Journal of Distributed Computing and Systems(JDCS), Vol. 4, Issue 2, 2021, Pages 7-12.

Increasing the accuracy of web suggestion system using fuzzy neural network and bio-algorithms

Zahra Abbasnejad*[1], Milad Ghahari Bidgoli[2]
[1]*Department of Computer Engineering South Tehran Branch, Islamic Azad University, Tehran, Iran.*
[2]*Department of Computer Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr, Iran.*

Article History:

Received: 13 October 2021
Received in revised form: 2 January 2022
Accepted: 10 January 2022
Available online: 17 March 2022

Abstract

The growing number of information on the web and the addition of different web pages and websites to this space has made users face problems. These problems appear to users when users are trying to obtain information on a particular topic, and finding all the pages that are suggested to them is a difficult and time consuming process. In the current research, a profile is first created based on the behavioral characteristics of users at different sessions that result from web server logs. These include things like the frequency of user page views, the length of time the user has been on different pages, and the date the page was viewed. We then group them using the clustering method, then fuzzy inference system, extract the fuzzy rules according to the interests of the users and their clusters, and after obtaining the users' movement patterns, they Insertneural network into vector format Other tools such as bio-algorithms can be useful by obtaining optimal parameters in optimizing predictions and increasing accuracy in fuzzy neural network. The evaluation criteria in this study is accuracy.

Keywords: data mining, web mining, user behavior patterns, neural network, fuzzy system, MFO algorithm.

I. INTRODUCTION

We live in an age of information, an age in which humans are more likely than ever to produce and disseminate information. In fact, the information we have is too much for us to analyze. Analyze. Therefore, methods and techniques are needed to efficiently access, share and extract data from the data and use this information. By the end of the twentieth century, Internet users had grown dramatically. These users have created a lot of information on the web and searched for information on the web. We are now witnessing the production of huge amounts of information on the web every day. Most internet users search for information on the web almost every day. Recently, the number of users and the volume of information in the web space has increased significantly due to the various contributions of users in their production. One of the tools available for this is to use web mining. Web mining is the use of data mining techniques to discover knowledge from resources in the field of web and is classified into different fields of research according to the type of resource being explored (1).

Among the personalization techniques that will be used in this research is application-based analysis that can be used to extract users' behavioral patterns and use them in predicting web pages. Appeared. Web mining is one of the research fields that uses data mining techniques to automatically discover and extract information from documents and web services. In fact, web mining is the process of discovering unknown and useful information and knowledge of web data. Web mining methods are divided into three categories based on what kind of data they are exploring: web content exploration, web structure exploration, and web usage exploration. During this research, after introducing web mining and examining its stages, the relationship between web mining and other research fields is investigated and the challenges, problems and applications of this research field are pointed out. Also, each type of web mining will be studied in detail, which in this project will focus more on web mining in the industry. For this purpose, models, algorithms and applications of each class are introduced. One of the important topics in exploring the use of the web is the clustering of web users (2). Web events can be a good source of web users 'behavioral patterns, and the analysis of these patterns will lead to a better understanding of users' tastes, and thus more appropriate and customized services such as the development of adaptive websites, personalized websites. And provide users with suggestion systems and improve the performance of web servers.

The interests of web users can be determined by the web pages they visit and the time spent on those pages. The page retrieval time parameter is an important parameter in analyzing the browsing behavior of web users. The proposed methods for clustering web users can be divided into two parts: traditional algorithms and evolutionary algorithms. Traditional algorithms have limited performance due to the existence of high-volume web events, and have low performance. In contrast, evolutionary algorithms are inspired by the natural behavior of organisms and are suitable for solving such problems.

*Corresponding Author: Mehrasa.Abbasnejad@gmail.com

II. EASE OF USE

A. LITERATURE REVIEW

Web mining: Web mining is the use of data mining techniques to identify and compile global web patterns, with the goal of better understanding the needs of web-based applications. Web mining is an application of data mining techniques to extract knowledge from the web. Web mining has been extracted for a great deal of degrees and techniques and has been suggested for a variety of applications including web search, categorization, personalization and more. Web analytics allows the use of automated devices to display and exit the required information and data from servers and web reports, and allows organizations to organize and unstructured information about browser activity, server logs, websites, and link structure, page content. And access different resources. The term web mining was first coined in 1996 by Etzioni in an article called The World Wide Web or Gold Mine. This article describes web mining as a task-oriented approach (3). In 2006, Agrawal proposed a bidding system that was based on various statistical methods and weighed on the desired features, and based on this, a choice was introduced to the user. In 2007, a web security recommendation system was developed. In this system, an agent was used and these agents collected information and then selected a sample for selection using fuzzy methods.
In the application mining technique for web personalization in the Etzion method (1996), using the improved C.4.5 algorithm, accuracy and memory were used in the improved processing time criteria. In Nasrawi's method (2003), he has improved the exact criterion using the NP method or the nearest index. In the section of hybrid systems for web personalization in the method of Rashidi et al. (2012) (4), they have increased the accuracy and coverage in the system by using a combination of content analysis and application mining and by using algorithm parsing. In the field of web personalization using fuzzy methods and clustering in the monotheistic method (2012)(5), using Dendagram calibration, has increased. In the same section in the method of Gastellano et al. (2011)(6), they have increased the accuracy and recall in the proposed system by using the application of mining.

III. METHAODS

A. *What is Web Mining*

Web mining and uncovers the hidden patterns in the large amount of data. It finds the unknown. relevant and useful information contained in the web documents. Web mining techniques are inspired from data mining techniques. It does not directly uses the data mining techniques due to diverse nature of web data which is available in the form of unstructured, semi-structured, and structured data. For analysis of web documents, There exist several mining tasks and algorithms in the literature. Unlike data warehousing, web has mixed type of data e.g. content data (text, audio, video, and graphics), structure data (hyperlinks, web graph), and usage data (web log data). On the basis of types of data used, web mining can be categorized as web content mining, web structure or link analysis mining, and web usage mining (7).

Web mining is an application of data mining techniques to extract knowledge from the web. Web mining has been extracted for a great deal of degrees and techniques and has been suggested for a variety of applications including web search, categorization, personalization and more. web mining involves three methods.

- Explore web content.
- Explore the structure of the web
- Explore web applications

Ways to explore web content: 1- Web agents 2- Database

B. *Personalization*

The nature of web personalization is the ability to adapt a website to the needs and interests of its users. Recognizing the interests of users can be based on the knowledge gained from the interests of previous users of the site (8). Typically, each user profile extracted from web log file data (website user records) reflects the common interests of a group of users of all tastes. These profiles are used to provide personalized suggestions. Clearly, the quality of user profiles is crucial to the performance of a personalization system.

Basic requirements of web personalization system:

- User identification: The personalization system interacts directly with the user and obtains the required information from his behavior. Therefore, there should be an appropriate mechanism to identify and differentiate users.
- Receive data used by users : The personalization system must be able to collect all the data related to the users. The type and volume of this data depends on the performance of the system.
- Data preparation : The collected data must be pre-processed to eliminate noise and convert it to a suitable format.
- Build user models : A key component of the system is the personalization of the user model, which includes information that the system maintains about the user's interests, knowledge, goals, and preferences. Model making can be done manually or automatically (9).

IV. DISCUSSION

Web mining is the use of data mining techniques to automatically extract information from documents and web services Due to the diversity of information in this field, web mining tasks can be divided into the following four categories.

- Finding information: In this section, information is retrieved and extracted offline or online from text documents on the web. Data is retrieved from web resources such as e-newsletters, newsgroups, HTML documents, and text databases.
- Information selection and preprocessing: In this section, after selecting the information, the initial processing is done automatically on the retrieved information.

- Generalization: In this section, general patterns are discovered on websites using data mining and machine learning methods.
- • Analysis: Finally, this section examines and validates the extraction patterns. A model that is verified is accepted and presented.

In this section, we first provide definitions of metadata generated and used by web servers. ”Fig.1” shows the structure of an HTTP transaction between an HTTP client and a server.

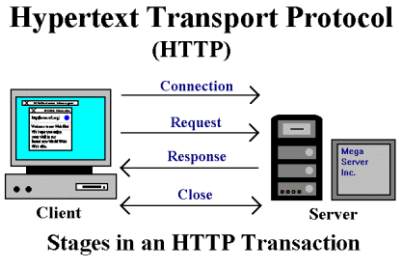

Fig. 1. HyperText Transport Protocol(HTTP)

A. *Recognition of data*

In an HTTP transaction, the underlying data is defined with the following metadata.

1) IP address of the customer's machin
2) User ID if the HTTP authentication process is done.
3) When the server processes the request.
4) HTTP method (POST, GET)
5) Request URI
6) Protocol and protocol version: such as “HTTP 1.1“, “HTTP 1.0“ and ...
7) HTTP status code that is sent to the client.
8) Response size in bytes
9) Referrer is the “URI“ from which customer reports are referred.
10) User agent which includes information: such as: browser name, its version and operating system on which the browser is running.
This data has 5 features which are:

1) Host: The host from which the request was sent. This part has two types Is hosting. If the host is available, its address, etc. In this case, the IP address. Relevant is shown. Due to the fact that there are two types of hosts in this section, IP addresses can be specified with the number 1 and host addresses with the number 2.

2) Exact time of request: This feature shows the exact time of request. The format of this feature is: ”DAY MON DD HH: MM: SS YYYY”, respectively. Which shows the day of the week, the name of the month, the day of the month, the hour, the minute, the second and the year.

3) Request: This feature indicates the user's request.

4) HTTP Request Response Code: The response code based on the HTTP protocol standard is described in the table(I).

5) Response byte rate: This feature also specifies the response byte rate.

TABLE I. RESPONSE CODE HTTP

Response Code	Status
200	Success
302	Found redirect status
304	Not Modified
403	Forbidden
404	Not Found

B. *Registration file information*

The information provided by this dataset includes: IP address, user ID, access time and date, HTTP request method, resource path on the web server, name of the page the user is viewing, The protocol used for the transfer, the status of the code returned by the web server and the number of bytes transferred. In some cases, instead of IP, the name of a site is displayed, which is for hosts who refer to the site, if they have a name, their name is saved, otherwise their IP is stored.
In this data set, the pages whose response size is marked with zero indicate an error in this request, which means that the user could not view this page. Such error pages, as well as requests for graphic files and requests for bots and web browsers, should be removed from existing data

C. *Create session vectors*

After removing the extra information, it is time for the most important step of data preprocessing. In this step, called session recognition, a list of pages that the user has viewed while logged in is obtained. In order to equate the situation with other similar tasks and to better evaluate and compare the method with other methods, in this research, as in most researches, a threshold of 30 minutes has been used to identify the sessions. In this research, a preprocessing software has been used to perform session building.
At the end of this phase, 73762 sessions were created from website log file. The output of this step is a file containing sessions created so that it can be used in later steps.

D. *Clear meeting*

Once the sessions are identified, it is time to clear the sessions. At this stage, pages that can be said to be present in all sessions (as the first and main pages of the site) as well as pages that are rarely seen in this set of sessions are usually removed. What is presented in this research as a suggestion to increase the accuracy of the proposing system is applied in this stage of creating the system. As mentioned in the previous chapter, in this system, a sequence of page views that are repeated in a session will be removed from the session as the wrong paths in link selection due to the length of time they are viewed. The procedure is that, for example, if in a duplicate sequence the page view time is less than a threshold, the sequence is removed from the session. In this study, this time is considered 5 seconds. According to the results obtained from 73762 sessions obtained in the previous stage, 26735 sessions have been removed and its number has been reduced to 47027 sessions (Table II shows the results obtained in this stage in different data sets). At this point, a large number of pages were removed from the sessions as erroneously viewed

pages and useless sessions, thus significantly reducing system load.

TABLE II. RESULT OBTAINED IN DIFFERENT DATA SET

Average length of sessions by applying the proposed method	Average length of sessions without applying the proposed method	Number of sessions after clearing	Total Sessions
4	4.5	47027	73762

E. Fuzzy system production

The first step required to generate fuzzy rules is for the data to have a target category and attribute. Given that the relevant data is unsupervised and does not have a target property, so using the ”X-Means” clustering algorithm, first obtain the optimal number of clusters, then add this number of clusters to the clustering algorithm. Provide Fuzzy ”C-Means” classification or make a fuzzy classification on the data. This allows the data to be cohesive and grouped.

One of the most important reasons for using the ”X-Means” clustering algorithm is that in algorithms such as ”K-Means” or ”C-Means” it is necessary to enter the number of clusters from the beginning. Therefore, considering that the researcher does not have any information about the number of optimal clusters from the beginning, it is necessary to first obtain the number of optimal clusters using an algorithm such as ”X-Means” and then according to the optimal number of k which indicates the number of clusters. - Optimal, the target attribution operation is performed. The following model shows an overview of ”X-Means” clustering to obtain optimal clusters from the data set (10).

In order to simulate this part, first the lower limit and the upper limit of the number of clusters are determined. The ”X-Means” clustering algorithm is then executed on the data according to the low to high limit. The clustering error is then calculated and compared with the previous error rate. This process continues until the number of optimal clusters is reached. How to achieve the optimal cluster is based on comparing the desired error with the next and previous errors. Therefore, in this study, the low limit of the number of clusters is zero and the upper limit of the number of clusters is 100. By performing the data clustering process using the ”X-Means” clustering algorithm, 4 clusters are determined as the optimal clusters”Fig.2”.

Fig. 2. Optimal Clusters

F. Moth-Flame-Optimization (MFO) Algorithm

The Butterfly Flame Algorithm ”MFO” is an innovative oven algorithm inspired by allegiance that mimics the traversal behavior of a type of impeller called the ”MOTH”. In this algorithm, the flares adjust their trajectory according to the flames with the flame. The main idea of this algorithm comes from the fact that the night owls adjust their trajectory according to the moonlight and move along a straight line (11). The implementation steps of the propeller flame algorithm ”MFO” are as follows.

1) Value the algorithm parameters
2) Making the initial population of Shaparak randomly
3) Evaluating the position of each shopper and calculating its competence
4) Steps 0 to 11 are repeated until the stop condition is met.
5) The number of flames is updated. The algorithm initially considers the number of flames equal to the number of flares. But during execution it reduces the number of flames to converge.
6) Quantifying the flames. To determine the flames, we consider the best experience of any shopper so far as a flame.
7) Parameter a is updated.
8) The position of each shopper is updated.
9) Calculating the updated the butterfly.
10) If the level of merit of the new shop is better than the best answer, the new shop will be considered as the best answer.

If the stop condition is not met, it skips to step 5, otherwise the program ends. Find the minimum Sphere function using the MFO algorithm: The Sphere function is a well-known and so-called benchmark function in optimization problems. And is one of the functions used to test the power of evolutionary algorithms. The mathematical form of this function is as follows.

$$f(x) = \sum_{i=1}^{d} (x)_i^2 \qquad (1)$$

The function form is as follows ”Fig.3”:

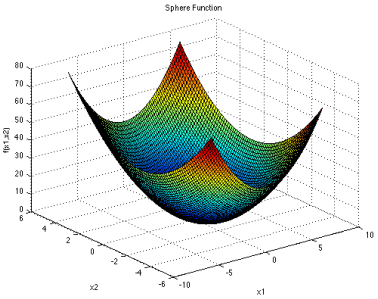

Fig. 3. Sphere Function

Calculate the number of flames according to the following equation:

$$flame\ no = \text{round}\ (N - 1 * \frac{N-1}{T}) \qquad (2)$$

At the beginning of the algorithm, we assumed the flame vector to be empty, so in the first iteration, the flames are

selected from the flares, and in subsequent iterations, the best experience of the flares and the existing flames is selected.
The output of the program for the Sphere function is as follows:
Best fitness = 3.5118e-18 Best solution found is: -4.9361e-10 -1.2759e-09 -6.8326e-11 -1.2629e-09 2.0139e-10

As described in the previous section, the number of optimalclusters was obtained. In this section, according to the number of optimal clusters, the process of determining the target property to produce fuzzy rules using the Mamdani system is performed.
Data clustering using the ”C-Means” clustering algorithm

Cluster 0 : 29 items
Cluster 1 : 250 items
Cluster 2 : 1076 items
Cluster 3 : 143 items

Validation in referral systems is often used on matters related to accuracy and recall. In this research, these criteria have been used to evaluate the system. Rate coverage is the number of times the model is able to predict a number of requests. In pattern recognition and data retrieval with binary classification, accuracy (also called positive predictive value) is associated with a fraction of retrieved samples, while retrieval (also known as sensitivity) is related. Is a fraction of the related samples that have been recovered.
The accuracy is equal to the number of correct system diagnoses on the number of retrieved sets:

$$Precision = \frac{|\{relevant\ pages\} \cap \{retrived\ pages\}|}{|\{retrived\ pages\}|} \qquad (3)$$

A system similar to that presented in Castellano's 2011 paper was used to evaluate and compare the system. The proposed system has also been compared with similar systems using ”KNN” and ”New-Biz” algorithms(table III).

TABLE III. RESULT

	KNN	**Naive Bayes**	**suggested method**
Accuracy	85.04%	88.19%	92.6%

V. CONCLUSION

The information on the web is growing day by day, and this process has led to the production of large volumes of interconnected pages that are not logically organized. Therefore, analyzing the exploration behaviors of web users and examining the real interests of users has become particularly important. By identifying the behavior of users, their interests can be discovered and using the obtained results, desirable information can be provided to users. Providing information that users are interested in viewing will make the web a user-friendly and engaging environment. By comparing the studied method, it was concluded that thestudied method had a better performance than othercompared methods in terms of accuracy.

Because the display of knowledge in the real world is somewhat uncertain and one user's interests are different from another, and on the other hand, typically, users do not express their interests clearly, obtaining a movement pattern of users in These will be difficult. In the studied method, using fuzzy clustering and modeling methods provided by fuzzy neural network, the structure has been shown to face different tendencies of users in which uncertainty has major problems in research. Done in this area, it helps us.

The purpose of the proposed method is to create a user profile, find their common movement patterns and provide a list of pages that the user is interested in viewing. Finding users' movement patterns is done by using web mining application and fuzzy clustering method and then extracting the relevant fuzzy rules. Also, the use of biological algorithms in this section has improved fuzzy rules. The proposed system uses a fuzzy neural network to find a suitable movement pattern for users and based on it predicts future requests of users. It can be concluded that the better performance of the proposed system is due to common interests that different movement patterns may have with each other. This highlights the importance of fuzzy clustering, which allows clusters to overlap.

In general, in order to improve the results and optimize therecommendation of web requests, in this research, the following steps are performed, which are: 1- Pre-processing and cleaning of data in order to convert unused data into logical data, 2- Preparation of optimal clusters using ”X-Means” optimal clustering algorithm, 3- Clustering data using ”C-Means” clustering algorithm using ”X-Means” algorithm, 4- Generating fuzzy rules with Use of fuzzy inference system, 5- Optimization of rules generated using biological algorithm 6- Application of fuzzy rules on neural-fuzzy neural networks and making suggestions using MFO algorithm. In this proposed method, it has improved 7.56% compared to ”KNN” algorithm and 4.41% compared to ”naive bayes” algorithm.
the results of the experiment show that the proposed algorithm has high accuracy in the number of pages proposed to users. therefore, comparing the results of other methods with the results of the proposed model which is a combination of biological, fuzzy and neural networks, show that our method is more favorable from the point of view of performance evaluation criteria and request - based applications

REFERENCES

[1] *Gadamsetty Vasavi and Dr. T. Sudha, "Information Extraction from Online Shopping Sites using Web Content Mining Methods and Techniques", International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT-2017).*

[2] *Anurag Kumar, Ravi Kumar Singh, "A Study on Web Structure Mining", International Research Journal of Engineering and Technology (IRJET),volume: 04 Issue: 1 , Jan -2017.*

[3] *O. Etzioni, The world wide web: Quagmire or goldmine, Communications of the ACM, 39(i1), pp. 65-68,1996.*

[4] *Rashidi, S., Harounabadi, A & Dezfouli, M. (2012). Prediction of users' future requests using neural network.Management Science Letters , 2(6), 2119-2124.*

[5] *A. K. Santra, S. Jayasudha, " Classification of web log data to identify interested users using naïve bayesian*

classification ", International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, pp. 381-387, 2012.

[6] G. Castellano, A. M. Fanelli and M. A. Torsello, "NEWER: A system for NEuro-Fuzzy Web Recommendation ",Applied Soft Computing, vol.11,issue 1, pp. 793-806, 2011.

[7] D. Pierrakos ,G. Paliouras , C. Papatheodorou ,C. D. Spyropoulos, " Web usage mining as a tool for personalization: a survey, User Modeling and User-Adapted Interaction ", ACM Journal, Vol. 13, Issue 4, pp. 311-372, 2003.

[8] Peng Guan, Yuefen Wang, "Personalized scientific literature recommendation based on user's research interest", Natural Computation Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) 2016 12th International Conference on, pp. 1273-1277, 2016. [11] Nawal Sael, Abdelaziz Marzak, Hicham Behja, "Web Usage Mining data preprocessing and multi level analysis on Moodle", Computer Systems and Applications (AICCSA) 2013 ACS International Conference on, pp. 1-7, 2013.

[9] M. K. Khribi, M. Jemni and O. Nasraoui, "Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval," 2008 Eighth IEEE International Conference on Advanced Learning Technologies, 2008, pp. 241-245, doi: 10.1109/ICALT.2008.198.

[10] Żelasko, D.; Książek, W.; Pławiak, P. Transmission Quality Classification with Use of Fusion of Neural Network and Genetic Algorithm in Pay&Require Multi-Agent Managed Network. Sensors 2021, 21, 4090. https://doi.org/10.3390/s21124090

[11] Ruba Abu Khurma, Ibrahim Aljarah, Ahmad Sharieh. An Efficient Moth Flame Optimization Algorithm using Chaotic Maps for Feature Selection in the Medical Applications. In Maria De Marsico, Gabriella Sanniti di Baja, Ana L. N. Fred, editors, Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2020, Valletta, Malta, February 22-24, 2020. pages 175-182, SCITEPRESS, 2020

How to cite: Z. Abbasnejad, M.Ghahari Bidgoli. Smaisim. Increasing the accuracy of web suggestion system using fuzzy neural network and bio-algorithms, Journal of Distributed Computing and Systems(JDCS), Vol 4, Issue 2, Pages 7-12, 2021.