Sequential Pattern Discovery method for Pattern classification in web usage Mining

G. Punithavathi[1], R. Sankarasubramanian[2]

[1]Ph.D., Research Scholar, Erode Arts and Science College, Erode-9, Tamilnadu.
[2]Principal, Erode Arts and Science College, Erode-9, Tamilnadu.

Received: 05.04.2024 Revised : 10.05.2024 Accepted: 22.05.2024

ABSTRACT

The arrival of novel technologies in our daily lives has been much emphasised compare to last decade. This has led to a multiplication of data mining tools aimed at obtaining knowledge from these data sources.Novel approach to classifying user navigation patterns and predicting users' future requests. The approach is based on the combined mining of web server logs and the contents of the retrieved web pages. The textual content of web pages is captured through extraction of characterN-grams, which are combined with Web server log files toderive user navigation profiles. The approach is implemented as an experimental system, and its performance is evaluated based on two tasks names as classification and prediction. This approach may be used to facilitatebetter web personalization and website organization. Pattern discovery from web page navigation sequences involves application of data mining and pattern classification methods to discover the patterns have subsequently accessed interesting page categories and user groups. In this paper, the proposed SPDM helps to discovering sequential patterns from web page navigation.

Keywords: WebUsage Mining, WebPage, WebServer logs, Discover Sequential Pattern weblogs.

INTRODUCTION

Web usage mining tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behaviour while the user interacts with Web. In[4] the potential strategic aims in each domain into mining goal as: prediction of the user's behaviour within the site, comparison between expected and actual Website usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories.

Web content and site topology will be employed as information sources throughout the data preparation process for web usage mining, which links web usage mining with web content and web structure mining. Furthermore, from usage mining to online content and structure mining, clustering in the pattern finding process serves as a bridge. Numerous studies in the fields of information retrieval, databases, intelligent agents, and topology have been conducted. These studies offer a strong basis for Web content and Web structure mining. Unsupervised techniques based on association rules are still disregarded in even recent surveys that discuss the application of data mining techniques in the domains of sentiment analysis and user-generated content [3]. These methods are particularly important nowadays because of their robustness when labelled data isn't available and their results are simple to understand. In the realm of big data, these issues are quite significant.

Web server log repositories, which maintain track of the various web users' usage habits, are an excellent source of information. The practice of determining browsing patterns by examining a user's navigational behavior is known as web usage pattern analysis. The web usage pattern analysis process receives its input from the web server log files, which contain visitor information. The web usage pattern analysis process receives its input from the web server log files, which contain visitor information. To enable the application of web usage mining techniques to these web logs, these log files are first pre-processed and transformed in to the necessary formats. The method of extracting valuable patterns from an academic institution's web server log file. Applications such as web traffic analysis, effective website management, site updates, system improvement and customisation, and business intelligence can all benefit from the results obtained.

RELATED WORKS

It is possible to acquire insight about framing techniques towards load balancing, data transfer and/or distribution, and other related topics by using web usage mining on web traffic and its behaviour.On the other side, online usage mining can be used to detect intrusions, frauds, breakthrough attempts, and other things in order to handle security, which is extremely important given the fast growing e- commerce and e-banking activities. This article provides a review and examination of current Web utilization mining patterns and innovations. Web use mining is used to find interesting client route designs and can be applied to numerous true issues, improving Web locales/pages, making extra point or on the other hand item proposals, client/client conduct templates, and so forth. [10]. Information gathering, information planning, route design revelation, design investigation and representation, and design applications are the five major tasks carried out by web usage mining. Every task is explained in detail, along with any innovations that are relevant [15].

A supervised machine learning technique called K- nearest neighbours(KNN)can be applied to regression and classification problems alike. In order to categorize an object into preset categories, KNN takes into account the similarity factor between new and available data. Numerous industries, including manufacturing, energy, airborne, environment, geology, maritime, geographic information systems (GIS), industry, machine engineering, health, marketing, electrical engineering, security, and manufacturing, have made extensive use of KNN.

An enhanced prefix span method has also been suggested for identifying websites that are often visited [8]. The approach is less difficult in terms of time and memory than the conventional prefix span algorithm. Studies using the MSNBC dataset at a minimum support level of 6% have led to the hypothesis that users are more likely to visit "Health" and "Summary" after watching "Weather" and "News" and "Tech" after viewing "Front page." It has also been determined that the most popular web sites are "Front-Page,""News,""Tech,""Local,""On-Air," "Weather,""Health,""Summary," and "BBS."

Using dynamic programming and the weighted A priori association rule mining approach, the next likely page was predicted [11]. Based on each page's visitation order, significant weights are allocated to the web pages. The navigation sequences' order is preserved with the help of the weighted A priori algorithm, and the development of ideal rules is aided by dynamic programming.

The MSNBC dataset has been used to evaluate the system in terms of both time and space. The experiments show that 33% support and 64% confidence are guaranteed by the regulations. Next, a different approach based on Particle Swarm Optimization (PSO) and weighted association rule mining has been introduced [11]. The approach has been tested with web log data that has a high volume of entries. The PSO technique has been utilized in this methodology to optimize the association rule mining algorithm's parameters, resulting in the evolution of optimal rule patterns. The parameters lift, correlation, support, and confidence make up the objective function that has to be maximized.

In order to create soft clusters that take into consideration both sequence and content similarity, a different method based on a rough set-based similarity upper approximation has been proposed [12]. The next page the user would prefer is predicted by applying singular value decomposition to a response matrix that is produced from the soft clusters.

Users and sessions are initially identified using data from web logs [1]. The number of times a certain user has viewed a particular web page is then provided via a user visit matrix hat is created. To group users of similar interest, the matrix is normalized and then sent to hierarchical clustering or Markov models.

PROPOSEDWORK

The Internet is becoming a non-resource of information. The evidence contained much too few options for the user to choose from and far too many suggestions. Choosing the right product or information has become a time-consuming and tedious task. To make the process of choosing the information easier, a recommendation system was developed. This might point to a website, but it might also point to another online tool or resource that the company could make utilize.

Discovering Sequential Patterns from Web Page Navigation

The patterns found in web server logs originating from web page navigation sequences are organized into user groups that enhance web personalization and intriguing page categories that are later visited. This improves cache optimization. Analysis is used to determine which page categories are fascinating enough for users to peruse and find to be very interesting. These pages have the greatest priority in the cache. Diverse techniques are employed to create user groups when it comes to user group patterns. An attempt is made to form groups so that people with similar interests are grouped together. Web page navigation sequence analysis involves assembling datasets, pre- processing information, building feature vectors, building learning models, evaluating results, and analysing and deriving desirable patterns.

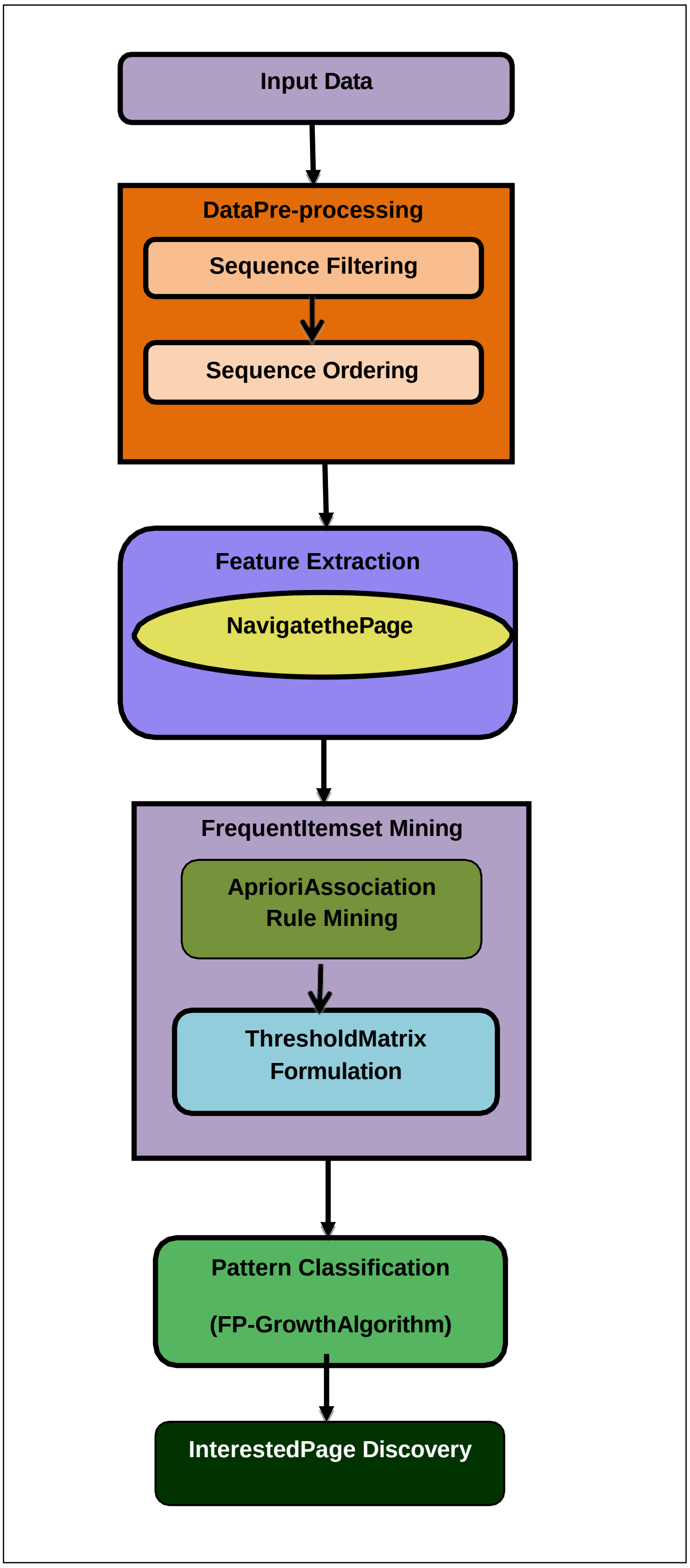

Fig.3.1 Flow Diagram of SPDM

Figure 3.1 depicts the suggested work for pattern identification. Certain methods are exclusive to a particular type of pattern, while others are ubiquitous for the identification of all patterns. The Dailies of India serve as a database representation for the user, suggesting to them potential browsing options that align with their needs at a given moment.

Data Filtering

Most of the patterns are found by sequence length- based filtering. The navigation sequences that visit only six distinct or similar page categories are screened throughout this filtering process, and the remaining sequences are discarded. Sequence ordering is carried out while grouping users via graph traversal. Different orderings of the original sequences are applied in order to examine the effects of ordering and select the optimal ordering for pattern identification. Pair wise sequence alignment is included because a small number of orderings may provide navigation sequences with different lengths. It gives information on the number of page categories that are parallel in two web page navigation sequences.

Feature Extraction

By automatically evaluating the repair verbatim, the bi-level feature extraction-based text mining for defect diagnosis can satisfy the aforementioned challenges. To achieve the required results, our key idea is to extract fault features at the syntax and semantic levels, respectively, and then fuse them. The suggested feature fusion of two levels may improve the accuracy of fault diagnosis for all fault classes, especially minority ones, given that the extracted features at each level have flaws and emphasize various aspects of feature spaces.

Frequent Item Set

The three subtypes of statistical approaches are univariate, multivariate, and hybrid [8]. When it comes to taking individual variables into account, univariate methods also referred to as feature filtering methods include information gain (IG), chi-square, occurrence frequency, log likelihood, and minimum frequency thresholds. While univariate methods are computationally efficient, they do not take into account attribute interactions. A feature vector with binary values (0 and 1) for visiting and non-visiting page categories is developed in order to identify the interesting page categories that are subsequently accessed. Furthermore, a class is required for forming user groups using supervised techniques. The same feature vector containing binary values for visit and non-visit to page categories is used for grouping based on two-class classification.

Frequent Pattern Classification

A data mining method called FP-Growth(Common Pattern Growth) discovers recurring patterns or item sets in a data set. It works by creating a compact representation of the dataset called an FP- Tree. Then, from the ground up, common patterns are built using the FP-Tree. In large datasets, the FP- Growth technique can efficiently identify common patterns and is very scalable. It is also more effective than the Apriori algorithm, which is another popular method for mining frequent item collections.

Finding patterns or correlations in a dataset that recur frequently is known as frequent pattern mining in data mining. Usually, to locate items or sets of objects that commonly appear together, vast datasets are analysed. Finding recurring patterns or item sets in a given dataset is the primary goal of frequent pattern extraction, a crucial task in data mining. It includes identifying sets of elements that regularly appear together in relational or transactional databases. This process can provide insightful views into the relationships and linkages between various elements or characteristics in the data.

Discovering Sequential Pattern

Formulate the problem of identifying a set of interesting sequences that are useful for explaining a sequence database. First we will define some preliminary concepts and notation. An item i is an element of a universe $U = \{1, 2, \dots, n\}$ that indexes symbols.

A sequence Sis simply an ordered list of items(e1,...,em)such that $ei\in U\forall i$. A sequence Sa =(a1,..., an) is a subsequence of another sequence Sb=(b1,..., bm), denoted $Sa \subset Sb$, if there exist integers $1 \le i1 < i2<...<in \le m$ such thata1=bi1,a2=bi2,.. .,an=bin (i.e., the standard definition of a subsequence).

Algorithm of SPDM

Step 1:	DFS-Pruning Node n=(s1,s2,….sn,In)
Step2:	Foreach(iSn)
Step3:	if ((s1,.......... , sk , {i}) is frequent)
	Stemp= Stemp{i},For each(i Stemp)
	DFS-Pruning((s1,…......................... ,sk,{i}),Stemp, All elements in Stemp greater than i)
Step4:	For each(I In)(9)if((s1,.....,sk» {i}) is frequent)
	if ((s1,..........., sk » {i}) is frequent)
	Itemp = Itemp {i}
	DFS-Pruning ((s1, ..., sk {i}), Stemp, all elements in Itemp greater than i
	Check whether a discovered sequences 'exists such that eitherss' ors's, and data base size L(Ds)=L(Ds')
	If such super-pattern or sub-pattern Exists then
	Modify the linkinL; Return(4)else insert s inL;
	scanDs once, find the set of frequent item set α such that α can be appended to form a sequential pattern s α.
	If no valid α available then, Return
Step5:	For each valid α do Call Clo Span(s α,Dsα,min supp, L
Step6:	End

All that's left of a sequence database is a list of sequences X (j). Further, we say that a sequence S is supported by a sequence X in the sequence database if S ⊂ X. Note that in the above definition each sequence only contains a single item as this is the most important and popular sequence type (cf. word sequences, protein sequences, click streams, etc.).

A multi set M is a generalization of a set that allows elements to occur multiple times, i.e., with a specific multiplicity #M. The architecture comprised of data collection, data pre-processing, formulation of feature vector, application of computational models, evaluation of the rules generated by these models and extraction of desired patterns.

RESULTSAND DISCUSSIONS

The performance of the proposed methodology in discovering the patterns is assessed through suitable performance evaluation techniques. Various performance metrics are employed for different methods. In the case of association rule mining, confidence, support, rule length and lift are the measures of interestingness that rank the association rules that have been generated.

X, the component on the left hand side of the rule, is referred to as the antecedent while Y, the component on the right hand side of the rule, indicates the consequent. The rules can be read as: if X holds true, then Y is said to be true. Many parameters are defined to rank the significance of a rule. Originally, sustenance specifies in what way regularly an item set seems in the dataset. The support of an item set (X) with regard to a set of transactions T is defined as the proportion of transaction t in the dataset which holds the item set X. In other words, it is the ratio of the number of transactions that hold the item set X to the total number of transactions.

$$Conf(X=Y)=\frac{support(X\cup F)}{support(X)}\text{-------}(1)$$

Yet another parameter, lift is defined as the ratio of the observed support to that of the expected support if X and Y have been independent.

$$(X=Y)= \frac{support(X\cup F)}{support(x)}\text{-----------}(2)$$

Using the values of TP, TN, FP and FN, the performance metrics, accuracy, recall and precisionare computed as follows. Accuracy denotes the total number of correctly predicted instances to the total number of instances in the case of describing patterns.

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\text{--------}(3)$$

A larger MSE indicates that the data points are dispersed widely around the central moment (mean), whereas a smaller MSE suggests the opposite. A Smaller MSE is preferred because it indicates that the data points are dispersed closely around its central moment, is calculated as

$$Mean\ Squared\ Error={}^{1}\sum_{n}(xi-yi)1-n\text{------------}(4)$$

A method for assessing machine learning modelsis called cross validation. It involves training several models on subsets of the available input data and comparing their performance on the complementing K*1-foldcross-validationi=k-1

$$cv_{(k)}=\frac{1}{k}\sum_{i=1}^{k}MSE_i\text{-------------------------}(5)$$

Use cross validation to detect over fitting ie., falling to generalize a pattern. A sing lek-fold cross validation is utilized with both a validation and test set. The total data set is split into k sets.

Mean Directional Accuracy (MDA) also known as mean direction accuracy is a measure of prediction accuracy of a forecasting method in statistics that compares the forecast direction to the actual realized direction. The following formula defines it:

MDA = 1 N ∑ t 1 sign (X t − X t − 1) == sign (F t − X t − 1) (6)

Where :is the actual observations time series.

Table 4.1. Pattern classification results for SPDM

Methods	CV (%)	MSE (%)	Accuracy (%)	MDE (%)
SVM	89.24	87.45	88.23	89.54
PDM	90.12	90.34	87.59	91.67
SPDM	90.23	90.16	92.67	95.12

In Table.4.1 explains the pattern classification results for sequential pattern discovery method with results for sequential pattern discovery method with existing methods. The proposed work Discovering Sequential Patterns gives 95.12% of accuracy and 92.67 values in F-measure. While comparing the existing methods the proposed method produce the high accuracy results.

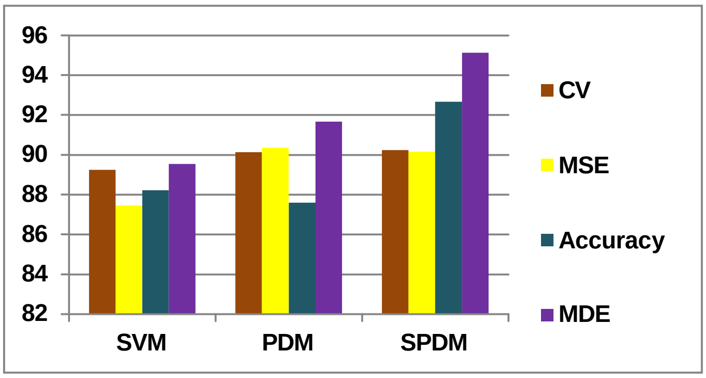

Fig.4.1 Pattern Classification results for SPDM

In Fig.4.1 explains the comparative chart on discovering patterns on web page navigation sequences and with existing methods. The Discovering Sequential Pattern method gives 95.12 % of accuracy and 92.67% values in F-measure. While comparing the existing methods the proposed method produce the high accuracy results.

CONCLUSION

Pattern discovery from web page navigation sequences involves application of data mining and graph theory methods to find the patterns viz., subsequently accessed interesting page categories and User groups. In this work, the processes involved in the SPDM method gives high accuracy rate while comparing the existing methods also discovering frequent patterns from web page navigation. Better website organisation and personalisation may be made possible using this strategy. Applying data mining and pattern classification techniques to online page navigation sequences allows for the pattern finding from accessed interesting page categories and user groups. The suggested SPDM approach facilitates the identification of sequential patterns from online page navigation very efficiently.

REFERENCES

[1] Al-asadi, TA &Obaid, AJ 2016, ‘Discovering similar user navigation behavior in web log data’, International Journal of Applied Engineering Research, vol. 1, no. 16, Pp: 8797-8805.

[2] A. Abdalla, T. Ahmed, M. Seliaman, Web usage mining and the challenge of big data: A review of emerging tools and techniques, in: I.R.M. Association (Ed.), Big Data: Concepts, Methodologies, Tools, and Applications,Vol.6,IGIGlobal,2016,Ch.42,Pp:

[3] 899–928.doi:10.4018/978-1-4666-9840-6.ch042.

[4] Dr.S.Brindha, Dr.S.Sukumarn, Relevance Pattern Discovery for Text Classification Using Taxonomy Methods” in International Journal for Science and Advance Research in Technology (IJSART)” Volume 4 Issue11–November2018 ISSN[online]:2395-1052.

[5] Dr.S.Brindha, Dr.S.Sukumaran, An Analysis on Big Data Interrogation Explore Problems and Tools, International Journal of Novel Research and Development (IJNRD), ISSN:2456-4184, Volume 5, Issue 6,June 2020.

[6] Corso, M.P.; Perez, F.L.; Stefenon, S.F.; Yow, K.- C.; GarcíaOvejero, R.; Leithardt, V.R.Q.Classification of Contaminated Insulators Using k-Nearest Neighbors Based on Computer Vision. Computers 2021, 10, 112. [CrossRef]

[7] Fan,G.-F.;Guo, Y.-H.;Zheng,J.-M.;Hong,W.-C. Application of the Weighted K-Nearest NeighborAlgorithm for Short-Term Load Forecasting. Energies 2019, 12, 916.

[8] Gajan, S. Modeling of Seismic Energy Dissipation of Rocking Foundations Using Nonparametric MachineLearningAlgorithms.Geotechnics2021,1, Pp: 534–557. [CrossRef]

[9] J. Gamalielsson, B. Lundell, S. Butler, C. Brax, T. Persson, A. Mattsson, T. Gustavsson, J. Feist, E. Lönroth, Towards open government through open source software for web analytics: The case of matomo, JeDEM-eJournal of e Democracy and Open Government 13 (2) (2021) Pp:133–153, https://doi.org/10.29379/jedem.v13i2.650

[10] Ismail A, Abdlerazek S, El-Henawy IM (2020) Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping. Sustain (switz). https://doi.org/10.3390/su12062403.

[11] S. Knight-Davis, Using awstats to analyze logs from ezproxy and from the public opac logs, in:Spring Forum: Collection Management and Technical Services Committees, 2017.

[12] Malarvizhi, M & Mary, SA 2017, ‘Discovery of effective relationships among web pages using hybrid weighted PSO’,Applied Mathematics and Information Sciences, vol. 11, no. 1,Pp. 235-241.

[13] Mishra, R, Kumar, P &Bhasker, B 2015, ‘A web recommendation system considering sequential information’, Decision Support Systems, vol. 76, Pp. 1-10.

[14] P. Shah, H.B. Pandit, A review: Web content mining techniques, Data Eng. Smart Syst. (2022) Pp:159–172, https://doi.org/10.1007/978-981-16-2641-8_15.

[15] S. Kumar, R. Kumar, A study on different aspects of web mining and research issues, IOP Conference Series: Materials Science and Engineering, vol. 1022, IOP Publishing, 2021, Pp: 012–018, https://doi.org/10.1088/1757-899X/1022/ 1/012018.

[16] V. Jain, K. Kashyap, An efficient algorithm for web log data pre-processing, in: Machine Vision and Augmented Intelligence Theory and Applications, Springer Singapore, Singapore, 2021, Pp: 505–514, https://doi.org/10.1007/ 978-981-16-5078-9_41.