International Journal of Engineering & Technology, 7 (3.27) (2018) 290-302

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Web Page Recommendation System Using Laplace Correction Dependent Probability and Chronological Dragonfly-Based Clustering

Jyoti Narayan Jadhav[1*], B. Arunkumar[2]

[1]*Research Scholar, Department of Computer Science and Engineering, Karpagam Academy of Higher Education, Karpagam University, Eachanari Post, Coimbatore, India.*

[2]*Assistant Professor, Department of CSE, Karpagam Academy of Higher Education, Coimbatore, India.*

E-mail:arunkumar.oct06@gmail.com

**Corresponding author E-mail:jnjadhav5@gmail.com*

Abstract

The web page recommenders predict and recommend the web pages to the users based on the behavior of their search history. The web page recommender system analyzes the semantics of the navigation by the user and predicts the related web pages for the user. Various recommender systems have been developed in the literature for the web page recommendation. In the first work, a web page recommendation system was developed using weighted sequential pattern mining and Wu and Li Index Fuzzy clustering (WLI-FC) algorithm. In this work, the Chronological based Dragonfly Algorithm (Chronological-DA) is proposed for recommending the webpage to the users. The proposed Chronological-DA algorithm includes the concept of the chronological for recommending the webpage based on the history of pages visited by the users. Also, the proposed recommendation system uses the concept of Laplacian correction for defining the recommendation probability. Simulation of the proposed webpage recommendation system with the chronological-DA uses the standard CTI and the MSNBC database for the experimentation, and the experimental results prove that the proposed scheme has better values of 1, 0.964, and 0.973 for precision, recall, and F-measure respectively.

Keywords: *Web page recommendation, dragonfly algorithm, history of pages, laplacian correction, recommendation probability.*

1. Introduction

In the recent years, data mining has gained significant popularity among the researchers, as it helps in maintenance and retrieval of a large amount of data. Besides, the data mining schemes identify significant relationships between the data patterns and convert the vast dataset into observable format. Digitization has helped many companies to store the data in the internet platform, and since more sources of information are available on the internet, finding the suitable/related information to a user query is challenging. Data usage mining is a recently developed scheme to retrieve the user information from the website [1].Various search engines have been developed in past decade to identify and gather useful information from the internet. Since the internet has a lot of data volume related to different fields, it is necessary for the search engine to provide the necessary webpage to the user based on his query. The problem on the internet due to big data can be overcome with the use of webpage recommendation systems. The webpage recommendation system suggests the webpage to the user as the related stories, related books, etc. on the websites [2]. The accuracy of the webpage recommendation system depends on the effectiveness of the domain knowledge the system has gained. The World Wide Web (WWW) has the collection of webpages in several domains and fields, like education, entertainment, shopping, news sites, etc., and hence, the recommendation scheme needs to categorize the web structure before suggesting the webpage. But, due to the vast data available on the internet, it is difficult for algorithms to classify the websites without gaining proper knowledge about the websites [16].

Several recommendation systems use the user session in the log file to identify the interests of the users, but they ignore the sequential information about each user session. In work [12], the probabilistic technique, like Markov model was utilized for building the recommendation system, and it had used the user session for prediction. The probabilistic-based techniques used for building the web recommendation scheme face issues such as the requirement of prior knowledge about the swapping probability, and they also requires domain knowledge [3]. Web Usage Mining (WUM) was one of the commonly used strategies by the recommendation systems for discovering the usage patterns of the users. WUM used the web mining and the machine learning based strategies to identify the patterns and the navigational activities of the users [10]. Web mining defined the way of retrieving and analyzing the information from the source of information [18]. The web mining was also referred to as data mining in the web and it allowed the user to get access to information available on the internet, data log in the internet providing the information about the services, and the structure of the web. Web mining identified the pattern/behavior of the user browsing through the internet platform [19] [4].

Various literature schemes have explored the WUM for improving the quality of the webpage recommendation schemes. Still, the accuracy of the prediction results provided by the existing recommendation algorithms was not up to the mark, as the

recommendation schemes focused on different characteristics of the webpage [11]. Hence, literature suggested in using different combination of pattern mining techniques. Web page recommendation system could be categorized as the decision support system, as it provided the webpage based on user's need [13]. Various commercial websites gained with the use of recommendation system as they helped the customer by recommending various products. E-commerce websites, such as Amazon, Flipkart have gained significantly by installing the web recommendation system as the backend process [12].The requests from the users can be categorized as implicit and explicit preferences, and the recommendation scheme needs to take note of the type of preference from the user before recommending the webpage [17]. The recommendation system falls under three categories, namely collaborative filtering approach, content-based and hybrid recommendation scheme [14].

In this work, the recommendation system is developed by proposing Chronological Dragonfly Algorithm, which is the integration of chronological concept in Dragonfly Algorithm (DA). Clustering is performed using the proposed Chronological Dragonfly Algorithm, where the update rule is designed based on the position information of the past solutions. Moreover, the recommendation probability in the first work [22]is replaced with the Laplace correction. The overall flow of the work is stated as follows: The weblog database that contains the log files is converted into transaction database, and the Prefix Span algorithm is employed to perform the sequential pattern mining. The weighted support is used to find the weighted sequential pattern based on the page duration and the frequency. Then, the proposed Chronological Dragonfly Algorithm clusters the sequence obtained using the maximal sequential pattern constraints based on the defined similarity measure, and the optimal cluster centroids are found. Finally, based on the user query, the recommended websites are provided to the user using the Laplace correction based recommendation probability.

The major contributions of this work towards the establishment of the web page recommendation are enlisted below:

- Firstly, this work introduces the chronological-DA by including the concept of the history of WebPages used in the existing DA, and the proposed Chronological-DA clusters the maximal weighted sequential patterns.
- Secondly, the recommendation probability is designed based on the Laplacian correction, and it finds the suitable webpage in the clusters and thus, recommends the suitable webpage based on user query.

The structure of this research work is organized as follows: Section 1 introduces the webpage recommendation system and the pattern mining concept. Section 2 revisits various literary works dealt with the webpage recommendation. The webpage recommendation based on the chronological-DA and Laplacian correction based recommendation probability is explained in section 3. Section 4 discusses the simulation results achieved by the proposed webpage recommendation system with the chronological-DA and the summary of this work is given in section 5.

2. Motivation

Literature Survey

This section briefs eight works which contributed to the development of the webpage recommendation system.

D.A. Adeniyi *et al.* [1] presented the automatic web usage data mining model by developing the Really Simple Syndication (RSS) model. The scheme made use of the K-Nearest-Neighbor (KNN) classification scheme for real-time training of log information from the users. The scheme also used the matching scheme and web recommender to provide the related webpage to the user based on their query. Thi Thanh Sang Nguyen *et al.* [2] proposed the webpage recommendation model by using the domain and web usage knowledge of the various websites. The technique made use of the ontology-based domain knowledge and the automatically generated semantic network for the recommendation. Besides, the authors had proposed the conceptual prediction model for automatic prediction of the semantics. The model provided improved precision and satisfaction, but neglected the information extraction schemes.

Katarya, R. and Verma, O.P., [3] presented the recommendation scheme based on the available sequential information of the web users. They have used the Fuzzy C-means (FCM) clustering for identifying the most visited webpage by the users, and they have also estimated the weight for each webpage. Do Couto, A.B.G., and Gomes, L.F.A.M., [4] proposed the web recommendation system based on the Dominance principle. The scheme defined rule sets for webpage returned by the browser according to the user query. Also, they had made use of several web mining strategies to identify the useful patterns from the log files.

Duwairi, R and Ammari, H [5] presented the modified version of the enhanced Cluster-based Association Rule Mining (CBAR) algorithm for recommending the webpage. The technique performed the recommendation online and hence, allowed dynamic updates. The technique removed the transaction details for training and hence, resulted in improvement of precision and recall, but not suitable for very large databases. Li, H., *et al.* [6] presented the optimized classification algorithm for classifying the webpage, and it was used for recommending the webpage. They have also developed the weight estimation algorithm based on the depth and breadth strategy for estimating the weight of the webpage. Though it is having improved accuracy, the algorithm suffered from class drift issues.

Manohar, E and Punithavathani, D.S., [7] presented the webpage recommendation for identifying suitable webpage, and the proposed scheme used the techniques such as weblog, web ranking, web rating and web review techniques. The scheme recommends the webpage to the user based on the success rate defined for each webpage. This work identifies the success rate of the webpage through the normalization but has reduced recommendation accuracy since the discovery knowledge about each webpage was low. Zhang, S *et al.* [8] presented the recommendation system for the users using the micro-blog. The scheme used the user relationship model to identify the interests of the user. Also, they have used the computing user authority algorithm for building the user graph, and it identifies the influential users. The scheme has failed to incorporate the bigdata challenge prevailing in the webpage recommendation.

Challenges

Various challenges involved in the design of the webpage recommendation system are enlisted below:

- A few challenges prevailing in the webpage recommendation system are learning the past web history of the user and identifying the domain knowledge of the website [2].
- Webpage recommendation system suffers from the 'new page problem', while the user uses the webpage not available in the recommendation block [2].
- Webpage recommendation system needs to consider the previously visited webpage by users to predict the suitable websites for the users. Also, it helps the service providers to redefine the latency, and online advertisements [15].
- One of the major challenges faced by the recommendation systems is that the web users are not aware of the credibility of the webpage suggested, as more websites pay the search engines to increase their respective ranks. Recommending this webpage may result in losing the customer satisfaction [9].

3. Proposed Methodology Web Page Recommendation System Based on Chronological-DA Algorithm

This work presents the webpage recommendation system by introducing the concept of chronology and Laplacian correction. Figure 1 presents the architecture of the proposed webpage recommendation system based on the proposed Chronological-DA algorithm. Initially, the database containing the log information of the various users is subjectedto pre-processing to create the transaction database, which provides the information about the webpage visited by the users. Then, based on the prefix scan algorithm, the information available in the transaction database is converted to sequential patterns, and thus, makes the mining process easy. Then, a weighted support is provided to the sequential patterns based on the frequency of the webpage visited by the user, and the session time. The useful webpage from the weighted pattern is retained by applying the maximal pattern constraint, producing the maximal weighted sequential pattern. The sequences are subjected to clustering based on the proposed Chronological-DA algorithm. When the query arrives from the user, the similarity measure is calculatedbetween the query and the cluster, and the clusterproviding the maximal similarity is considered as the optimal cluster to which the query belongs. Finally, the webpage is recommended to the user by calculating the recommendation probability between the optimal centroid and the user query. The sequence providing the maximum recommendation probability is recommended to the user.

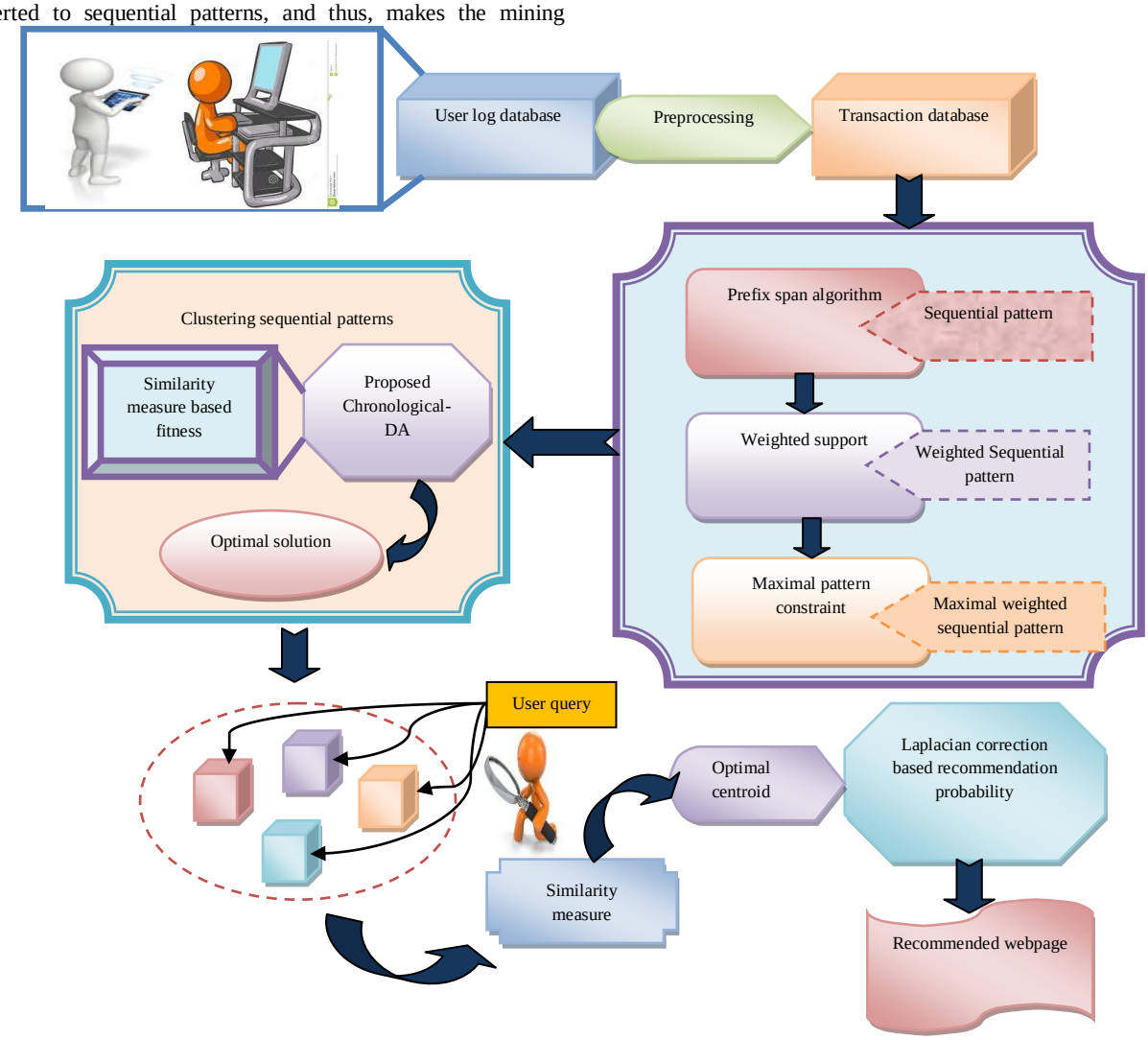


Figure 1: Architecture of the proposed webpage recommendation system based on chronological-DA algorithm

Construction of Maximal Weighted Sequential Pattern

Consider the log database L having the log information of various users, providing the information about the webpage visited, session time, data type, etc.

Preprocessing of log files using the transaction database

The log file database has many unwanted and redundant information, and thus, to gather the useful information from the database, preprocessing of database is done. The log file in the database requires pre-processing since only certain information from the database is required for constructing the webpage recommendation system. Thus, the pre-processing is done to gather the data, such as webpage visited and the session time spent the user in each webpage. The users in the database can be represented as,

$$J = \{J_1, J_2, J_3, \mathrm{K}, J_m, \mathrm{K}, J_n\}; \qquad 1 \le m \le n \qquad (1)$$

where, J_m refers to the m^{th} user in the transaction database. Pre-processing of the database provides the information about the webpage visited by the user. The user in the transaction database visits more than one webpage, and it can be represented in the following expression,

$$H = \{H_1, H_2, H_3, \mathrm{K}, H_k \mathrm{K}, H_l\}; \quad 1 \le k \le l \quad (2)$$

where, H_k refers to the k^{th} webpage visited by the users, and the value of k varies from 1 to l. The transaction database resulted through the pre-processing of log files provides the information about the pattern of webpage visited by the users, and thus, can be represented as follows,

$$P = \{p_{mk}\} \quad (3)$$

where, P indicates the transaction database, and the term p_{mk} is the data value in the transaction database providing the relation between the users and webpages visited, and thus it provides the transaction information for the users. The webpage visited by the user varies between $1 \le k \le l$. Hence, the database provides the log information of l number of web pages. The webpage visited by the users vary from one user to another, and hence, the information session time of each page visited by the user need to be noted. The session time can be obtained from the session log in and log out time. Thus, for each transaction information available in the transaction database, session time data is calculated and it is expressed as follows,

$$S = \{s_{mk}\} \quad (4)$$

where, S indicates the session time database and the term s_{mk} indicates the time spent by the m^{th} user on the k^{th} webpage. Formulating the session time database makes the webpage recommendation system to be more compactable, as they provide the interest pages of the users.

Sequential pattern mining based on the prefix span algorithm

The transaction database constructed through preprocessing can be made suitable for building the webpage recommendation system by incorporating the prefixspan algorithm [25]. Applying the prefix span algorithm to the transaction database produces the sequential patterns for mining the webpage. Thus, both the webpage visited by the user and the session time available in the transaction database are applied as input to the prefix span algorithm. As the patterns visited by the database grow, for the online session, the prefixspan algorithm constructs a projected database for every change in the pattern. The sequential pattern identified by the prefixspan algorithm is expressed as follows,

$$M_j = \langle M_1, M_2, M_3, \mathrm{K}, M_j, \mathrm{K}, M_k \rangle; \quad 1 \le j \le k \quad (5)$$

where, M_j indicates the j^{th} sequential pattern and it has k subsequences. The subsequences in the sequential patterns is considered to be a subset of webpage and it is represented as $M_j \subseteq H$. The sequential patterns construction for a sequence $g = \langle g_1, g_2, g_3, \mathrm{K}, g_a \rangle$ and another sequence $h = \langle h_1, h_2, h_3, \mathrm{K}, h_b \rangle$ can be related as follows: the sequence g is the subsequence of h if $a \subseteq b$ and $a \le b$ and $g_a \subseteq h_b$. The sequence provided by the prefix span algorithm is expressed in the following equation,

$$M = \{M_1, M_2, M_3, \mathrm{K}, M_c \mathrm{K}, M_d\}; \quad 1 \le c \le d \quad (6)$$

where, M_c indicates the subsequence in the sequence M and the maximum length of the sequential pattern is considered to be d. The expression for the subsequence is given by the following equation,

$$M_c = \langle M_c^1, M_c^2, M_c^3, \mathrm{K}, M_c^r \rangle \quad (7)$$

where, M_c^r indicates the r^{th} webpage pattern in the subsequence.

Constructing weighted support for sequential patterns

The sequential patterns obtained through the prefix span algorithm require a projected database for every change or growth in the sequential pattern. Hence, the complexity of the process is high, and this is avoided using the weighted support for the sequential patterns. The weighted support created for the sequential patterns is the combinations of the frequency of occurrence of the patterns and the time spent by the user for a particular sequential pattern. The weighted support designed for the sequential pattern is explained as follows,

$$L(M_c) = \frac{1}{2}\left\{\frac{f(M_c)}{R} + \frac{1}{|M_c|}\sum_{m=1}^{|M_c|} T_c^m\right\} \quad (8)$$

where, $L(M_c)$ defines the weighted support for the sequential pattern M_c, R indicates the total number of transactions, $f(M_c)$ refers to the frequency of subsequence occurring in the sequence M_c, T_c^m indicates the time spent by the m^{th} user and the expression for T_c^m is expressed as follows,

$$T_c^m = \frac{\sum_{c=1}^{R} T_c . f(M_c)}{\sum_{c=1}^{R} T_c^m . f(M_c)} \quad (9)$$

The expression for the frequency of the subsequence $f(M_c)$ is given as follows,

$$f(M_c) = \begin{cases} 1; & \text{if subsequence } M_c \text{ is present in } M \\ 0; & \text{else} \end{cases} \quad (10)$$

Here, the frequency gets the maximum value 1, when the subsequence M_c is in M: orelse it has the value as 0. The weight induced in the sequence increases the length of the sequence, and thus, it can be reduced by introducing a threshold e. Here, the threshold is chosen to be 0.7, and sequencesproviding the weights less than the threshold are retained. Finally, the sequences extracted by applying the weighted support is expressed as follows,

$$M_U << \{L(M_c) > e\} \quad (11)$$

where, M_U refers to the weighted sequences, the term $L(M_c)$ indicates the weighted support applied to the sequence M_c, and the term e indicates the threshold for obtaining the sequences.

Constructing maximal pattern constraint

Here, the maximal pattern constraint is applied to the sequential patterns for finding the most suitable webpage patterns from the maximal patterns, and this can be achieved through applying the maximal pattern constraint [26]. The maximal pattern constraint reduces the complexity of the webpage recommendation system as it finds the set of most visited webpage by the user. Also, the maximal sequential patterns retrieved by the user finds the pattern with the maximum number of Web pages. The maximal sequence can be obtained from the set of sequence based on the frequency of occurrence of the sequences. Thus, the subsequence M_c can be declared to the maximal subsequence, if the frequency of occurrence of subsequence f has the same value as another subsequence. The sequential patterns found by the application of the maximal pattern constraint is expressed by the following expression,

$$W_V = \{W_1, W_2, W_3, \mathrm{K}, W_A, \mathrm{K}, W_B\}; \quad 1 \le A \le B \quad (12)$$

where, W_A represents the A^{th} maximal weighted sequential pattern and it varies in the range $1 \le A \le B$.

Clustering the Maximal Weighted Sequential Pattern Using the Proposed Chronological-DA

The next step involved in the design of the webpage recommendation system is developing the proposed Chronological-DA algorithm for clustering the sequential patterns obtained. Since the maximal weighted sequential patterns for every user in the log file have different characteristics, it is necessary to cluster the patterns into several groups/clusters. Here, the proposed Chronological-DA algorithm clusters the sequential patterns according to their properties.

Solution encoding

Consider the sequential patterns are clustered into G number of clusters, and the proposed chronological-DA algorithm identifies the optimal centroid for each cluster, and thereby, find the suitable patterns to be grouped in each cluster. Figure 2 presents the solution encoding of the proposed Chronological-DA algorithm.

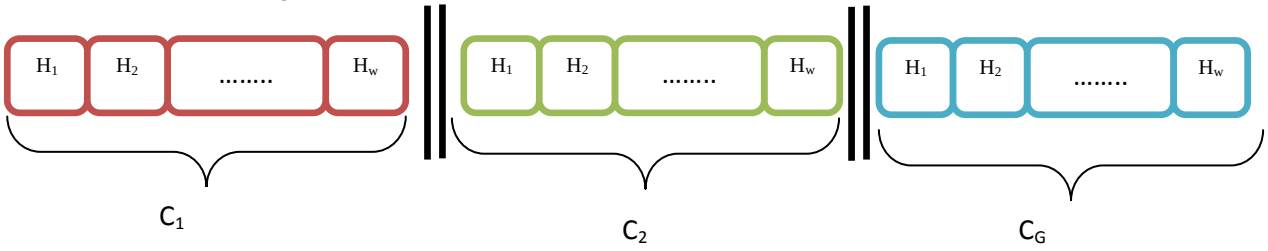


Figure 2: Representation of solution vector of chronological-DA

Fitness for clustering the maximal weighted sequential pattern

The proposed Chronological-DA algorithm uses the maximal fitness function as defined in the following expression,

$$\text{Fitness} = \sum_{o=1}^{C_G} \sum_{A=1}^{B} \text{Similarity}\left(W_A, C_o\right) \quad (13)$$

where, $\text{Similarity}\left(P_j, K_i\right)$ indicates the similarity measure between the sequence and the centroid. It is necessary that the proposed Chronological-DA needs to assign the sequences with the maximal similarity measure to the group. Thus, the fitness measure includes the similarity measure between the sequence and the centroid chosen for the analysis.

Algorithmic description of the proposed chronological-DA algorithm

The proposed Chronological-DA algorithm aims to identify the suitable web patterns for each cluster. Thus, for clustering the maximal sequential pattern based on the fitness measure, the proposed Chronological-DA includes the past solutions in the present iteration for achieving the optimal results. This makes the webpage in the cluster more related to the history of the web pages visited by the users. The proposed Chronological-DA includes the properties of existing DA [20], which identifies the optimal solution based on the behavior of the dragonflies. The steps involved in the proposed Chronological-DA algorithm are explained as follows:

1. **Initialization of population:** DA considers the solution as the position of dragonflies, and hence, initially, the position of the dragonflies is randomly chosen. As the sequential patterns need to be clustered, the maximal weighted sequential patterns are randomly chosen as the position of dragonflies, and it is expressed as

$$D = \{D_1, D_2, \mathrm{K}, D_i \mathrm{K}, D_N\} \quad (14)$$

where, D_i refers to the i^{th} solution and the maximum size of the solution vector is N.

2. **Fitness evaluation:** The optimization procedure identifies the best solution based on the defined fitness measure as mentioned in equation (13). In this step, the fitness of the randomly assigned solutions is identified.
3. **Update the position based on DA:** The existing DA defines the optimization procedure as the behavior of the dragonfly. The dragonfly alters its position during the search of food, enemy position, etc. The various factors influencing the position of the dragonfly as mentioned in DA are separation, alignment, cohesion, attraction to food, and distraction due to the enemy. Thus, the expression for position update of solutions based on DA is expressed as follows,

$$D(t+1) = D(t) + \Delta D(t+1) \quad (15)$$

where, $D(t)$ refers to the position of the dragonfly at the iteration t, and $\Delta D(t+1)$ refers to the change in position, expressed as,

$$\Delta D(t+1) = \left(z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i\right) + K.\Delta D(t) \quad (16)$$

where, the terms z_1, z_2, z_3, z_4, and z_5 are the weights corresponding to the separation, alignment, cohesion, attraction, and distraction, respectively, and the terms u_i, v_i, w_i, x_i, and y_i refers to the separation, alignment, cohesion, attraction, and distraction of the i^{th} solution. The expression of each factorisgiven below,

$$u_i = -\sum_{j=1}^{N} D - D_j \tag{17}$$

where, D_j indicates the position of the j^{th} neighbor solution.

$$v_i = \frac{\sum_{j=1}^{N} q_j}{N} \tag{18}$$

where, q_j indicates the velocity of the j^{th} neighbor solution.

$$w_i = \frac{\sum_{j=1}^{N} D_j}{N} - D \tag{19}$$

$$x_i = D^+ - D \tag{20}$$

$$y_i = D^- - D \tag{21}$$

where, D^+ and D^- indicate the attraction and the distraction of the position of the i^{th} solution due to the food and the enemy sources. Finally, the updated solution based on the DA is specified as,

$$D(t+1) = D(t) + (z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.\Delta D(t) \tag{22}$$

4. **Find the chronological solutions and update the position:** In the proposed Chronological-DA, the position attained by the DA in the past iteration $(t-1)$ is considered for the analysis, and is expressed as,

$$D(t) = D(t-1) + (z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.\Delta D(t-1) \tag{23}$$

where, $D(t-1)_{\text{indicates}}$ the position of the solution at the iteration $(t-1)$. Now, replace $D(t)$ in equation (22) with the equation (23) to get the required position of solution, as expressed below,

$$D(t+1) = D(t-1) + (z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.\Delta D(t-1) + (z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.\Delta D(t) \tag{24}$$

Rearranging the above equation yields the position update based on DA with the chronological solutions and it is specified as,

$$D(t+1) = D(t-1) + 2(z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.[\Delta D(t-1) + \Delta D(t)] \tag{25}$$

5. **Update the position based on chronological-DA:** The proposed Chronological-DA provides equal weightages to the solution update from the DA algorithm as expressed in (22), and modified solution of DA with the chronological solutions as expressed in (25). Thus, the solution update based on the proposed Chronological-DA depends on the average of the equations specified in (22) and (25),respectively, and it is expressed as,

$$D(t+1) = \frac{D(t+1) + D(t+1)}{2} \tag{26}$$

Now, substitute the equations (22) and (25) in equation (26) to obtain the required solution update based on proposed Chronological-DA.

$$D(t+1) = \frac{1}{2}\left[\begin{matrix} D(t) + (z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.\Delta D(t) + \\ D(t-1) + 2(z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + K.[\Delta D(t-1) + \Delta D(t)] \end{matrix}\right] \tag{27}$$

Rearranging the above equation yields the solution update based on proposed Chronological-DA, and it is given as,

$$D(t+1) = \frac{1}{2}\left[\begin{matrix} D(t) + D(t-1) + 3(z_1.u_i + z_2.v_i + z_3.w_i + z_4.x_i + z_5.y_i) + \\ K.[\Delta D(t-1) + 2\Delta D(t)] \end{matrix}\right] \tag{28}$$

6. **Evaluate the fitness of solution and find the best solution:** Here, the fitness of the updated solution is identified, and the solution with high fitness is retained as the best solution at the end of the iteration.
7. **Termination:** The position of the solution gets updated based on proposed Chronological-DA for an increase in the iteration t, and at the end of the iteration T_{max}, the algorithm gets terminated.

Recommendation of Webpage Based on User Query

The clusters identified from the proposed Chronological-DA makes the recommendation process easier, since the similarity measure used in this work compares the user query with the clusters and identifies the optimal cluster for recommending the webpage. Figure 3 presents the recommendation scheme along with the proposed laplacian correction based recommendation probability.

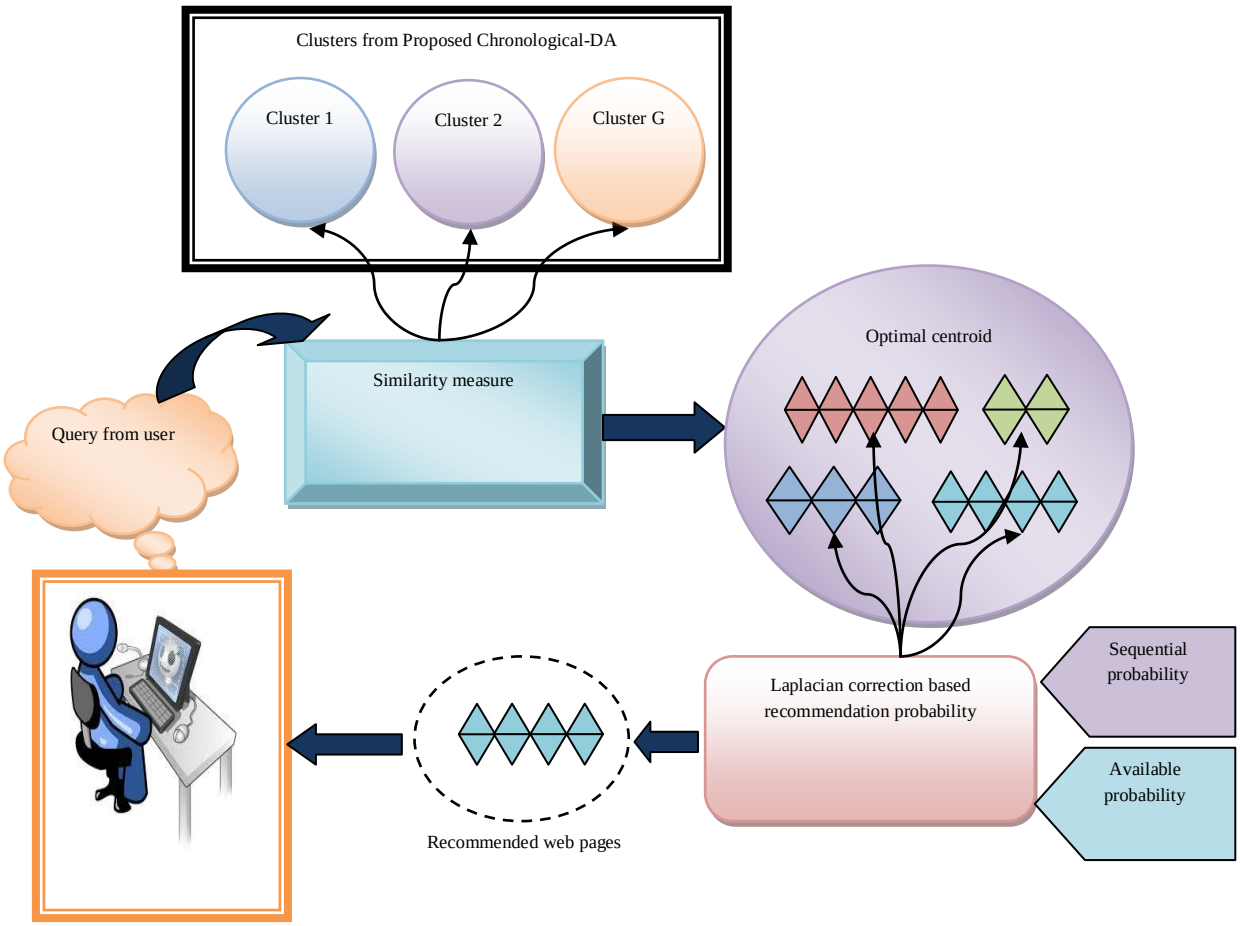


Figure 3: Recommending webpage for user query based on laplacian correction based recommendation probability

Identifying optimal centroid based on user query

The similarity between the user query and the sequential patterns in the clusters are calculated, and the clusterproviding the maximum similarity measure is considered as the optimal cluster regarding the user query. Consider the user query Q from the user, and the similarity measure identifies the similarity between the query and each cluster. The expression for the similarity measure is expressed as follows,

$$\text{similarity}(Q, C_o) = \left[\frac{Q \cap C_o}{Q \cup C_o}\right] \tag{29}$$

And finally, the cluster having maximum similarity with the user query is retained to be optimal cluster, and it is expressed as,

$$C_O = \{H_1, H_2, H_3, K\ H_\lambda, K\ , H_N\} \tag{30}$$

Laplacian correction based recommendation probability

Here, the recommendation probability is defined based on the Laplacian correction parameter. The recommendation probability considered comprises available probability and sequential probability. The sequences attaining probability greater than the recommendation probability are finally recommended to the users. Here, the sequences present in the optimal cluster obtained from the previous step are chosen for the analysis, and then, the recommendation probability of each sequence is calculated based on the following expression,

$$P^Z(H_\lambda \| Q) = \frac{P^E + P^F}{2} \tag{31}$$

Here, expression for the available probability and sequential probability are calculated based on the Laplacian correction [21], and it is expressed as follows,

$$P^E = \frac{\text{total subsequence occurrence of query } Z \text{ with length 1 present in } H_\lambda + 1}{\text{total number of subsequence in } H_\lambda \text{ with length } 1 + \text{dom(c)}} \tag{32}$$

$$P^F = 1 - \frac{Y + 1}{\text{total number of subsequence in } H_\lambda \text{ with length } 2 + \text{dom(c)}} \tag{33}$$

where, Y indicates the total gap between the items in the subsequence in H_λ on the user query Q. If the web pages in the query are not present in the subsequence of the optimal centroid, the gap takes the maximal length of the sequence. In the final stage, the recommendation of the web page is obtained based on the proposed recommendation probability P^R. The recommendation probability of each subsequence of the optimal centroid is measured. The proposed webpage recommendation system recommends the webpage from the subsequence of the optimal centroid having the highest recommendation probability.

4. Results and Discussion

This section presents the simulation results of the webpage recommendation system based on the proposed Chronological-DA

algorithm. The results of the proposed Chronological-DA are measured against the various existing techniques for the standard databases.

Experimental Setup

Experimentation of the webpage recommendation scheme with the proposed chronological-DA is analyzed under different conditions and simulated in the JAVA tool. The experimentation requires the PC with the configuration of Windows 10 OS, Intel I3 processor, and 4 GB RAM.

Values fixed to the parameters: Cluster size G = 5, Maximum iteration T_{max} = 100.

Database description

The experimentation of the webpage recommendation system with proposed Chronological-DA has utilized the two standard databases, such as MSNBC [24] and CTI [23], and the description to these databases is given below:

1. **MSNBC dataset:** The MSNBC dataset contains the weblog information of various users visiting the pages of the MSNBC website for one day. The database provides log information of average of 989818 users and 5.7 visits for each page available in the MSNBC database. The category available in the pages has 10 to 5000 URLs.
2. **CTI dataset:** Similarly, the CTI webpage has log information of visits to the CTI site. It also includes the statistics about the user sessions.

Evaluation metrics

The evaluation of the proposed `Chronological-DA algorithm for recommending the webpage can be done through the metrics, such as precision, recall, and F-measure, and the mathematical expression for each metrics is defined below:

1. **Precision:** Precision defines the measure of the fraction of the web page that is successfully recommended based on the query.

$$\text{Precision} = \frac{|H_\tau \cap H_R|}{H_R} \tag{34}$$

where, the term H_τ expresses the relevant web page based on the user query, and the term H_R expresses the recommended web page.

2. **Recall:** Recall defines the measure of the fraction of the relevant web pages present in the recommended web page that is relevant to the user query.

$$\text{Recall} = \frac{|H_\tau \cap H_R|}{H_\tau} \tag{35}$$

3. **F-measure:**The F-measure defines the harmonic mean of the precision and the recall metric. Equation (32) expresses the F-measure metric.

$$F\ \ measure = 2*\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{36}$$

Comparative models

The proposed Chronological-DA is analyzed by determining the performance of the various comparative models, such as Fuzzy C Means (FCM) + K Nearest Neighbor (k-NN) clustering [1], FCM + recommendation probability [3], WLI+k-NN, and WLI-FC + recommendation probability [22]. Each model considered in this work for the analysis is suitable for establishing a webpage recommendation system, and description of each work is given below:

FCM + k-NN [1]: In this work both the FCM and k-NN clustering schemes have been deployed for recommending the webpage.

FCM + recommendation probability [3]: Here, the FCM model clusters the database, and designs a recommendation probability for identifying the suitable webpage for retrieval.

WLI +k-NN: Here, WLI and k-NN model are utilized for the recommendation process.

WLI-FC + recommendation probability[22]: Here, a recommendation probability is developed based on the page availability and sequential arrangement for recommending the webpage form the optimal clusters found by WLI-FC.

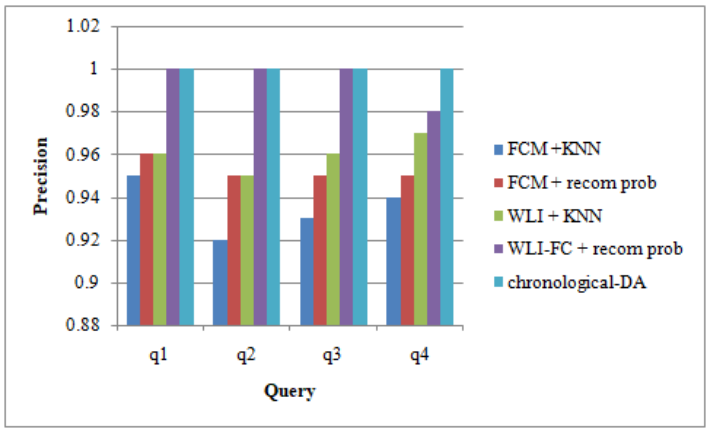
Comparative Analysis of Proposed Chronological-DA Algorithm

The performance of the proposed Chronological-DA algorithm is compared against the performance of the various existing works, such as FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability for the MSNBC and the CTI datasets. Here, the analysis is done by varying queries, and data sizes, and the performance of the comparative models are measured based on precision, recall, and f-measure.

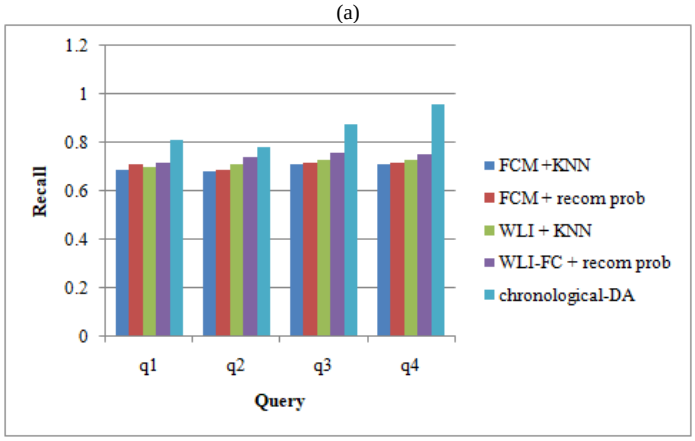
Analysis based on MSNBC database

i) Varying the user queries

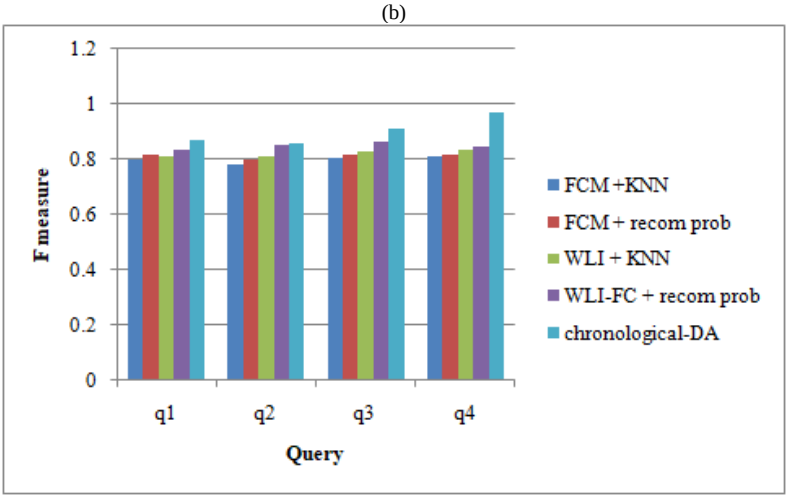
Figure 4 presents the comparative analysis of the proposed Chronological-DA algorithm for varying user queries in MSNBC database. Analysis based on precision, as depicted in figure 4.a, shows that the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability algorithms have the precision value of0.94, 0.95, 0.97, and 0.98, respectively, for the query q_4. Meanwhile, the proposed Chronological-DA algorithm has a high precision value of 1 for the same query. Similarly, analyzing the algorithms based on recall metric, as shown in figure 4.b, shows that the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability models have recall value of 0.71, 0.72, 0.73, and 0.75, respectively for q_4. It is clear from the analysis that the proposed Chronological-DA has higher recall value of 0.96 for q_4. Again, the performance based on f measure states that the proposed Chronological-DA algorithm has high f measure value of 0.968 when the user provides the query q_4, and thus, outperforms other comparative techniques.



(a)



(b)

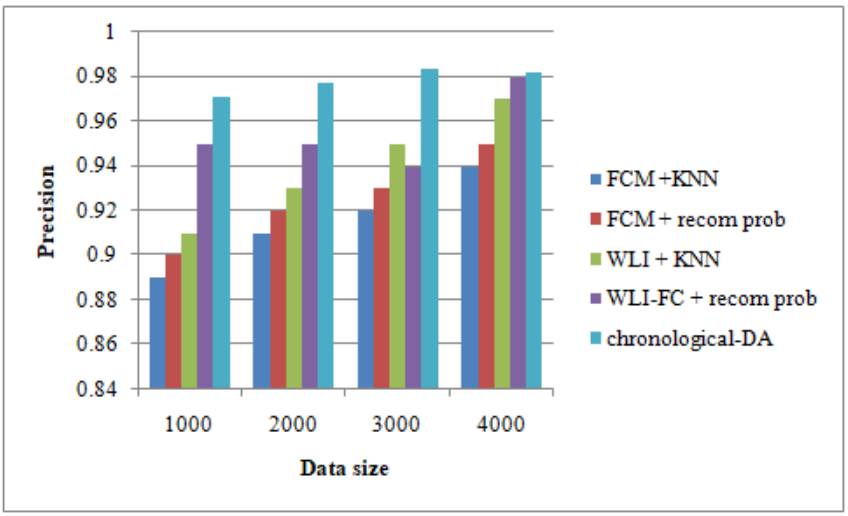


(c)

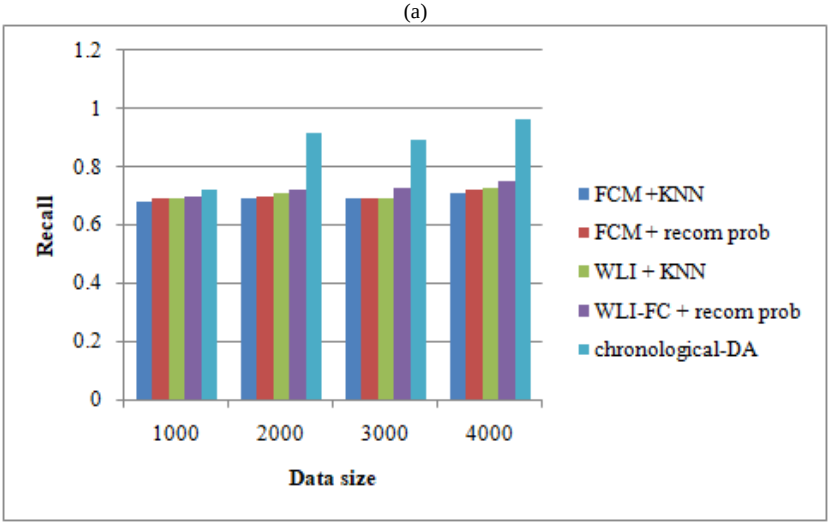
Figure 4: Comparative analysis for varying the query from the user in MSNBC database based on (a) precision, (b) recall, and (c) F measure

ii) Varying the data sizes

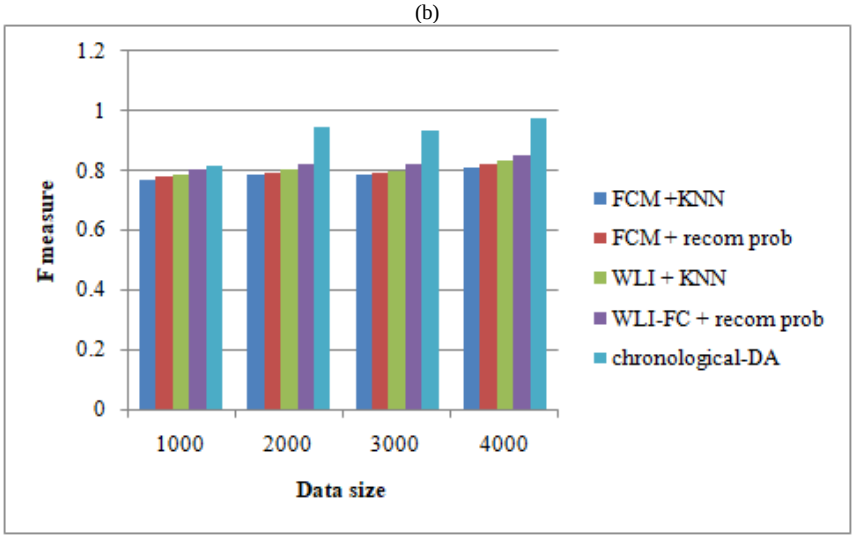
Figure 5 presents the performance of comparative models for varying data sizes based on different evaluation metrics. Precision analysis of the models, as shown in figure 5.a,demonstrates that the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC+recommendation probability techniques have precision values of 0.94, 0.95, 0.97, and 0.98, respectively, for the data size of 4000. The proposed Chronological-DA algorithm obtained improved performance over existing techniques with the precision value of 0.9816. Analysis based on recall, and f measure shown in figure 5.b, and figure 5.c depicts that the proposed Chronological-DA algorithm achieved improved values of 0.964 and 0.973,respectively for both the measures when the data size is 4000.



(a)



(b)



(c)

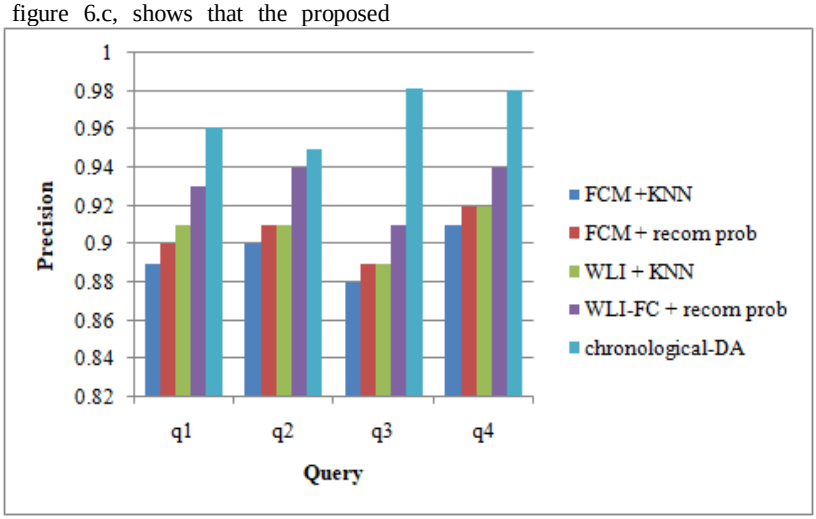
Figure 5: Comparative analysis for varying data size in MSNBC database based on (a) precision, (b) recall, and (c) F measure

Analysis based on CTI database

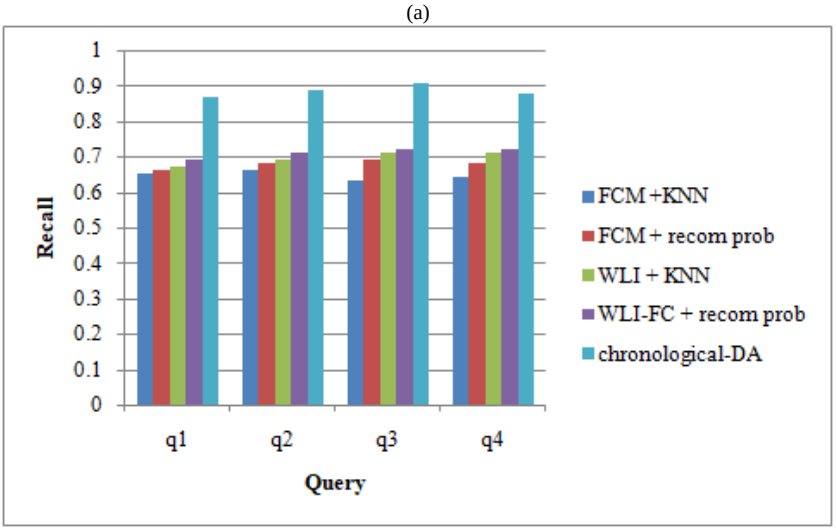
i) Varying the user queries

Figure 6 presents the comparative analysis of the proposed Chronological-DA for the CTI dataset for varying query from the users. Precision analysis, as depicted in figure 6.a, shows that the proposed Chronological-DA has obtained the improved precision value of 0.9806 then the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability models, which have achieved values of 0.91, 0.92, 0.92, and 0.94, respectively, for the query q_4. While analyzing the performance based on recall as shown in figure 6.b, the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability models have the recall value of 0.64, 0.68, 0.71, and

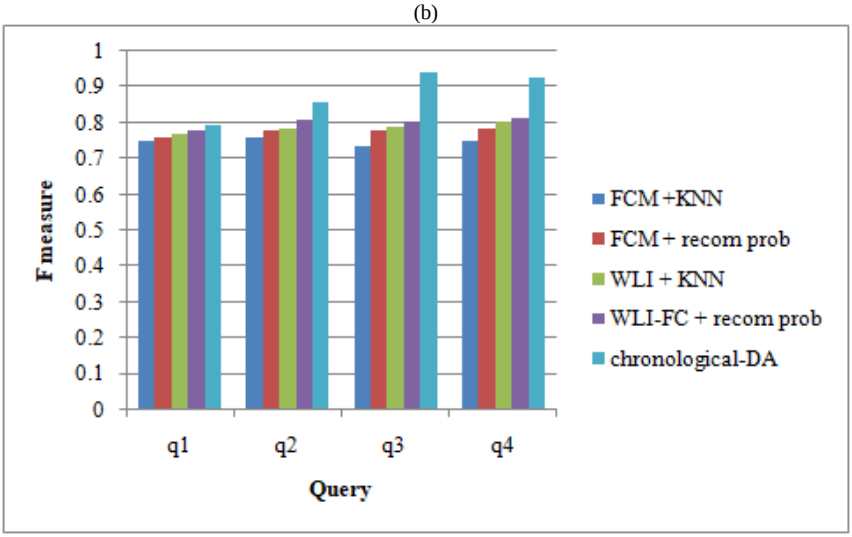
0.72, respectively, for the user query q_4. But, the proposed Chronological-DA algorithm has better recall value of 0.875 than the existing models for the same query. Analysis based on f measure, as depicted in figure 6.c, shows that the proposed Chronological-DA algorithm has improved f measure value of 0.9249 for the query q_4.



(a)



(b)



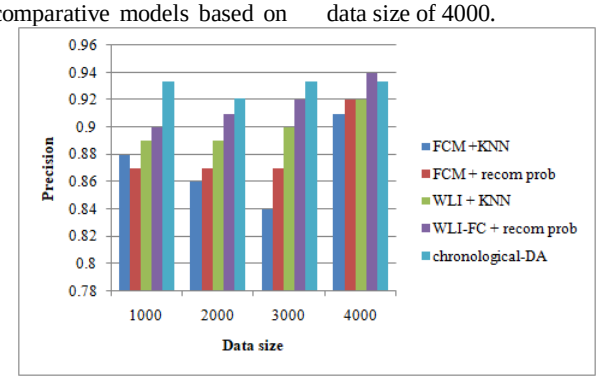
(c)

Figure 6: Comparative analysis for varying the query from the user in CTI database based on (a) precision, (b) recall, and (c) F measure

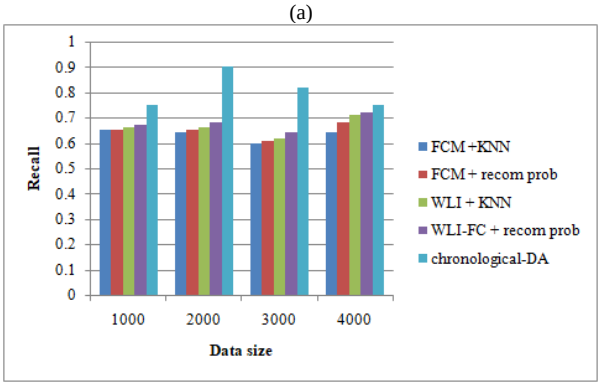
ii) Varying the data sizes

Figure 7 presents the performance of comparative models for varying data sizes of CTI database based on different evaluation metrics. Precision analysis of the models, as shown in figure 7.a, shows that the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability techniques have precision values of 0.84, 0.87, 0.9,

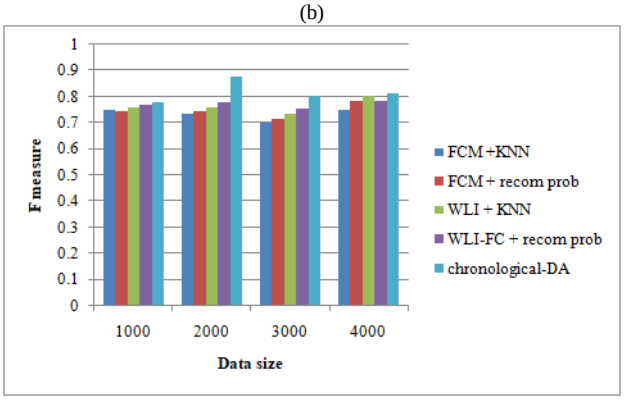
and 0.92, respectively, for the data size of 3000. The proposed Chronological-DA algorithm obtained improved performance over existing techniques with the precision value of 0.933. Analysis based recall shown in figure 7.b depicts that the proposed Chronological-DA algorithm achieved improved values of 0.9,for recall than the existing techniques for the data size 2000. Figure 7.c shows the performance of the comparative models based on the f measure metric for varying data sizes. Here, the existing FCM + k-NN, FCM + recommendation probability, WLI + k-NN, and WLI-FC + recommendation probability models have achieved the f measure value of 0.751, 0.782, 0.801, and 0.785 for the data size = 4000 respectively. The proposed Chronological-DA algorithm achieved overall better f measure value of 0.815 for the data size of 4000.



(a)



(b)



(c)

Figure 7: Comparative analysis for varying data size in CTI database based on (a) precision, (b) recall, and (c) F measure

Comparative Discussion

Here, comparative discussion of the various existing techniques is presented based on their performances against the performance of the proposed Chronological-DA algorithm. As shown in table 1, for the MSNBC database,[27] the proposed Chronological-DA algorithm achieved overall best performance with the values of 1, 0.964, and 0.973 for the precision, recall, and f measure, respectively. Similarly, for the CTI database, the proposed Chronological-DA algorithm has the values of 0.981, 0.9, and 0.943 for precision, recall, and f measure respectively. [28]

Table 1: Comparative Discussion of the Proposed Chronological-DA Algorithm

Dataset	Evaluation metric	Methods for comparison				
		FCM + k-NN	FCM + recom prob	WLI + k-NN	WLIFC + recom prob	Proposed Chronological-DA
MSNBC	Precision	0.94	0.95	0.97	0.98	1
	Recall	0.71	0.72	0.73	0.75	0.964
	F-measure	0.808	0.819	0.833	0.849	0.973
CTI	Precision	0.88	0.89	0.89	0.91	0.981
	Recall	0.64	0.65	0.66	0.68	0.9
	F-measure	0.734	0.777	0.789	0.803	0.943

5. Conclusion

This work introduces the web page recommendation system by introducing a novel optimization algorithm, namely chronological-DA. The database containing the log information is converted into

maximal pattern sequence through sequential pattern mining. Then, the patterns are clustered into different groups, where the optimal centroid of the clusters is identified with the proposed Chronological-DA algorithm. When the query is requested from the user, it is compared with the clusters and suitable cluster related to the query is identified based on the similarity measure. Finally, recommendation probability designed based on Laplacian correction parameter is used for recommending the suitable webpage from the optimal clusters. The simulation environment for experimentation of the proposed Chronological-DA based webpage recommendation system requires JAVA simulation tool, and two standard databases, namely MSNBC and CTI, have been utilized for the analysis. For the MSNBC database, the proposed Chronological-DA algorithm achieved overall best performance with the values of 1, 0.964, and 0.973 for the precision, recall, and f measure, respectively. Similarly, for the CTI database, the proposed Chronological-DA algorithm has the values of 0.981, 0.9, and 0.943 for precision, recall, and f measure, respectively.

References

[1] Adeniyi DA, Wei Z & Yongquan Y, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", *Applied Computing and Informatics*, Vol.12, No.1, (2016), pp.90–108.

[2] Nguyen TTS, Lu HY & Lu J, "Web-page recommendation based on web usage and domain knowledge", *IEEE Transactions on Knowledge and Data Engineering*, Vol.26, No.10,(2014), pp.2574-2587.

[3] Z Yesembayeva (2018). Determination of the pedagogical conditions for forming the readiness of future primary school teachers, Opción, Año 33. 475-499

[4] G Mussabekova, S Chakanova, A Boranbayeva, A Utebayeva, K Kazybaeva, K Alshynbaev (2018). Structural conceptual model of forming readiness for innovative activity of future teachers in general education school. Opción, Año 33. 217-240

[5] Katarya R & Verma OP, "An effective web page recommender system with fuzzy c-mean clustering", *Multimedia Tools and Applications*, Vol.76, No.20, (2017), pp.21481-21496.

[6] Do Couto ABG & Gomes LFAM, "Multi-criteria Web Mining with DRSA", *Procedia Computer Science*, Vol.91, (2016), pp.131-140.

[7] Duwairi R & Ammari H, "An enhanced CBAR algorithm for improving recommendation systems accuracy", *Simulation Modelling Practice and Theory*, Vol.60, (2016), pp.54-68.

[8] Li H, Xu Z, Li T, Sun G & Choo KKR, "An optimized approach for massive web page classification using entity similarity based on semantic network", *Future Generation Computer Systems*, Vol.76, (2017), pp.510-518.

[9] Manohar E & Punithavathani DS, "Hybrid Data Aggregation Technique to Categorize the Web Users to Discover Knowledge About the Web Users", *Wireless Personal Communications*, Vol.97, No.4, (2017), pp.5289–5303.

[10] Zhang S, Zhang S, Yen NY & Zhu G, "The Recommendation System of Micro-Blog Topic Based on User Clustering", *Mobile Networks and Applications*, Vol.22, No.2, (2017), pp.228-239.

[11] Serrano W & Gelenbe E, "The Random Neural Network in a Neurocomputing Application for Web Search", *Neurocomputing*, (2017).

[12] Castellano G, Fanelli AM & Torsello MA, "NEWER: A system for NEuro-fuzzy WEb Recommendation", *Applied Soft Computing*, Vol.11, No.1, (2011), pp.793-806.

[13] Göksedef M & Gündüz-Öğüdücü Ş, "Combination of Web page recommender systems", *Expert Systems with Applications*, Vol.37, No.4, (2010), pp.2911-2922.

[14] Mishra R, Kumar P & Bhasker B, "A web recommendation system considering sequential information", *Decision Support Systems*, Vol.75, (2015), pp.1-10.

[15] Wu S, Jiang M, Gao X & Wei G, "Webpage Recommender System concerning high dimensional and sparse features", *Proceedings of 8th International Conference on Information Science and Digital Content Technology*, (2012), pp.109-112.

[16] Wang C, Kalra A, Borcea C & Chen Y, "Revenue-Optimized Webpage Recommendation", *Proceedings of IEEE International Conference on Data Mining Workshop*, (2015), pp.1558-1559.

[17] Sejal D, Kamalakant T, Tejaswi V, Anvekar D, Venugopal KR, Iyengar SS & Patnaik LM, "WNPWR: Web navigation prediction framework for webpage recommendation", *Proceedings of IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, (2015), pp.120-125.

[18] Kolekar P & Wakhade S, "A novel approach to provide Web page recommendation using domain knowledge and web usage knowledge", *Proceedings of International Conference on Communication and Electronics Systems*, (2016), pp.1-5.

[19] Su AJ, Hu YC, Kuzmanovic A & Koh CK, "How to Improve Your Google Ranking: Myths and Reality", *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, (2010), pp.50-57.

[20] Chiu PH, Kao GYM & Lo CC, "Personalized blog content recommender system for mobile phone users", *International Journal of Human-Computer Studies*, Vol.68, No.8, (2010), pp.496-507.

[21] Zheng N & Li Q, "A recommender system based on tag and time information for social tagging systems", *Expert Systems with Applications*, Vol.38, No.4, (2011), pp.4575-4587.

[22] Mirjalili S, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems", *Neural Computing and Applications*, Vol.27, No.4, (2016), pp.1053–1073.

[23] Störr HP, Xu Y & Choi J, "A compact fuzzy extension of the Naive Bayesian classification algorithm", *Proceedings In Tech/VJ Fuzzy*, (2002), pp.172-177.

[24] Jadhav JN & Asaithambi M, "Web Page Recommendation System Using Weighted Sequential Pattern Mining and WLI Fuzzy Clustering", *Journal of Advanced Research in Dynamical and Control Systems*, (2017), pp.42-59.

[25] CTI dataset from facweb.cs.depaul.edu/mobasher/classes/ect584/resource.html, 2017.

[26] MSNBC dataset from https://archive.ics.uci.edu/ml/machine-learning-databases/msnbc-mld/msnbc.data.html, 2017.

[27] Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U & Hsu MC, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth", *Proceedings of the 17th international conference on data engineering*, (2001).

[28] Ren JD & Sun YF, "Interactive mining of Maximal Constrained frequent Patterns", *Database Engineering and Applications Symposium*, (2004).