (i.e., Home/Consumables, Digital/Mobile Devices, Home Appliances /Cooling Equipment, Home Appliances or Home and Kitchen Appliances). Besides, they will see the items of topic 22 in the third step (i.e., Car/Accessories; Car/Consumables; Home and Kitchen Appliances; Digital/Mobile Devices; Jewelry and Sports Facilities).

Table 6
Top Five Topics of Each Cluster

Cluster	First Topic	First Value	Second Topic	Second Value	Third Topic	Third Value	Fourth Topic	Fourth Value	Fifth Topic	Fifth Value
1	Topic 7	193	Topic 4	50.8	Topic 22	36.9	Topic 13	16.8	Topic 3	14.1
2	Topic 21	177	Topic 4	69.1	Topic 3	42.7	Topic 17	31	Topic 22	28.9
3	Topic 25	153	Topic 3	93.5	Topic 21	43.1	Topic 1	307	Topic 4	29.8
4	Topic 1	92.5	Topic 4	42.2	Topic 22	35.9	Topic 23	33.4	Topic 21	25.4
5	Topic 11	154.3	Topic 18	35.2	Topic 4	15.7	Topic 3	11.6	Topic 21	10.9
6	Topic 21	224.8	Topic 18	171.2	Topic 6	55.4	Topic 17	51	Topic 4	36.2
7	Topic 25	169.1	Topic 7	59.8	Topic 21	35.5	Topic 11	32.1	Topic 4	26.7
8	Topic 7	110.1	Topic 18	60.8	Topic 4	34.8	Topic 22	24.1	Topic 21	14.8
9	Topic 18	144.5	Topic 2	131.5	Topic 4	90.7	Topic 21	76	Topic 3	50.6
10	Topic 3	149.6	Topic 4	134.7	Topic 22	87.8	Topic 21	68.3	Topic 23	46.4
11	Topic 4	9.5	Topic 21	96.2	Topic 16	86.7	Topic 22	79.6	Topic 14	76.3
12	Topic 2	180.1	Topic 18	88.3	Topic 16	30.4	Topic 14	28.2	Topic 21	27.6
13	Topic 11	48.8	Topic 18	45.6	Topic 21	33.7	Topic 4	31.6	Topic 25	27.5
14	Topic 1	119.8	Topic 25	89.4	Topic 4	42.2	Topic 11	41.4	Topic 7	15.4
15	Topic 22	269.8	Topic 4	152.9	Topic 23	34.4	Topic 13	27.6	Topic 18	27.3
16	Topic 4	172	Topic 18	130.6	Topic 25	104.3	Topic 22	27.6	Topic 20	23.6
17	Topic 11	343.3	Topic 18	171.5	Topic 4	101.8	Topic 7	56.3	Topic 23	39.1
18	Topic 16	125	Topic 25	68.3	Topic 18	58.5	Topic 21	16	Topic 14	15.8
19	Topic 22	101.2	Topic 25	82.7	Topic 4	73.1	Topic 7	28.8	Topic 23	21.8
20	Topic 23	127.2	Topic 18	33.8	Topic 4	20.9	Topic 3	16.4	Topic 25	12.3
21	Topic 13	197	Topic 4	75.5	Topic 22	59.2	Topic 21	57.7	Topic 3	44.7
22	Topic 8	116.7	Topic 4	46.8	Topic 13	32.9	Topic 22	0.1	Topic 21	27.3
23	Topic 12	143.1	Topic 14	30.7	Topic 18	28.7	Topic 19	24.6	Topic 21	21.2
24	Topic 9	70.4	Topic 25	55.2	Topic 18	45.5	Topic 14	27.3	Topic 16	15.2
25	Topic 23	141.7	18	140.5	Topic 14	105.5	Topic 4	37.5	Topic 16	31.9

Besides, some of the results of user grouping are as follows:

- A significant percentage of people who visit this website are looking for home appliances, including kitchen and electrical appliances.
- The second favorite item among this website's users is mobile because it is seen in a high percentage of groups.

- Some groups include people who do not have a specific purpose in their search and randomly select pages.
- Users who are looking for a camera and video equipment are also interested in digital and mobile devices.
- Users who are interested in the car and its equipment are also interested in digital and mobile devices.

With this information, new users can be placed in each of these 25 groups according to the pages they refer to, and their favorite topics can be extracted as well. In addition, their next navigational behaviors can be predicted. In addition to extracting users' favorite topics, their behavioral parameters have also been extracted to build the model. Some of the results of users' navigational behavior are as follows:

- Most of the users have an average of one to two sessions per day. This means that many users of this website conduct their surveys purposefully.
- Users with the highest number of sessions have spent more time since their last visit to the website. It can be stated that these users are not loyal customers of the website.
- Users who visit the website purposefully are people who are generally looking for home appliances and digital devices.
- Considering the targeted customers, the website manager should consider that the users have access to the desired product in their first little navigation. This is crucial in designing a website and choosing its layers to access each product.

Conclusion

Today, e-commerce websites need to know more about their customers to increase their satisfaction and loyalty, which is possible by analyzing and understanding their behavior and interests. Studying users' needs and interests and anticipating their online behaviors is part of web personalization. Therefore, this study presented a general framework of web personalization for a Persian website using an integration of web content and usage mining. Integrating content and usage mining will lead to identifying different groups of customers to predict new customers' future behavior and interests. Previous studies such as Gurbas et al. (2013) have tried to add content features to the results of usage mining to increase the accuracy and quality of patterns and facilitate their understanding. However, researchers have generally combined the results of usage mining with the web anthology. However, similar to the results of Herwanto (2016), the results of our study indicated that integrating web usage and content mining will allow removing subjectivity from profile data and keeping it updated. Besides, including semantic information in creating a navigation pattern more accurate recommendations can be provided. This indicates an increase in the quality of the patterns.

The data used in this study was the real data from a Persian language retail website. Since the syntax and semantics of the Persian language are different from English and other languages, recognizing nouns and pronouns, parts of speech tagging, finding word boundaries, tracing, or manipulating characters are also different (Habib, 2021). Besides, the research design and methods used in this study were the combinations of the methods used in previous studies. For example, LDA was used for content mining that is similar to the studies of Abdi Ghavidel et al. (2015), which used LDA for Persian metaphor classification and its frequency

prediction; Du et al. (2020), Yang & Zhang (2018), Herwanto (2016) and Shotorbani et al. (2016) which used LDA for content mining. Besides, similar to Habib (2021) and Shotorbani et al. (2016), the SVM was used for classification. However, no similar model that combines both content and usage mining for predicting customers' behavior was used in this study, which distinguishes it from previous studies and makes it impossible to compare the results with previous studies.

This study will help the literature in the field of web personalization in three ways: first, in this research, five parameters were introduced and calculated to identify users' behavioral patterns; Second, to extract users' interests, topic modeling of web pages along with web mining techniques was used, which had not been done before in a Persian-language website and is the innovation of this study. The third contribution of this study to the literature in this field is the use of a dependency distribution algorithm for clustering user navigational patterns that have not been used in previous studies and is one of the innovations of this study.

The importance of this study is in combining web server logs with the content of webpages that can be used as a basis for products and services recommendations. The most important result of the proposed system is that the patterns of users' navigation are understandable; the subsequent analyses will be much simpler.

Owing to the internet's nature and the lack of physical customers, meeting customers' needs and improving the quality of services require accurate knowledge of customer priorities; customers, however, are generally not interested in long queries or filling out forms. Therefore, online retailers must gather customers' preferences from the interactions and information provided by the sales process; it is essential to know how their customers use their websites. Inferring useful results requires in-depth data analysis. Thus, the integration of web usage and content mining is crucial for optimally designing the structure of sales websites, increasing the attraction of potential customers, and retaining existing ones. Because e-commerce has become a competitive space for online retailers, electronic marketers should focus on increasing customer satisfaction. For this purpose, e-business and electronic commerce websites should have quick and accurate access to customers' needs, customize web pages accordingly, and provide personalized products and services. The combination of the results of web content and usage mining increases the accuracy and quality of the recommender systems, helping them achieve more personalization that is effective. Therefore, online retailers can use the results of this study to better personalize their websites. In addition, the results can help organizations gather customer data, suggest products, send personalized messages to customers, and more.

In designing a website, having a clear understanding of user-profiles and site goals does not seem to be enough. Site designers also need to have proven knowledge of the user's navigational behaviors and interests and the way they choose to visit pages. Therefore, website visitor behavior analysis and the content analysis of the website would be a powerful tools that can be used to gather valuable tips on measuring a website's success in achieving its expected goals.

User interest profiles obtained from a combination of content and usage mining can be used to predict the link and personalize the website's content. Moreover, goods and services can be recommended based on the user's interests. The results are beneficial for the studied website, and other websites can use the results of this study to find ways to satisfy customers, make data-driven decisions, and improve and develop their decision-making strategies. It also helps managers promote their products and services and provides customers with personalized items

to gain new customers and maintain their existing customers.

The results can be used for better and wiser website design. In addition, web designers can use this information to optimally design their web pages according to users' interests and provide links that are more relevant to customers' interests. Web developers can also use the results of this study to improve their content and page displacement.

The proposed model enables the company to identify its valuable customers and recommend appropriate goods and services. The results of this study can help website managers to understand the behavior of website users better.

Limitations and Recommendations for Future Studies

In this study, users' access to the cached pages and information about the time spent on these pages were not recorded so it is a suggestion for future research. In addition, the model's accuracy was low due to the low volume of website content and the weak relation between the contents of the web pages. Therefore, doing the same study on a website with more related items and content is recommended.

Moreover, in this study, several parameters have been extracted to determine the users' navigational behavior in referring to the website. However, it is suggested that future studies develop these parameters and identify users more accurately by extracting more features.

Furthermore, topic modeling was used in this study. Future research could use other techniques in natural language processing (NLP and ontology-based methods) for content mining. In addition, this study used a subjective tag to categorize users. It is suggested that future studies use several subjective labels to model users. Besides, the proposed model can be used on other online shopping websites.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Akanfe, O., Valecha, R. & Rao, H. R. (2020). Design of an inclusive financial privacy index (INF-PIE): A financial privacy and digital financial inclusion perspective. *ACM Transactions on Management Information Systems*, 12(1), 1-21. https://doi.org/10.1145/3403949

Al Shalabi, L. & Shaaban, Z. (2006, May). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *2006 International conference on dependability of computer systems* (pp. 207-214). IEEE. https://doi.org/10.1109/DEPCOS-RELCOMEX.2006.38

Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M. & Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. *Expert Systems with Applications*, 40(18), 7478-7491. https://doi.org/10.1016/j.eswa.2013.07.040

Beheshtian-Ardakani, A., Fathian, M. & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision.*

Science. Letter, 7(1), 39-54. https://doi.org/10.5267/j.dsl.2017.4.005

Bharti, P. M. & Raval, T. J. (2019, June). Improving web page access prediction using web usage mining and web content mining. In *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1268-1273). IEEE. https://doi.org/10.1109/ICECA.2019.8821950

Bin, W., & Zhijing, L. (2003, September). Web mining research. In *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003* (pp. 84-89). IEEE. https://doi.org/10.1109/ICCIMA.2003.1238105

Blei, D. M. (2012). Probabilistic topic models. *Communication of the ACM*, 55(4), 77-84. https://doi.org/10.1145/2133806.2133826

Briedienė, M., Kilpys, V. & Krilavičius, T. (2019). Media analysis that reflects the spread of antichristian opinion. In CEUR Workshop proceedings [electronic resource]: IVUS 2019, International conference on information technologies, Kaunas, Lithuania, 25 April 2019. Aachen: CEUR-WS, 2019, Vol. 2470, (pp. 125-129). https://hdl.handle.net/20.500.12259/102061

Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J. & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. *Expert Systems with Applications*, 39(12), 11243-11249. https://doi.org/10.1016/j.eswa.2012.03.046

Chitraa, V. & Thanamani, A. S. (2011). A novel technique for session's identification in Web usage mining preprocessing. *International Journal of Computer Applications.* 34(9), 23-27. Retrieved from https://b2n.ir/p35850

Dai, H. (2005). Integrating semantic knowledge with Web usage mining for personalization. In *Web Mining: Applications and Techniques* (pp. 276-306). IGI Global.

Das, R. & Turkoglu, I. (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with* Applications, 36(3), 6635-6644. https://doi.org/10.1016/j.eswa.2008.08.067

Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation. *Engineering Applications of Artificial Intelligence*, 87, 103279. https://doi.org/10.1016/j.engappai.2019.103279

Eirinaki, M. & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology*, 3(1), 1-27. https://doi.org/10.1145/643477.643478

El Aissaoui, O., El Madani, Y. E. A., Oughdir, L., & El Allioui, Y. (2019). A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments. *Education and Information Technologies*, 24(3), 1943-1959. https://doi.org/10.1007/s10639-018-9820-5

Fong, A. C. M., Zhou, B., Hui, S. C., Hong, G. Y. & Do, T. A. (2011). Web content recommender system based on consumer behavior modeling. *IEEE Transactions on Consumer Electronics*, 57(2), 962-969. https://doi.org/10.1109/TCE.2011.5955246

Grace, L. K., Maheswari, V. & Nagamalai, D. (2011, January). Web log data analysis and mining. In *International Conference on Computer Science and Information Technology* (pp. 459-469). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17881-

8_44

Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M. & Alhajj, R. (2013). Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems*, 49, 50-62. https://doi.org/10.1016/j.knosys.2013.04.014

Habib, M. K. (2021). The challenges of Persian user-generated textual content: A machine learning-based approach. https://arxiv.org/abs/2101.08087

He, Y., Chen, Q., Wang, X., Xu, R., Bai, X., & Meng, X. (2010, March). An adaptive affinity propagation document clustering. In *2010 The 7th 7th International Conference on Informatics and Systems (INFOS)* (pp. 1-7). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/5461817

Herwanto, G. B. (2016). *Kombinasi Web Usage Mining Dan Web Content Mining Untuk Menghasilkan Profil Navigasi Pengunjung Website*. Doctoral dissertation, Universitas Gadjah Mada. Retrieved from http://etd.repository.ugm.ac.id/penelitian/detail/93650

Jalali, M., Mustapha, N., Sulaiman, M. N. & Mamat, A. (2010). WebPUM: A web-based recommendation system to predict user future movementshttps://www.sciencedirect.com/journal/expert-systems-with-applications, 37(9), 6201-6212. https://doi.org/10.1016/j.eswa.2010.02.105

Kalaignanam, K., Kushwaha, T. & Rajavi, K. (2019). How does web personalization create value? Lower cash flow volatility or enhanced cash flows. Lower cash flow volatility or enhanced cash flows (May 8, 2019). Kenan Institute of Private Enterprise Research Paper, (19-21). http://dx.doi.org/10.2139/ssrn.3384432

Khatri, R. & Gupta, D. (2015, July). An efficient periodic web content recommendation based on web usage mining. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)* (pp. 132-137). IEEE. http://dx.doi.org/10.1109/ReTIS.2015.7232866

Kumar, S. & Kumar, R. (2021). A study on different aspects of web mining and research issues. In *IOP conference series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012018). IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012018/pdf

Kumar, V., & Thakur, R. S. (2018). Web usage mining: Concept and applications at a glance. In *Handbook of Research on Pattern Engineering System Development for Big Data Analytics* (pp. 216-229). IGI Global. https://doi.org/10.4018/978-1-5225-3870-7.ch013

Liu, B. & Chen-Chuan-Chang, K. (2004). Editorial: Special issue on web content mining. *ACM SIGKDD Explorations Newsletter*, 6(2), 1-4. https://doi.org/10.1145/1046456.1046457

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

McLeod, S. A. (2019, May 17). *Z-score: definition, calculation and interpretation*. *Simply Psychology*. Retrieved from https://www.simplypsychology.org/z-score.html

Mobasher, B., Dai, H., Luo, T., Sun, Y. & Zhu, J. (2000, September). Integrating web usage and content mining for more effective personalization. In *International conference on electronic commerce and web technologies* (pp. 165-176). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44463-7_15

Modi, S. & Jagtap, S. (2018). Browser integrated Web content filtering using natural language

processing. *Aeu-Int J Electron C (IJMECE),* 6(4), 108-112.

Nkwo, M., Orji, R., Nwokeji, J. C. & Ndulue, C. (2018). E-commerce personalization in Africa: A comparative analysis of Jumia and Konga. In *PPT@ Persuasive* (pp. 68-76).

Pappas, I. O., Kourouthanassis, P. E., Giannakos, M. N. & Lekakos, G. (2017). The interplay of online shopping motivations and experiential factors on personalized e-commerce: A complexity theory approach. *Telematics and Informatics*, 34(5), 730-742. https://doi.org/10.1016/j.tele.2016.08.021

Pierrakos, D., Paliouras, G., Papatheodorou, C. & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 13(4), 311-372. https://doi.org/10.1023/A:1026238916441

Prathyusha, P., Kumari, B. A., Sumathi, M. & Rangaswamy, K. (2019). The phases of web usage mining for classification and pattern analysis. *Journal of the Gujarat Research Society*, 21(16), 1082-1090.

Rathi, P. & Singh, N. (2019). An efficient algorithm for data preprocessing and personalization in Web usage mining. *International Journal of Computer Sciences and Engineering*, 7 (5), 160-164.

Salman, R. H., Zaki, M., & Shiltag, N. A. (2020). A studying of web content mining tools. *Al-Qadisiyah journal of pure science*, 25(2), 1-16. https://doi.org/10.29350/qjps.2020.25.2.1067

Salonen, V. & Karjaluoto, H. (2016). Web personalization: The state of the art and future avenues for research and practice. *Telematics and Informatics*, 33(4), 1088-1104. https://doi.org/10.1016/j.tele.2016.03.004

Senkul, P. & Salin, S. (2012). Improving pattern quality in Web usage mining by using semantic information. *Knowledge and Information Systems*, 30(3), 527-541. https://doi.org/10.1007/s10115-011-0386-4

Shanthi, S. (2017). Survey on Web usage mining using association rule mining. *International Journal of Innovative Computer Science & Engineering*. 4(3), 65-67. Retrieved from file:///C:/Users/Reza/Downloads/103-Article%20Text-198-1-10-20200515.pdf

Sharma, S. & Rana, V. (2020). Web search personalization using semantic similarity measure. In *Proceedings of ICRIC 2019* (pp. 273-288). Springer, Cham. https://doi.org/10.1007/978-3-030-29407-6_21

Shotorbani, P. Y., Ameri, F., Kulvatunyou, B. & Ivezic, N. (2016, September). A hybrid method for manufacturing text mining based on document clustering and topic modeling techniques. In *IFIP Advances in Information and Communication Technology book series* (IFIPAICT,volume 488) (pp. 777-786). Springer, Cham. https://doi.org/10.1007/978-3-319-51133-7_91

Sujatha, V. (2012). Improved user navigation pattern prediction technique from web log data. *Procedia Engineering*, 30, 92-99. https://doi.org/10.1016/j.proeng.2012.01.838

Sun, L. & Zhang, X. (2004, April). Efficient frequent pattern mining on web logs. In *Asia-Pacific Web Conference* (pp. 533-542). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24655-8_58

Sunikka, A. & Bragge, J. (2012). Applying text-mining to personalization and customization research literature-Who, what and where? *Expert Systems with* Applications, 39(11),

10049-10058. https://doi.org/10.1016/j.eswa.2012.02.042

Taherizadeh, S. & Moghadam, N. (2009). Integrating web content mining into web usage mining for finding patterns and predicting users' behaviors. *International Journal of Information science and Management (IJISM)*, 7(1), 51-66.

Velásquez, J. D. (2013). Combining eye-tracking technologies with Web usage mining for identifying website Key objects. *Engineering Applications of Artificial Intelligence*, 26(5-6), 1469-1478. https://doi.org/10.1016/j.engappai.2013.01.003

Van Aartsen, B., El-Gayar, O. & Noteboom, C. (2020). A systematic review of web usage mining techniques and future research options. In *MWAIS 2020 Proceedings*. Retrieved from https://aisel.aisnet.org/mwais2020/25/

Vijaiprabhu, G. & Meenakshisundaram, K. (2020). User Identification in Web Usage Mining using Data Mining Techniques with Nominal Distance Function and KD Tree. *International journal of analytical and experimental modal analysis*, 12 (3), 273-281. Retrieved from http://www.ijaema.com/gallery/31-ijaema-july-4138.pdf

Yadav, M. P., Feeroz, M. & Yadav, V. K. (2012, July). Mining the customer behavior using web usage mining in e-commerce. In *2012 third international conference on computing, communication and networking technologies (ICCCNT'12)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCNT.2012.6395938

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H.... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems volume*, 14(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Xing, W., Lee, H. S. & Shibani, A. (2020). Identifying patterns in students' scientific argumentation: Content analysis through text mining using latent dirichlet allocation. *Educational Technology Research and Development*, 68(5), 2185–2214. https://doi.org/10.1007/s11423-020-09761-w

Yang, S. & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Information, Control and Computer Sciences*, 11(7), 525-529. https://doi.org/10.5281/zenodo.1317350

Yilmaz, H., & Senkul, P. (2010, December). Using ontology and sequence information for extracting behavior patterns from web navigation logs. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 549-556). IEEE. https://doi.org/10.1109/ICDMW.2010.44