International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 8866-8876
© Research India Publications. http://www.ripublication.com

Dominance Rank Fuzzy Clustering and Distributed Probability Graph for Web User Behaviour Mining

Ulaganathan. N

HOD, Assistant Professor, National Institute of Fashion Technology - Tirupur Exporters Association College of Knitwear Fashion
East of Tirupur Exporters Knitwear Industrial Corporation, Small Industries Development Corporation, Mudalipalayam, Tirupur, (Affiliated to Bharathiar University), Coimbatore-641046, Tamilnadu, India.

Abstract

Web usage mining examines the navigation patterns in web access logs and extracts the past unknown and valuable information to accessed web pages. This strategies helps for different web-oriented applications such as website framework called Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) is designed. The main objective is in investigating the relation of cognitive styles through Dominance Rank with other types of navigation behavior and therefore the user's interactions with other web objects in a significant manner. Initially, the URLs are visited by the user and date and time of the visits is collected through server log files. After that, Separation of relevant and irrelevant information regarding web user is performed using Dominance Rank model. Then the Fuzzy clustering is performed on the relevant information to group the web pages which are similar to the user interests with available data in log files. Finally, Distributed Probability Graph Arc (DPG) model is used to provide optimized latency and reducing the cache utilization. Experimental evaluation is performed with ECommerce Web Logs to show the performance of DFC-DPG framework in terms of true positive rate, clustering efficiency, latency and cache utilization with number of data regarding the web user and web user queries

Keywords: Web usage mining, Dominance rank model, Fuzzy clustering, Distributed probability graph arc,Cache utilization,Latency

INTRODUCTION

With the progression of the internet and at the same time with the popularity of the web has increased a great deal of significance among the researchers to web usage mining. However, large voluminous data is accessible in the web that failed to provide the necessary user information. Web usage mining extracts the most essential information from huge voluminous data in the web logs, possessing information pertaining to accessed web pages. Due to this huge voluminous data, it is advantageous in handling relatively small group of related and essential data, instead of dealing with valuable data all together. Clustering techniques are employed with the few techniques for partitioning the data.

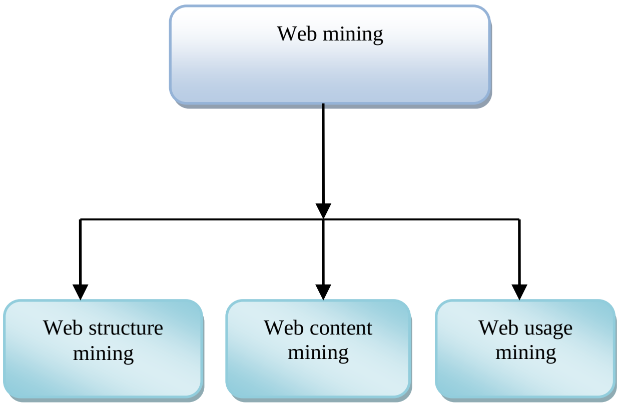

Figure 1: Types of Web mining

Figure 1 shows the different category of Web mining. Web mining is the application of data mining methods to determine and extract the information from the internet (i.e. web). Web mining is generally partitioned into three different categories such as web structure mining, web content mining and web usage mining according to the types of data to be mined. Web structure mining employs graph theory for analyzing the node and association structure of a web site. Web structure mining is the process of finding the structure information from the web. Web content mining is the other types of web mining which is used for extracting useful information from the contents of web documents. Finally, the Web usage mining monitors web access logs navigation patterns with the objective of extracting previous unknown and essential information to predict the future access of the web user. Clustering is applied to identify the user queries with similar characteristics. That means a group of users related navigation patterns are clustered on a web site. Therefore, the proposed DFC-DPG framework uses the web usage mining for extracting the most essential information from the web services through clustering process. There are several web mining techniques has been presented to analyze the web user behavior.

Web service ranking approach was designed in [1] depends on collaborative filtering (CF) by evaluating the user behavior with their past access to collect the potential user behavior. However, clustering was not performed to group the similar user interested data. A Hybrid Sequence Alignment Measure (HSAM) was performed in [2] for measuring the distance between session pair on the basis of user navigated paths to cluster web sessions and measure the cluster quality. However, latency of the web user was not reduced by improving the server performance.

A comprehensive adaptive interactive system was introduced in [3] for modeling the users' cognitive styles with the pattern of user navigation and click stream data. However, the navigation metrics showed promising results in terms of navigation behavior, cognitive styles of users with other types of navigation behavior, remains an open issue. A new approach called MiND (Mining Neubot Data) was introduced in [4] for automatic and effective clustering of users with a similar Internet access behavior. But Internet access parameters were not developed with different frequency schemes.

The user browsing pattern analysis was performed in [5] for determining the more accurate browsing behavior using k means clustering. However, time taken for prediction was not discussed. An improved K-means clustering algorithm was designed in [6] for discovering the internet user behavior analysis. But the clustering efficiency and cache utilization was not improved.

A neuro-fuzzy based hybrid model was developed in [7] for detecting variations in user browsing behavior in web. However, the performance of latency for responding the user query was not discussed. Dynamic recommendation technique was introduced in [8] to the whole web users particularly unregistered ones of E-commerce site. It also reduced the false positive rate but it failed to evaluate the true positive rate.

Web Log Expert tool was presented in [9] for analyzing the user behavior patterns on web with the help of Web Access Logs and Log Analyzer. However, it failed to recognize the customer behavior analysis. Fuzzy Semantic Search Engine (FSSE) was developed in [10] by using fuzzy logic to collect similarities of terms in the web. It also used to classify the web pages and selects the appropriate domain for searching web pages. However, the cache utilization of storing the multiple user information was not investigated.

The certain issues observed from the above said reviews such as failed to perform clustering, more latency, and more cache utilization and failed to recognize the customer behavior analysis. In order to overcome such kind of issues, Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is introduced.

The major contribution of the paper is described as follows,

Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is developed for extracting the most essential information from huge data in the web logs. The information (i.e. data) related with the web user is collected through a web server log files which comprises a list of activities regarding the web user.

Dominance Rank model is applied for separating the relevant and irrelevant data through the spearman rank correlation measure. If the rank correlation coefficient provides positive correlation, then the data is said to be a relevant to web user. Otherwise, it is irrelevant data to the web user.

Therefore, relevant data regarding the web user is selected and remove the irrelevant data.

Fuzzy clustering is applied on the relevant data for grouping the web pages with similar user interests from web usage data available in server log files based on the user queries. By applying fuzzy clustering, the centroid and fuzzy membership of data to the cluster center is measured with Euclidean distance. The data points are clustered based on minimum distance between them.

Distributed Probability Graph Arc (DPG) model is used for constructing the graph based on history of the web user activities. The density of the session is used to provide the relationship among the web pages. This helps to identify the interests of user while navigate through a web site.

The rest of the paper is organized in following structure. In Section 2, Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is explained briefly with neat diagram. In Section 3, experimental settings are described and the analysis of results is presented in Section 4. Section 5 introduces the reviews related to the research works. The conclusion of work is presented in section 6.

Dominance fuzzy clustering and distributed probability graph framework

Web mining is the process used for extracting the accurate information from the huge number of data in internet. There are several data mining techniques like clustering, classification, rule generation are mostly used for mining the significant information from the internet. But the navigation behavior and perceived latency and cache utilization remains major challenging issue. In order to overcome such kind of issues in World Wide Web (i.e. internet), Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is developed. The flow processing diagram of the DFC-DPG framework is shown in figure 2.

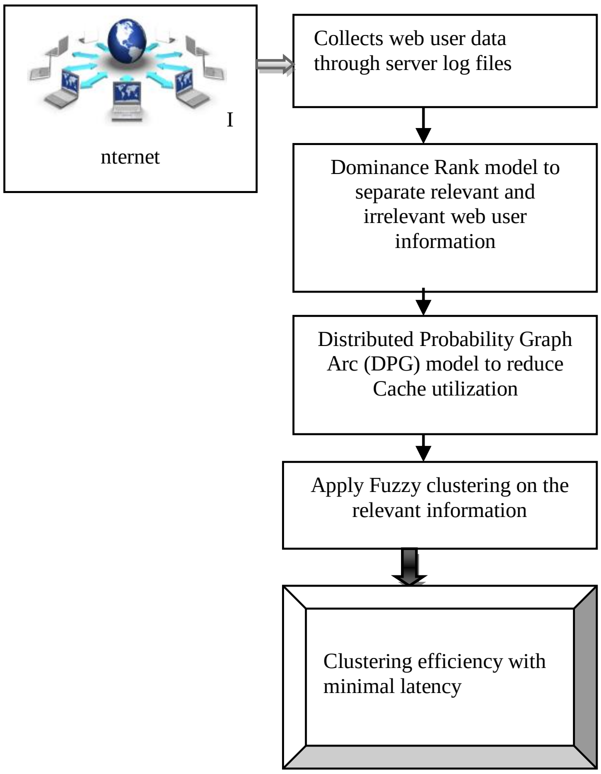
I
nternet
Collects web user data through server log files
Dominance Rank model to separate relevant and irrelevant web user information
Distributed Probability Graph Arc (DPG) model to reduce Cache utilization
Apply Fuzzy clustering on the relevant information
Clustering efficiency with minimal latency

Figure 2: Flow processing diagram of the Dominance Fuzzy Clustering and Distributed Probability Graph framework

Figure 2 illustrates the Flow processing diagram of the Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework. Initially, the data related with the web users are collected through server log files. After that, the Dominance Rank model is applied to separate the relevant and irrelevant web user data. Followed by, the Fuzzy clustering is applied for grouping the users having similar access sequence (i.e. user sessions). Finally, Distributed Probability Graph Arc (DPG) model is applied to reduce Cache utilization and latency through the analysis of past access of web user to extract the future access of the web user. Brief discussion about the DFC-DPG framework is presented in next subsections.

Collection of web user information

The first step in the design of DFC-DPG framework is collects the web user data through the log files. Web server log file is the text file with one line for every web user queries. Each line in the log file has following information like host making the request, timestamp, requested URL, HTTP reply code and bytes in reply which is visited by the user and so on. The other log files as Parse Log which is extracted from the web server log file. The extracted fields are IP address, hostname, date, time and request. These data are stored in a database for easier and effective data handling.

A web server log is a log files automatically generated and preserved by a server containing a list of actions it performed. The web user information is collected from the server log filed on a web site. An example line of the access log in common log format is described as below,

$$123.456.789.110 - [DD/MM/YYYY:HH:mm:ss - 0400]"GET/HTTP/1.0"200\ 3245$$

The above log file format contains the Client IP address (i.e. 123.456.789.110) , User ID, Access date (DD/MM/YYYY), Access time (HH: mm: ss). Http request contains the number of method as shown in table 1.

Table 1: HTTP request method

S. No	HTTP request method	Description
1	GET	Request to read data from server
2	HEAD	Request to read a web page header
3	PUT	Request to server for storing the data
4	TRACE	Request to resend the received request
5	DELETE	Request to delete data
6	CONNECT	Request to connect another host
7	OPTIONS	Request to ask the server to return the list of supported request

Table 1 shows the different types of HTTP request methods used in server log file. These requests are used for mining process at different Server level. The other log files as resource path on the Web server, Protocol used for the transmission (Hyper Text Transfer Protocol). Status code returned by the web server (200) this is indicated as OK. The other status code like 100 it indicates Continue the process, 300 represents Multiple Choice, the code 400 indicates Bad Request, 403 indicates Forbidden, 404 represents Not Found, 503 denotes a Out of Resources and so on. Number of bytes being transmitted is indicated as 3245. This data about the web user is visited by the site is collected to extract the relevant data and remove the irrelevant data.

Dominance Rank model

Once the web user data is collected, Dominance Rank model is applied to separate the relevant and irrelevant data from the collected web user data. Due to huge amount of data on the web in the form of text, image, video, audio and so on. It is very hard to discover relevant data for a web user. Dominance Rank model is used to remove the unwanted data such as image files, script files, HTTP response code. The relevant data about the web user such as hit ratio (i.e. time to fetch the data), number of visits, time spent. The irrelevant data are bad request, unauthorized access, not found, null request and so on. This information is separated by applying Dominance Rank model through spearman rank correlation measure.

The Spearman rank correlation is applied to identify the degree of relation between the data regarding the web user. Spearman rank correlation is measured as follows,

$$\rho = correlation\ (D_1, D_2) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

From (1), where, d_i is the difference between the rank of data and 'n' is the number of data $D_1, D_2, .. D_n$ and ρ is the correlation coefficient,(D_1, D_2) represents the data infrastructures. A perfect Spearman correlation coefficient (ρ) provides the results as positive correlation '1' when the two data being evaluated are related to identify the relevant and irrelevant data regarding the web user. If the coefficient provides negative correlation, then the data is not related to web user. From the correlation measure, the data which is irrelevant to the web user and related to the web user is identified. Dominance ranking algorithm is described as follows.

```
Input: Number of collected data ', Spearman correlation coefficient '$\rho$'

Output: Separate irrelevant and relevant information
Step 1: Begin
Step 2:   For each collected data
Step 3:          Calculate the Spearman rank correlation coefficient using (1)
Step 4:       If( positive correlation '+1')
Step 5:          Identify data related to web user
Step 6:       else If (negative correlation '-1')
Step 7:          Identify data irrelevant to web user
Step 8:       End if
Step 9:   End for
Step 10: End
```

Figure 3: Dominance ranking algorithm

Figure 3 denotes the Dominance ranking algorithm for detecting the relevant data and irrelevant data regarding the web user. If the coefficient provides positive correlation '+1', then the data are related to web user. Otherwise, the data are irrelevant to the web user. Therefore, the data which are relevant regarding the web user is selected and remove the irrelevant data.

Fuzzy clustering approach

After the relevant data extraction from the Dominance Rank model, Fuzzy clustering approach is performed for grouping the similar user interest's web pages available in server log files based on the user query. In general, clustering is the process of grouping which user visits the similar pages of a web site based on their relationship. The main aim of clustering is that, the webpage's within a group is similar to one another and dissimilar from the web pages in other groups. In DFC-DPG framework, the user sessions are clustered based on the order in which users visit different pages of a web site. The user sessions are generated based on the host name and time fields. These sessions are stored in a database. In general, a session contains number of web pages visited by a user in the sequence within a specified time. For example, a user visits pages P1, P3, P5 of a web site in a series, then, the session is represented as follows,

$$S = (P1, P3, P5) \tag{2}$$

From (2), Where S is the sessions. The cluster analysis is used to group the frequent pages browsed by the users and the results produce who are interested to use similar kind of pages for improving the web usage quality. Therefore a fuzzy clustering algorithm (FCA) is used to construct clusters with uncertain boundaries. Therefore, clustering allows that one object belongs to multiple clusters with some fuzzy membership degree. In fuzzy clustering, data points (i.e. users visit similar kind of pages of a web site) are possibly be a member of multiple clusters. The number of fuzzy cluster is initialized. The FCA partitions the number of data points (i.e. web pages) $P = \{p_1, p_2, p_3 \ldots\ldots, p_n\}$. The set of clusters are denoted as,

$$C_k = \{c_1, c_2, c_3, \ldots., c_k\} \tag{3}$$

Any data point P includes a set of coefficients providing the degree in the kth cluster. By applying fuzzy, the cluster centroid is the mean of all the data points, weighted by their degree of associated with the cluster. The centroid of the cluster is calculated as follows

$$v_i = \frac{\sum_{i=1}^{n}(\mu_{ij})^m p_i}{\sum_{i=1}^{n}(\mu_{ij})^m} \tag{4}$$

From (4),v_i denotes a center of cluster , μ_{ij} represents the fuzzy membership of ith data to jth cluster center. 'n' denotes a number of data points (i=1,2,…n and j=1,2,3…n) where each data point μ_{ij} reveals the degree to which data point p_i

belongs to cluster C_j. Therefore, the fuzzy membership function is calculated as follows,

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{D_{ij}}{D_{ik}}\right)^{\frac{2}{m-1}}} \qquad (5)$$

From (5),D_{ij} represents the Euclidean distance between of all session's ith and to their jth cluster center. From (5), m denotes a fuzzifier determines level of cluster fuzziness. The fuzzy membership function converges to $\mu_{ij} \in [0,1]$. Then the Euclidean distance is measured as follows,

$$D_{ij} = \|p_i - C_j\|^2 \qquad (6)$$

Therefore, the FCA aims to minimize an objective function is described as follows,

$$\arg\min_C \sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}{}^{m}\|p_i - C_j\|^2 \qquad (7)$$

As a result, Fuzzy clustering algorithm is used to group the user who are visited the similar kind of pages from web usage data available in log files based on the user queries. The algorithmic representation of fuzzy clustering is shown in figure 4.

```
Input: number of user queries, set of data points
$P = \{p_1, p_2, p_3 \ldots\ldots, p_n\}$ and
$v_i = \{v_1, v_2, v_3, \ldots. v_n\}$ , number of cluster
$C_k = \{c_1, c_2, c_3, \ldots. c_k\}$
Output: Clustering efficiency
Step 1: Begin
Step 2:   For each data points
Step 3:     Define number of cluster
Step 4:       Randomly select the cluster centroid
Step 5:         Measure the fuzzy center ($v_i$) using (4)
Step 6:       Calculate the fuzzy membership using (5)
Step 7:         Repeat the process step 5 until the minimum 'j' value is attained
Step 8: end for
Step 9: end
```

Figure 4: Fuzzy clustering algorithm

Figure 4 shows the algorithmic description of fuzzy clustering process. For each data points, the number of cluster is defined. Then the cluster center is assigned to the entire cluster. After that, the fuzzy membership function is measured based on Euclidean Distance between the data points. The data points are grouped with minimum distance between them. This process is repeated until all web pages are clustered. This helps to improve the clustering accuracy.

Distributed Probability Graph Arc model

To provide optimized latency, a heuristic probabilistic framework is designed using Distributed Probability Graph Arc (DPG) model. The DPG model is said to be distributed in nature since the user level deployment is performed at the server, whereas the client hosts manages pre-fetching at the client's cache across the network. A DPG model performed based on web user' session is transformed into the structure of a graph, which are represented as density of a session based on a graph theory. In general, the web users visit the web pages and their behavior includes information collection and browsing. The history of the web user activities is used to construct a relationship network among web sites.

Let us consider the web pages $wp_1, wp_2, \ldots wp_n$ are visited at the similar session and wu_1 and wu_2 are web users. The adjacency concept is applied to provide more visited pages in web sites. In Distributed Probability Graph Arc (DPG) model, a graph contains two components vertex and arc. The graph is a set of next web pages and it is transformed into an adjacent matrix. An approach of adjacent matrix and a directed graph of the session in web server are shown in figure 5.

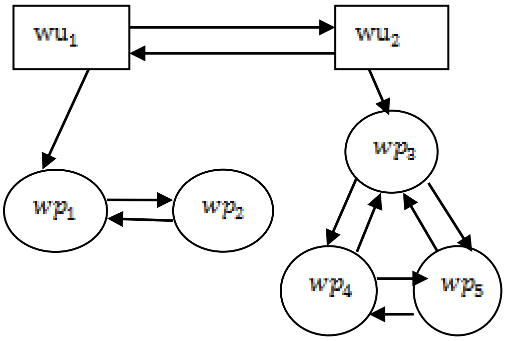

Figure 5: Graph arc model

	wu_1	wp_1	wp_2	wu_2	wp_3	wp_4	wp_5
wu_1	0	1	0	1	0	0	0
wp_1	0	0	1	0	0	0	0
wp_2	0	1	0	0	0	0	0
wu_2	1	0	0	0	1	0	0
wp_3	0	0	0	0	0	1	1
wp_4	0	0	0	0	1	0	1
wp_5	0	0	0	0	1	1	0

Figure 6: Adjacent matrix

Figure 5 and 6 shows the graph arc model and adjacent matrix of web user visited the web pages. An adjacent matrix is an nxn square matrix (a_{ij}). Then the distributed probability is described as,

$$A = a_{ij} = \begin{cases} 1 & if\ wp_i\ vistited\ \ adjacent\ to\ wp_j \\ 0 & otherwise \end{cases} \tag{8}$$

From (8), if two web pages wp_i and wp_j are adjacent, then the probability is said to be 1. Otherwise, the probability is 0. A directed graph consists of vertex and arc. A vertex is a web pages and arc (i.e. link) means a relation between the two web pages.

The arrangement of network in a session is required to analyze quantitatively. The density of a graph is used for modeling the structure. The density of a network is defined as the number of pages among web pages is determined by how many web pages visited and the degree to which one web page connected to the others. The density represents the degree connected actually by maximum relation which is measured as follows,

$$Density = \frac{a}{v(v-1)/2} \tag{9}$$

From (9), 'a' denotes a number of arcs and v denotes a number of vertexes in a network. The range of density is lies between 0 to 1. Then the density of the session is measured as follows,

$$Density\ of\ Session = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(t_{ik}+t_{jk})}{n(n-1)} \tag{10}$$

From (10), n is number of web pages in a session. t_{ik} and t_{jk} denotes duration time in a session. Higher density of the session provides the more relationship among the web pages. Therefore, it shows the user interest of web pages. The group of similar web pages is stored in a server log files. This helps to reduce the Cache utilization. Cache defines the collection of web pages of the similar type stored in a server. As a result, a Probability Graph Arc model uses the past access of the web user to extract the future access of the web user. As a result, the Probability Graph Arc model minimizes the latency.

EXPERIMENTAL EVALUATION

An effective Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is experimented using Java language with the use of ECommerce Web Logs to extract the useful information in the internet. This paper describes the implications of the new trends in web development languages on log file analysis for e-commerce. Electronic commerce or ecommerce is a term for any kind of business, or commercial transaction, which involves the transfer of information across the Internet. The performance evaluation of DFC-DPG framework compared with existing approach Web service ranking approach [1] and HSAM [2]. The following metrics such as true positive rate, clustering efficiency, latency and cache utilization are evaluated to show the performance of DFC-DPG framework.

RESULTS AND DISCUSSION

Result analysis of Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is described in this section. The DFC-DPG framework is compared against the existing Web service ranking approach [1] and HSAM [2]. The results are discussed with the factors true positive rate, clustering efficiency, latency and and cache Utilization. Experimental results are compared and analyzed with the help of table and graph.

Impact of true positive rate

True positive rate is defined as the ratio of the number of (i.e. no. of) relevant data is identified to the number of information available in website. The formula for true positive rate is expressed as follows,

$$TPR = \frac{No.of\ relevant\ data\ correctly\ identified}{No.of\ data} * 100 \tag{11}$$

From (11),TPR is the true positive rate and It is measured in terms of percentage (%).

Table 2: Tabulation of true positive rate

Number of data	True positive rate (%)		
	DFC-DPG	Web service ranking approach	HSAM
10	80.35	70.36	62.47
20	83.24	73.65	65.88
30	84.20	76.42	69.78
40	85.69	78.15	72.36
50	86.78	82.65	75.85
60	88.10	85.46	78.65
70	90.65	87.46	82.24
80	91.47	88.98	85.47
90	92.78	89.12	86.45
100	94.58	90.45	87.65

Table 2 describes the true positive rate with three different techniques DFC-DPG framework, Web service ranking approach [1] and HSAM [2]. Among the three different methods, the proposed DFC-DPG framework increases the performance results in terms of true positive rate than the existing methods.

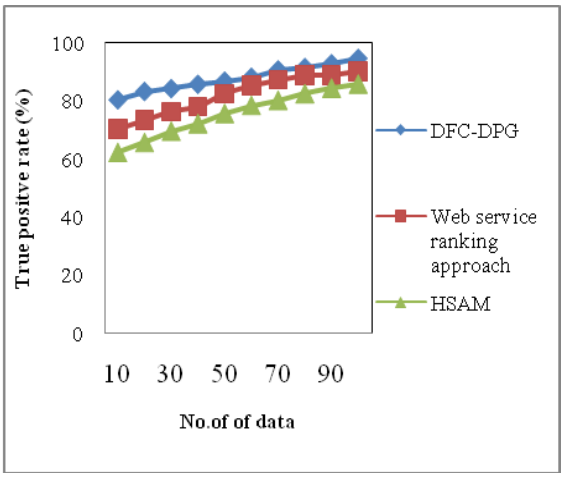

Figure 7: Measure of true positive rate

As shown in figure 7, true positive rate is measured with respect to number of data. From the figure, it is clearly evident that the proposed DFC-DPG framework correctly identified the relevant data regarding the web user. This DFC-DPG framework improves the performance of true positive rate than the other two existing methods. This is because, the URLs visited by the web user and the date, time of the visits are collected through server log files. After that, Dominance ranking algorithm is applied for identifying the relevant data and irrelevant data regarding the web user. The Spearman rank correlation is applied in Dominance ranking model to identify the degree of association between the information regarding the web user. If the correlation coefficient value provides positive correlation, then the data are related to web user. Otherwise, the data are irrelevant to the web user. Therefore, the data which are relevant data is correctly identified in DFC-DPG framework regarding the web user is selected and removes the irrelevant data regarding the web user. Therefore, the true positive rate is considerably increased by 7% and 16% compared to existing Web service ranking approach [1] and HSAM [2] respectively.

Impact of clustering efficiency

Clustering efficiency is defined as the ratio of the number of web pages are correctly clustered to the total number of web pages based on the web user query. It is measured in terms of percentage (%). The formula for clustering efficiency is defined as follows,

$$Clustering\ efficiecny = \frac{No.of\ webpages\ clustered}{total\ No.of\ web\ pages} * 100 \qquad (12)$$

Table 3: Tabulation for clustering efficiency

No. of user queries	Clustering efficiency (%)		
	DFC-DPG	Web service ranking approach	HSAM
5	89.36	72.68	63.47
10	90.52	73.12	64.58
15	91.65	75.68	66.89
20	92.54	78.14	70.20
25	93.48	80.36	73.65
30	94.25	82.65	75.46
35	95.68	85.97	78.65
40	96.35	88.46	82.65
45	97.12	89.32	85.98
50	98.65	93.65	88.96

Table 3 describes clustering efficiency using number of web user queries which is varied from 5 to 50. Therefore, the clustering efficiency is considerably improved in proposed DFC-DPG framework than the existing Web service ranking approach [1] and HSAM [2] respectively.

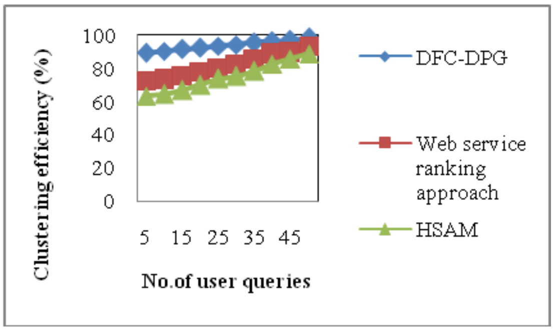

Figure 8: Measure of clustering efficiency

Figure 8 shows the performance analysis of clustering efficiency with number of web user queries. From the results, the proposed DFC-DPG framework provides better performance in terms of clustering efficiency when compared to two other existing methods. This is because with the application of fuzzy clustering algorithm. Fuzzy clustering is applied on the relevant data for grouping the user who are visited the related web pages based on the user queries. In fuzzy clustering, the centroid of cluster and fuzzy membership of data is measured with Euclidean distance. Therefore, the data points are clustered based on minimum distance between them. This process is repeated until all the data points are grouped. As a result, clustering efficiency is significantly improved by 15% and 26% compared to existing Web service ranking approach [1] and HSAM [2] respectively.

Impact of latency

Latency is defined as the amount of time required for responding the user query from the web. It is measured in terms of milliseconds (ms). The formula for latency is measured as follows,

$$Latency = No.of\ \ user\ queries * Time\ (resposing\ the\ user\ query) \quad (13)$$

Table 4: Tabulation for latency

Number of user queries	Latency (ms)		
	DFC-DPG	Web service ranking approach	HSAM
5	12.65	15.32	18.44
10	15.45	20.36	22.30
15	18.32	22.54	26.65
20	20.10	26.65	31.12
25	22.65	31.10	35.85
30	25.12	35.85	40.65
35	27.65	38.65	42.32
40	28.65	42.12	47.52
45	32.10	45.65	50.54
50	35.47	48.21	52.31

Table 4 describes the performance analysis of latency for different number of user queries. From the table, that the proposed DFC-DPG framework achieves better performance with minimal latency than the other existing Web service ranking approach [1] and HSAM [2] respectively.

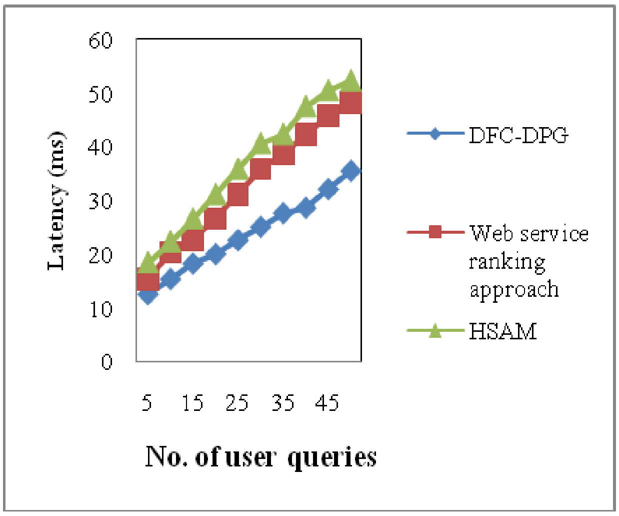

Figure 9: Measure of Latency

Figure-9: describes the performance analysis of Latency with no.of user queries. While increasing the no.of user queries, the latency of the responding time is increased. But comparatively, the responding time of web is significantly reduced using proposed DFC-DPG framework. This is because, a Distributed Probability Graph Arc (DPG) model is applied in DFC-DPG framework.

A DPG model performed based on web user' session is changed into the graph format which is represented as session density. The history of the web user activities provides the relationship among the web pages. The graph arc model contains vertex and arc to show the web user activities.

When the user request the query to web, the responding time of web is considerably reduced by predicting the web page accessed by the user and it is pre fetched to reduce the latency.

Therefore, the latency is considerably reduced by 26% and 35% compared to existing Web service ranking approach [1] and HSAM [2] respectively.

Impact of cache utilization

Cache utilization is referred as amount of memory used for storing the similar web pages in server log files based on the user queries. It is measured in terms of kilo bytes (MB).

$$CU = n * Memory\ (storing\ the\ similar\ web\ pages) \quad (14)$$

From (14), CU represents Cache Utilization and 'n' is number of similar web pages.

Table 5: Tabulation for Cache utilization

No. of user queries	Cache utilization (MB)		
	DFC-DPG	Web service ranking approach	HSAM
5	11.32	13.36	15.32
10	13.25	16.54	18.69
15	15.65	21.10	25.36
20	20.32	25.69	31.21
25	22.58	28.36	36.54
30	25.80	33.65	42.45
35	32.65	39.56	48.65
40	36.75	42.12	51.36
45	38.12	47.69	55.32
50	41.65	51.36	58.21

Table 5 describes the cache utilization for storing similar web pages. The number of web pages regarding the user queries are stored in cache is reduced in proposed DFC-DPG framework. Therefore, the proposed DFC-DPG framework significantly improves the performance results with less cache utilization than the state-of-the-art methods.

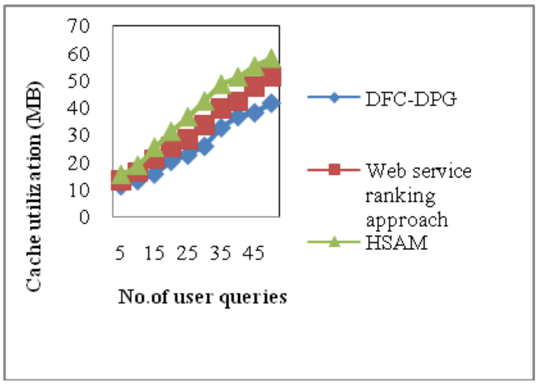

Figure 10: Measure of Cache utilization

Figure 10 illustrate the measure of cache utilization with respect to no.of user queries. The proposed DFC-DPG framework performs relatively well when compared to two other methods existing methods. The rate of accuracy using the proposed DFC-DPG framework is increased with the application of Distributed Probability Graph Arc (DPG) model that efficiently analyze the user with the visited pages sequentially and the related web pages about the web user are stored in memory. The cache utilization is significantly minimizes the memory while maintaining the user behavior profile. Therefore, cache utilization is reduced by 19% and 33% compared to existing Web service ranking approach [1] and HSAM [2] respectively.

As a result Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework improves the performance of clustering efficiency, true positive rate and minimum latency, cache utilization.

RELATED WORKS

A new technique was developed in [11] to effectively offer improved Web-page recommendation by semantic-enhancement through Web usage information of a website. However, the relevant information extraction was not performed effectively using Web-page recommendation system. The DFC-DPG framework is significantly extracting the relevant information regarding the web user of a website.

A hybrid approach Apriori algorithm and Decision tree was developed in [12] to get HTML and XML contents from a web page. But it was not improved the efficiency of this hybrid approach. The DFC-DPG framework effectively improves the clustering efficiency through fuzzy clustering.

Apriori algorithm and frequent-pattern tree was designed in [13] for extracting usage patterns electronic commerce and also improving the quality of Internet information services to end users. However, clustering on grouping the sessions was not performed based on the user visit the pages of a web site. The DFC-DPG framework perform efficient clustering to partitions the sessions.

In [14], a new approach was developed to retrieve information from a given e-commerce website and gathering the data from the website's structure in specified locations. However, the website's users were not predicted effectively. The DFC-DPG framework web pages are predicted effectively using Distributed Probability Graph Arc (DPG) model.

An ant colony optimization-based algorithm was introduced in [15] for predicting the web usage patterns. However, the effectiveness when collecting information was not improved to predict real user behavior. The DFC-DPG framework performs web user information collection with the help of server log files.

An effective particle swarm optimization mining algorithm was developed in [16] using feedback model of user to offer records of best-matching WebPages for each user. But, it failed to examine the possibility of getting optimal solution. The DFC-DPG framework improves the performance analysis of web usage mining.

The Hybrid model was introduced in [17] using Markov model and Hidden Markov Model to provide the user records of web pages of their interest. However, it takes more response time for prediction. The DFC-DPG framework effectively reduces the response time (i.e. latency) by using Distributed Probability Graph Arc model.

Markov model and all-Kth Markov model was introduced in [18] for estimating the web user Behavior. But it failed to use other features in the session's logs by extracting the features to improve prediction accuracy. The DFC-DPG framework effectively uses the session's logs to perform clustering for improving the accuracy.

WebBluegillRecom-annealing dynamic recommender system was developed in [19] using simulated annealing and swarm intelligence for detecting the interesting data recommended for the users. However, it failed to track user profiles effectively. The DFC-DPG framework effectively tracks the user profile who visited the similar kind of pages in web.

Association Rule Mining was introduced in [20] for mining the web user interesting relations behavior. But, the latency of the responding time was not analyzed. The DFC-DPG framework takes minimum amount of time to respond the user requested query.

As a result, the proposed DFC-DPG framework is capable of obtaining the interests of user while navigating through web sites.

CONCLUSION

An efficient framework called Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) is developed to improve the analysis of web user. The DFC-DPG framework uses four processes to perform the web user behavior analysis. At first, the web user data is collected from web server log file based on the web user queries. Then the collected information is separated in the form of relevant and irrelevant data. This separation is achieved by correlation measure through Dominance Rank model. After that, Fuzzy clustering is performed on the extracted relevant data for grouping the data with similar user interests from web usage data available in log files. Finally, Distributed Probability Graph Arc (DPG) model is applied to transform the web user' session into the graph. This helps to reduce the latency for the user by predicted page is pre-fetched. Experimental evaluation is carried out with the parameters such as true positive rate, clustering efficiency, latency and cache utilization with ECommerce Web Logs. The result shows that the DFC-DPG framework improves true positive rate, clustering efficiency with minimum latency and cache utilization than the state-of-the-art methods.

REFERENCES

[1] Guosheng Kang , Jianxun Liu, Mingdong Tang , Buqing Cao , Yu Xu, "An Effective Web Service Ranking Method via Exploring User Behavior", IEEE Transactions on Network and Service Management , Volume 12, Issue 4, , 2015 , Pages 554 - 564

[2] G. Poornalatha, S. Raghavendra Prakash," Web sessions clustering using hybrid sequence alignment measure (HSAM)", Social Network Analysis and Mining, Springer, Volume 3, 2013, Pages 257–268

[3] Marios Belk, Efi Papatheocharous, Panagiotis Germanakos, George Samaras," Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques", The Journal of Systems and Software, Elsevier, Apr 2013

[4] Tania Cerquitelli , Antonio Servetti, Enrico Masala, "Discovering users with similar internet access performance through cluster analysis", Expert Systems With Applications, Elsevier, Volume 64,2016, Pages 536–548

[5] Vedpriya Dongre, and Jagdish Raikwal, "An Improved User Browsing Behavior Prediction Using Web Log Analysis", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May 2015, Pages 1838- 1842

[6] S.Padmaja and Ananthi Sheshasaayee, "Clustering of User Behavior based on Web Log data using Improved K-Means Clustering Algorithm", International Journal of Engineering and Technology (IJET), Volume 8, Issue 1, 2016, Pages 305-310

[7] G. Shivaprasad N. V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining", Procedia Computer Science, Elsevier, Volume 54, 2015, Pages 327 – 334

[8] Prajyoti Lopes and Bidisha Roy, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users", Procedia Computer Science, Volume 45, 2015, Pages 60-69

[9] Neha Goel, C.K. Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool", International Journal of Computer Applications, Volume 62, Issue 2, 2013, Pages 29-33

[10] Salah Sleibi Al-Rawi, Rabah N. Farhan, Wesam I. Hajim, "Enhancing Semantic Search Engine by Using Fuzzy Logic in Web Mining", Advances in Computing, Volume 3, Issue 1, 2013, Pages 1-10

[11] Thi Thanh Sang Nguyen , Hai Yan Lu , Jie Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 10, 2014, Pages 2574 - 2587

[12] Rupinder Kaur, Kamaljit Kaur, “An Improved Web Mining Technique to Fetch Web Data Using Apriori and Decision Tree”, International Journal of Science and Research (IJSR), Volume 3 Issue 6, 2014, Pages 2094-2098

[13] Rahul Mishra, Abha choubey, Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data”, International Journal of Computer Science and Information Technologies, Volume 3, Issue 4, 2012, Pages 4662 – 4665

[14] Joao Pedro Dias, Hugo Sereno Ferreira, “Automating the Extraction of Static Content and Dynamic Behavior from e-Commerce Websites”, Procedia Computer Science, Elsevier, Volume 109, 2017, Pages 297-304

[15] Pablo Loyola n, Pablo E.Roma´ n, JuanD.Velasquez, “Predicting web user behavior using learning-based ant colony optimization”, Engineering Applications of Artificial Intelligence , Volume 25, 2012, Pages 889–897

[16] Lu Dai,WeiWang, and Wanneng Shu, “An Efficient Web Usage Mining Approach Using Chaos Optimization and Particle Swarm Optimization Algorithm Based on Optimal Feedback Model”, Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2013, August 2013, Pages 1-8

[17] Meera Narvekar and , Shaikh Sakina Banu, “Predicting User’s Web Navigation Behavior Using Hybrid Approach”, Procedia Computer Science, Elsevier, Volume 45 , 2015 , Pages 3 – 12

[18] Mamoun A. Awad and Issa Khalil, “Prediction of User's Web-Browsing Behavior: Application of Markov Model”, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Volume 42, Issue 4, 2012, Pages 1131 – 1142

[19] Anna Alphy and S. Prabakaran, “A Dynamic Recommender System for Improved Web Usage Mining and CRM Using Swarm Intelligence”, The Scientific World Journal, Hindawi Publishing Corporation, Volume 2015, April 2015, Pages 1-16

[20] Amit Dipchandji Kasliwal and Girish S. Katkar, “Web Usage mining for Predicting User Access Behaviour”, International Journal of Computer Science and Information Technologies, Volume 6 , Issue 1, 2015, Pages 201-204