JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
31st August 2018, Volume 3, Issue 2, JSEIS, CADMEI Copyright © 2016-2018
www.jseis.org

AN UNSUPERVISED APPROACH FOR USER BEHAVIOUR CLUSTERING OF WEBSITES USING THE NAVIGATION PATTERNS OF WEB USERS

[1] SYED TAUHID ZUHORI, [2] JAMES MILLER

[1,2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
Email: [1]zuhori@ualberta.ca, [2]jimm@ualberta.ca

ABSTRACT

Web traffic and e-commerce activities are increasing rapidly day by day. Hence, understanding the behavior of users based on their interactions with a website is becoming important. Web usage mining is needed for that. It works on web clickstream data in order to extract usage patterns. There are two major challenges involved here: One is preprocessing the raw data to provide an accurate picture of how a website is used. Other is to present the rules and patterns that are potentially interesting to the users. This paper proposes and develops an architecture for performing that. Firstly, we clean the web server logs by using a traditional clustering approach. Then, we apply a Discrete Time Markov Chain approach to generate a model of the user behavior. For generating the nodes for the model, we use a technique (regular expressions) to find out the atomic propositions. Then we find a directed graph as an output of a DTMC inference process. Next, we apply spectral clustering on that directed graph, which works on the affinity of the graph nodes and divides the nodes into clusters. Finally, we use graph traversal algorithms and discover the navigation patterns of web users for each cluster. To evaluate the approach, we use server log files from the website www.ualberta.ca. This approach is very useful to simplify better web personalization and website organization. It automatically finds out clusters of usage patterns undertake by the users, and makes this data available to the web designer. Hence the web designer will know the interests of the user and this will help them to develop a more personalized space for users.

Keywords: website personalization; atomic propositions; Discrete time Markov chain inference process; navigation patterns; spectral clustering;

1. INTRODUCTION

At present, the World Wide Web (WWW) is growing at an astounding rate not only measured by the sheer volume of traffic but also by the size and complexity of websites. For this reason, more concern is needed on some issues such as Website design, Web Server design and simply navigating through a website. As the amount of knowledge in the WWW has grown explosively, it is becoming much more difficult for users to access relevant information efficiently. So, the design phase for a website is becoming more complex. An important input to these design tasks is an analysis of how a website is being used. Some straightforward strategies can be included in this analysis, such as: page access frequency and the common traversal path of a user. In this paper, we investigate the second strategy, by deducing the common traversal path of each arbitrary user. This information can be used to restructure a website, so that the new website can give a better service to web users. The main challenge for website designers is to organize the content of a website in such a way so that it can successfully meet the needs of its users. Our approach, the automatic clustering of web users' navigation patterns, can provide a useful tool to solve this challenge faced by website designers. The fundamental characteristic of our approach is that it can extract the navigation profiles that can capture the similar behaviour of website users. This profile can also be used for predicting the navigation behaviour of the current, and future, website users. Our automated system has benefits on both sides. If we think from a user perspective, the clustering of navigation patterns can enhance the quality of personalized recommendations. Depending on the recommendations, the links of the most visited pages will then be inserted dynamically for display that can help web users to access their favourite information efficiently. On the other side, if we think from the website designer perspective, the clustering of navigation patterns can guide designers and they can organize the content of their website according to the desire of their users. Materially, organizing or developing a website is both static and reactive. Whereas the navigation patterns of users will be learned periodically and the change in their navigation interest can be captured regularly. Therefore, managing websites will be dynamic and proactive. As a result, the visitor of the website will be interested to become consumers or users of the site.

Our aim is straightforward. We partition the users into a cluster. So only the users with similar behaviours are placed in the same cluster. Then, for each cluster, we display the behaviour patterns as an output within that cluster. For clustering the user, we use the server log files as raw input. The characteristics of the raw data can be described as i) the server log file is converted into a set of sequences, one sequence for each user session, ii) each sequence is represented as an ordered list of discrete symbols, and iii) each symbol represents one of the several possible categories of web pages requested by the user.

In this paper, we propose a system which can take this large dataset as input, and finally cluster user behaviour patterns that can track the movement of the user throughout the website. Firstly, we clean the web server logs by using a known clustering approach. After that, we apply a Discrete Time Markov Chain approach. It can generate a model of user behaviour. Initially, we use a technique for finding atomic propositions [Carlo et al., 2014]. Regular expressions are used to find out these atomic propositions. The atomic propositions are unique links on any website. Then, we find a directed graph as an output of the Discrete Time Markov Chain inference process. After that, we apply spectral clustering on that directed graph. The clustering approach divides the nodes into clusters. Finally, we use graph traversal algorithm that can find out the navigation patterns of the users. We illustrate this process by using server log files from the website www.ualberta.ca.

Our paper makes four key contributions:

- We have built a dynamic system that has the ability to adapt to the changes in the unique links in the website. Also, the system can incorporate these changes in the directed graph model of the website.
- Our proposed system is designed in a way such that it clusters the unique links based on the frequency of visits made by the users to the unique links of the website.
- Our proposed method helps in website designing and this is aided further by profiling the user behaviour based on the percentage of visits for every unique link in the website.
- Finally, the system finds out the navigation patterns of the user and all this is done using the basic of information available to us – "log files".

The rest of this paper is organized as follows: In section 2, we review recent research on the topics of user behaviour clustering. Section 3 describes the architecture of the proposed system for clustering user navigation patterns. The experimental results are presented; and a discussion about the evaluation of these results is presented in Section 4. Finally, Section 5 summarizes the paper with conclusions.

2. RELATED WORK

We examine research work related to our study in this section. Our work is related to web mining. It can be categorized into three active research areas depending on what components of web data are mined. The first one is Content Mining which is the process of extracting important information from the content of websites such as contents of documents or their descriptions related to websites. The next one is Structure Mining that uses links and references within web pages. After analysing that, it can obtain the fundamental topology of the interconnections between web objects. The final one is usage mining that studies on user access information from log server data and can also extract interesting usage patterns. Our paper is based on this type of web mining. Hence we include research work related to mining for user navigation patterns.

Igor et al. [2000] propose a methodology for visualizing and exploratory analysis of the navigation patterns on a website. As a test bed, they used server logs of individual browsing records of thousands of users at the msnbc.com site. They divided the web site users into clusters such that only users with similar navigation paths through the site are placed into the same cluster. They generated a thousand of clusters. Each cluster represents a navigation path and also number of visits of each particular link of that navigation path. They applied a Model-based clustering technique that uses a collection of statistical models to group the users of same interests instead of using the attribute values for two different users. Initially they assigned k empty clusters. When a user arrives at the website then they assign it to one of K clusters with some probability and when a user is in a cluster then their behaviour is generated from model specific to that cluster. For assigning the users in a cluster they consider two things; i) the number of navigation paths of the user ii) the probability distribution of each links of the Website on that navigation path. Finally, for each cluster, they displayed the user's interest on each links of the website and analysed the behaviour of the users within that cluster. Yunjuan et al. [2001] proposed an approach to cluster website users into different groups. Their technique

generated common user profiles too. They used a concept of mass distribution in Dempster-Shafer's theory. They assigned probabilities for single pages based on their co-occurrence in sessions. These groups were refined while using Dempster-Shafer's theory of combination. The modified or refined groups of pages are common user profiles. The advantage of their process is dynamicity. Dempster's rule for the combination of evidence can allow an expression of uncertainty with respect to aggregated components. Thus, this rule is suitable for clustering pages into groups.

Haibin et al. [2007] propose an automatic classification of web user navigation patterns. They apply a novel approach to classify these patterns. With this classification, their approach can predict a user's future requests. Their approach was based on the combined mining of web server logs and the contents of retrieved web pages. They used a character N-gram-based approach for representing the content of web pages. Then they combined it with user navigation patterns. They built user navigation profiles, composed of a collection of N-grams. Miao et al. [2012] presented a technique to capture Web users' behaviour based on interest oriented actions of users. In their approach, they utilized Random Indexing, an incremental vector space technique, that can identify latent factors or hidden relationships among web users' navigational behaviour. This technique can allow for continuous web usage mining. They utilized a web server log as input and "pre-processed" it. Then they extracted the page according to user interest. After that, they split the pages according to their URL to find segments of pages as output. Next, they generated an indexed vector that can store the navigation patterns of each user of the website. As example, if there are U users of a system and N navigation paths then the indexed vector will U*N. Finally, they applied a k-means algorithm on that vector for clustering of the users and hence they deduce common navigation paths of the users. Schur et al. [2013] presented a fully automated tool that can mine the explicit behaviour models of enterprise web applications. The name of their system was ProCrawl. They focus their application was supporting system testing and maintenance. This system requires the URL of a web application, the website user's login credentials, a scope definition (the part of the web application to be observed), and a start event that denotes the starting point from where the observation can start. Their system handles rich web 2.0 applications, including dynamic technologies such as AJAX. Their resulting behaviour model was a finite state automaton in which the nodes denote abstract individual states of web applications. Their model can be directly used for model-based regression testing. Schur et al. [2015] updated their proposed "ProCrawl" system by analysing the multi-user workflow models. For inferring decision rules from the input data, they presented a machine learning approach. Their approach can support the execution of business related process. Multiple interacting roles of the user were typically involved there. They worked on a system of ordering process of a seller and a vendor or a system of reviewing process where an author, a reviewer and a program chair are involved. They were centered on data stored in a shared data store that is manipulated when executing the processes. The data was collaboratively edited through browser-based clients. This functionality of the clients was varied depending on the user role. Gang et al. [2016] presented the design and evaluation of a practical and scalable clickstream tool. They use this tool for user behaviour analysis. Their system used a similarity matrix between clickstreams. Using this matrix, they build similarity graphs. These similarity graphs can capture behavioural patterns between users. The edges of the graph can capture similarity distances between the clickstream of the user. The clusters represented the user groups with similar behaviour. They used a hierarchical clustering approach to detect the most popular behavioural patterns. They also used an iterative feature pruning technique. Using this technique, they removed the dominant feature from each subsequent layers of the user. For evaluating their system, they presented case studies on two large-scale clickstream traces (142 million events) from real social networks.

In this paper, we focus on the clustering of navigation patterns of web users according to their past visits. We analyse the server log files and apply clustering on the navigation patterns. The novelty of our approach is that it can create clusters of navigation patterns that assists in understand similar user behaviour. Previous research on this area used clickstream data, visualized the navigation patterns of the web user, created clusters of web users, captured web user behaviours, but they did not focus on the clustering of user navigation patterns. In the existing methods, the user behaviour alone is found out by identifying the user's interest on certain links. On the other hand, the proposed method tracks how the user navigates through the website and uses this data to study the user behaviour. By this approach, the real behaviour of the user is understood and clustering is done based on this data. Navigation clustering denotes the way in which a group of users of similar type navigates through a website in a session. Hence, we can understand the pattern on behaviour which is very important for web personalization. By this method, we can find out the total navigation history of a user for a session. Some previous works have used large sized log files (Gang et al. used a log file that contains 142 million events). But they analysed the website manually and find out the unique links. In our

proposed system, we use a regular expression for finding the unique links of the website. Therefore, our system can work if the website is changed or updated.

3. THE NEW SYSTEM

In this section, we discuss the overall architecture of the proposed system. We start our work with web server log pre-processing by cleaning the irrelevant data, identifying the sessions and users. At the next step, we give semantics to URLs occurring in the log file by providing them unique names by means of the set of atomic propositions that denote relevant user actions. Our proposed method consists of four major modules: weblog pre-processing, inferring the model, spectral clustering and building navigation patterns.

3.1 Weblog pre-processing

The aim of weblog pre-processing is to reformat the original weblogs. In web server logs, the web server usually registers all users' access activities on the website. In our work, we include data cleaning, user differentiation, and session identification that are the same for any web usage mining problem. For cleaning the dataset, we the requests executed by automated programs such as web robots, spiders, and crawlers. In case of HTTP status code, we only consider successful entries that are between 200 and 299. We delete unknown request methods and consider only "Get" and "Post". After that we differentiate the users. If the same user makes two distinct arrivals to the site, then it should be determined accurately. To achieve this, a unique identifier is associated with a particular user. Every time a user visits the site, this identifier is used to identify the user and differentiation is made accordingly. This is accomplished with the help of a client ID, an IP address, so subsequent visits to the same site can be associated with the same user. Finally, we identify the sessions for user. We use Time-based expiration at our system that occurs either when there is inactivity for specified time duration or at the end of the day. As soon as a user arrives on the site, back counting of the time constraint gets started. For every activity of the user, the constraint value gets updated and the session is active till the allowable time (for inactivity) becomes zero. Any activity of the user after that is counted as a new session. However, this rule is different at the end of the day. Even if the time constraint is not zero, the session will be ended at the end of the day and a new session will be counted for the start of the next day.

3.2 Inferring the model

After the pre-processing step, we then generate the model of a system based on user access sessions. Our inference process analyses the log file of the application and infers a set of discrete time Markov chains (DTMCs). There are two basic steps of this model generation. At first, we identify atomic propositions that are the nodes of our model. Then for the given set of atomic propositions AP, our process infers a set of discrete time Markov chains (DTMCs).

3.2.1 Identifying the atomic proposition

In the first steps of the model generation process, we work on the semantics of the rows of the groomed log file by means of a set of atomic propositions (AP) that indicate which entries can be assumed as valid for a certain entry in the log file. As an example, the proposition homepage is associated with a row in the log file. Carlo et al. [2014] proposed a special code fragment, called filter, for finding out the atomic propositions from the log files. This filter is parameterized with a regular expression that selects the groups of URLs which match with the regular expression. Our approach scans the groomed log file and it invokes the filters with matching parameters on each row. Then finally it associates the atomic propositions returned by the filters to entries in the log file. Table I shows some atomic propositions associated with the URLs of the www.ualberta.ca applications.

Table. 1 The relevant URL's of the Website

URL	Description	Atomic Proposition
/home/	Homepage of any website	Home
/account/login/	The page by which all personal information can be tracked.	Profile
/Libraries/	The page represents the data of the Libraries	Libraries
/photos/	The page shows the photos of different events	photos
/people/	The page represents the people information related to the website.	people
/events/	The page which shows the past and future events news.	events

3.2.2 Discrete Time Markov Chain inference process

Discrete Time Markov Chains are finite state automata augmented with probabilities: each state is characterized by a discrete probability distribution that regulates the outgoing transitions. Let, Xt be a random variable. It represents the state of a system at time t, where t=0, 1, 2, …. A stationary Markov chain is a special type of discrete-time stochastic process with the following assumptions:

- The probability distribution of the state at time t+1 depends only the state at time t, and does not depend on the previous states leading to the state at time t;
- A state transition from time t to time t+1 is independent of time.

Suppose Pij is the probability that the system is in a state j at time t+1 and is in a state i at time t. If the system has a finite number of states, 1, 2, 3,…,s, then the Discrete Time Markov Chain can be defined by an SXS transition probability matrix. There is an initial probability distribution Q that is represented by a 1XS matrix. Where qi is the probability that the system is in the state, i at the time, t=0. And the total probability of a definite state is 1. So,

$$\sum_{j=1}^{j=s} P = 1$$

The probability that a sequence of states X1, …, XT at time 1, …, T occurs in the context of the stationary Markov chain is computed as follows:

$$P(X_1, X_2, \ldots\ldots, X_T) = q_{x_1} \prod_{t=2}^{T} P_{X_T} - 1X_t$$

The transition probability matrix, P and the initial probability distribution of a stationary Markov Chain, Q can be learned from the observations of the system state of the past. If the observation of the system states X0,X1,X2,……,XN-1 at time t=0,1, 2,…,N-1, we find out the transition probability matrix and the initial probability distribution by following equations:

$$\mathrm{P_{ij}} = \frac{N_{ij}}{N_i} \text{ and, } q_i = \frac{N_i}{N}$$

Where, Nij is the number of observation pairs Xt and Xt+1 with Xt in state i and Xt+1 in state j. Ni is the number of observation pairs Xt and Xt+1 with Xt in state i and Xt+1 in any one of the states 1, …, s and N is the total number of observations.

The transition probability matrix is set depending on atomic propositions. If there are n atomic propositions, then the transition probability matrix will be nxn. Our proposed inference engine examines propositions associated with filters and the current rows of log files. In our groomed log file, only the atomic propositions with the user IP address are included. The atomic propositions, which match with the log file set as a destination state. Transitions extracted from the log file are used by the inference engine to update two sets of variables that are initially set to zero: A set of variables, countij for each pair of states' (si, si)∈ s *s and a set of variables, ti for each state, si ∈ s. The engine increments both variables countij and ti; it increments the variable, counti,j for each transition from a state, si to a state, sj and variable, ti for each transition whose source state is si, independent of its destination state. The variable, ti indicates the number of times the users exited state, si. On the other hand, the variable, counti,j represents the number of times the users moved from a state, si to a state, sj . For each row in the log file the inference engine updates the variables. The inference engine uses these variables to compute the (i,j) entry of the stochastic matrix, P that represents the probability of traversing the edge from a state, si to a state, sj. The computation is done by the following equation:

$$\mathrm{P(s_i, s_j)} = \frac{\mathrm{count_{ij}}}{\mathrm{t_i}}$$

So, in general, the probability, P(si, sj) is computed as the ratio between the number of traversals of the transitions from the state, si to state, sj and the total number of traversals of the transitions exiting state, si.

3.3 Spectral Clustering

Clustering is one of the most widely used techniques for exploratory data analysis. In our approach, we use Spectral Clustering algorithms for dividing the data into groups proposed by Jordan et al. [2010]. At pre-processing, we calculate the adjacency matrix and produce the affinity matrix from the adjacency matrix. From the Affinity Matrix, we can find out the affinities between all pairs of edges of a matrix. Since this matrix is used for similarity measurement; we use the Gaussian distance, for entry (i,j)for estimating the Affinity Matrix. After finding the Affinity Matrix, A we find out the spectral representation of it. Then, we find out the Degree Matrix, D. Then we generate the Laplacian Matrix, L by subtracting A from D. The Laplacian Matrix is an n x n matrix, L, can be defined as:

$$L_{ij} = \begin{cases} \deg(V_i); & if\ i = j \\ -1; & if\ i! = j\ and\ V_i\ is\ adjacent\ to\ V_j \\ 0; & Otherwise \end{cases}$$

Next, we find out the symmetric version of the Laplacian Matrix. The symmetric Laplacian Matrix can be defined as:

$L_{SYM}=D^{-1/2}\,LD^{-1/2}=I-D^{-1/2}AD^{-1/2}$

Where, Matrix, I is the Identical Matrix. The elements of LSYM are given by:

$$L_{sym_{i,j}} = \begin{cases} 1; & where\ i = j\ and\ \deg(V_i) \neq 0 \\ \dfrac{1}{\sqrt{\deg(V_i)\,\deg(V_j)}}; & if\ i \neq j\ and\ V_i\ is\ adjacent\ to\ V_j \\ 0; & Otherwise \end{cases}$$

After that we find out the Eigen Vector. If T is a linear transformation of a vector space, V over a field, F into itself and v is a vector in V that is not the zero vector, then v is an eigenvector of T, if T(v) is a scalar multiple of v. This condition can be written as: $T(v) = \lambda v$. Where λ is a scaler in the field, F. For finding the Eigen Vectors, we can consider n-dimensional vectors that are formed as a list of n scalars. Suppose we want to find out an Eigen Vector of Matrix A then we use the characteristics equation: $|A - \lambda I| = 0$. For further processing, we need only the largest k Eigen Vectors. Actually, these largest Eigen Vectors have the dominating behavior over the whole Eigen Vectors. As a final stage of the spectral representation, we construct the Normalized Matrix of the k-largest Eigen Vectors by using the following equation:

$$U_{ij} = V_{ij} / \left(\sum_j {V_{ij}}^2 \right)^2$$

3.4 Building navigation patterns

Initially, we use k-means clustering for finding clusters within the Normalized Matrix. It divides the matrix elements into k clusters such that a metric relative to the centroids of the clusters are minimized. Next, we estimate the optimal number of clusters. The optimal value of k is derived using the Elbow method proposed by Trupti et al. [2013]. Suppose there are n independent samples (28 unique links at our system), clustered into k clusters C1 ,C2 , ...,Ck. At first, we compute the sum of the squared error (SSE) for some values of k which is defined as the sum of the squared distance between each member of the cluster and its centroid. Then, we "plot" k against the SSE and we see that the error decreases as k gets larger; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly that produces an "elbow effect" in the graph. The "plotting" and "Elbow location" is implemented by package [https://github.com/Sanjay015/Optimal_Clusters]. After finding out the optimal number of clusters, we derive the common navigation paths of each cluster. Using these navigation paths, we represent the users' interest. We propose an algorithm which can derive the navigation paths of each cluster that requires the starting node and ending node of the graph as input.

4. EVALUATION AND RESULTS

In this section, we discuss our results and their evaluation process. We start with the discussion of the experimental data set in Section 4.1 that is collected from University of Alberta. Web log pre-processing results are then presented in Section 4.2. After that we represent the result of the Discrete Time Markov Chain inference process in Section 4.3 Then we evaluate our clustering results for different numbers of users in Section 4.4. Finally, we represent our clustering results in Section 4.5 by representing the cluster of user's navigation paths (common user profiles), common web links of the clusters and the member list of each cluster. The results show that our proposed automatic system of clustering navigation patterns provides accurate navigation paths.

4.1 Dataset and Environment

The dataset used for the experiments allows us to analyse web log data and web pages. We have conducted our experiments on an Apache server log access file; this file belongs to the University of Alberta's official website / application, www.ualberta.ca. Figure 1 shows the screen shot of the home page of the website. The University of Alberta icon leads to the home page and the links at the top of the homepage also lead to a separate page with detailed information and links.

Figure. 1 Screen shot of University Alberta websites homepage

4.2 Weblog Pre-processing

Table 2 provides some statistics of the experimental dataset after preprocessing. In this dataset, 7,290 clean entries are picked out and there are approximately 600 different users who accessed the Web server in January 24,2016 to 26,2016. Although 1686 sessions were identified by the session-duration-based method, only 1250 of them contain more than 2 requests, we regard this as a minimal visit. In the 3 days' period, the maximum visit time of a user on a link is "Bear Tracks" at 7 minutes. On the other hand, the minimum stay time of a user on a link is "Home" at 0.01 Minutes. The most visited links by the user is "Bear Tracks" and the least is "Photos".

Table. 2 Statistics of the experimental dataset

Attributes	Day 1	Day 2	Day 3
Total access entries	2341	4421	3461
Clean access entries	1090	3907	2293
Total number of Unique links	28	28	28
Total numbers of user	110	308	182
Identified sessions(session-duration-based)	381	730	575
Identified sessions (≥2 requests, session-duration-based)	280	620	350

Highest accessed link	Bear Tracks	Bear Tracks	Home
Lowest Access link	Photos	Photos	Photos
Maximum stay time of user on a link	4 Min	7 Min	5 Min
Minimum stay time of user on a link	0.01 sec	0.01 sec	0.01 sec

We extract 28 Atomic propositions (Unique Links) from the web server log file. Table 3 provides the basic description of the extracted atomic propositions that represents the purpose of the propositions.

Table. 3 Statistics of the experimental dataset

Atomic Proposition	Purpose of this link	Atomic Proposition	Purpose of this link
Module	It represents the content of the courses offered by the University of Alberta.	FAQ	It represents the frequently asked questions with their answers.
Customize	It represents the customized services such as workshops, presentations and online instructions to University and non-university groups on a range of academic topics.	People	It represents the information of the people related to the University. All types of people involved with the University such as the faculty members, Students, administrative stuffs are included in thin part.
Email & Apps	It represents the link of University web mail and other applications such as help desk, profile manager, my gadget etc.	News	It represents the news related to the University, research, conferences, workshops etc.
University Calendar	It represents the University Calendar that includes Academic schedule and holidays in Canada.	Menu	It represents different menus such as the personal information menu, the menus of Faculties and Programs, Departments etc.
Emergency	It represents some emergency services such as Student emergency loans, Graduate student's association bursary, emergency management etc.	Media	This link provides guidelines for staff, faculty, students and alumni who manage social media accounts on behalf of the University of Alberta.
Search	It gives the facility of searching anything related to University such as any faculty member, any course materials etc.	Libraries	This link provides information related to the University libraries. User can search Books, ejournals, newspapers, magazines, conference proceedings etc. that are available at the libraries of the University
Photos	It represents the photos related to the University such as photos of classrooms, Seminars, Workshops etc.	Full Web	It provides the service to Mobile users to access full website of the University.
Video	It represents the videos related to the University such as videos of Seminars, Workshops etc..	Transit	It represents the information about the Edmonton transit system, Universal transit pass (U-Pass), Transit schedule etc.
eClass	It represents an eLearning environment that is customizable and scalable to meet the needs of instructors and students in a wide variety of courses.	Athletic	It represents the information about the Fitness Centre of the campus, Campus recreational activities, information about the sports program offered by the athletics group etc.
ONEcard	Using this link, students can apply for an ONEcard and get all services	Bear Tracks	It provides the service links, such as Academic services, Financial services,

	related to it such as refill fund in the card, give a payment, Manage the card system, check the account balance etc.		Personal information updates, to students, applicants, employees and instructors in a safe and secure online environment.
Home	It represents the home page of the University.	Maps	It provides various maps of related to the university environment.
Login	It gives the facility of signing in to the University web site using a Campus Community ID (CCID) and password.	Events	It represents the information about the events that take place in and around the University campus, including all faculties.

4.3 DTMC inference process results

We generate our model using a Discrete Time Markov Chain inference process. We get a directed graph as an output of the DTMC process. Figure 2 represents the directed graph of DTMC of University of Alberta website. This directed graph is a visual aid that helps us in understanding how different components are related in the inference process. From the graph, it is evident that "Home" component is literally linked with every other component in the process. Other than "Home", "Email & Apps", "Menu", "Media", "Search", "About" and "University Calendar" are the components that seem to be more connected. "Social", "Video", "Athletic", "Transit" and "Photos" are the least connected components in the system. Other components in the system, though not highly connected, have considerable amounts of connections with the other components in the system. There are 30 nodes in our model. But we consider 28 nodes for clustering as there are two nodes for "Start" and "End". Since there are 356 edges in our resultant graph, we represent the transition probabilities using a simply colour-encoding instead of numbers. We represent the high probabilities by Black, Medium high probabilities by Green, Medium low probabilities by Blue and Low probabilities by Red colour.

4.4 Evaluation of clustering results

The clustering result evaluation is reported for 5 groups of users; Figure 3 represents the evaluation of clustering results. There are 600 users and the groupings are done by sub-sampling the user community, specifically users representing 20%, 40%, 60%, 80% and 100% of users are considered. Except the last group, where the whole population is considered, 10 different (random) variations of the groups are considered and the mean value of the performance is considered for evaluation purposes. For measuring the performance, we consider the accuracy of the clustering process. We manually pre-label the 600 users with resulting clusters starting at (k=2) system and ending with cluster (k=9) system. We assign the cluster number for all 600 users. Then we apply our approach on all sub-samplings; this essentially establishes a ground-truth for the evaluation. Subsequently, we find out the Accuracy by comparing the system generated output with the manually labelled output. As seen from the graph, for cluster (k=2), system the performance was highest for 80% of users and the next highest being the 20% users. In cluster (k=3) system, for 40% of users and 60% users the performance is same and 100% of users recorded the lowest performance. The performance of Cluster (k=4) system shows nearly equal performance for the first four groups and 100% of users record the lowest performance. There seems to be similar trend across all the cluster systems and 100% users record the least performance across all cluster systems. For 20% of users, Cluster (k=2) system shows the highest performance with an average value of 98.33. Again, for 40% of users Cluster (k=2) system shows the highest performance with an average value of 97.92. For other groups, too cluster (k=2) system shows the highest performance. However, perhaps the most important result is the demonstration of the stability of the results, regardless of the value of k or the amount of sub-sampling, the variation of the performance is finite. This suggests that the performance of the process is not significantly impacted by the selection of these parameters. However, a (small) trend may exist that the accuracy declines as k increases; but given that only 8 data points exist, it seems (numerically) inappropriate to provide definitive guidance on this topic. We also represent the basic information of each cluster in Table 4. We represent the number of nodes and the number of arcs in each cluster. Table 4 clearly demonstrates that the graph partitioning "result" keeps chunking as k is increased. After finding out the clustering accuracy of our system we analyse each cluster. There are 28 unique links in the website and they have a total visitation count of 7290. Using the Elbow Method (Trupti et al. [2013]) estimates that k = 9 is the optimal partitioning.

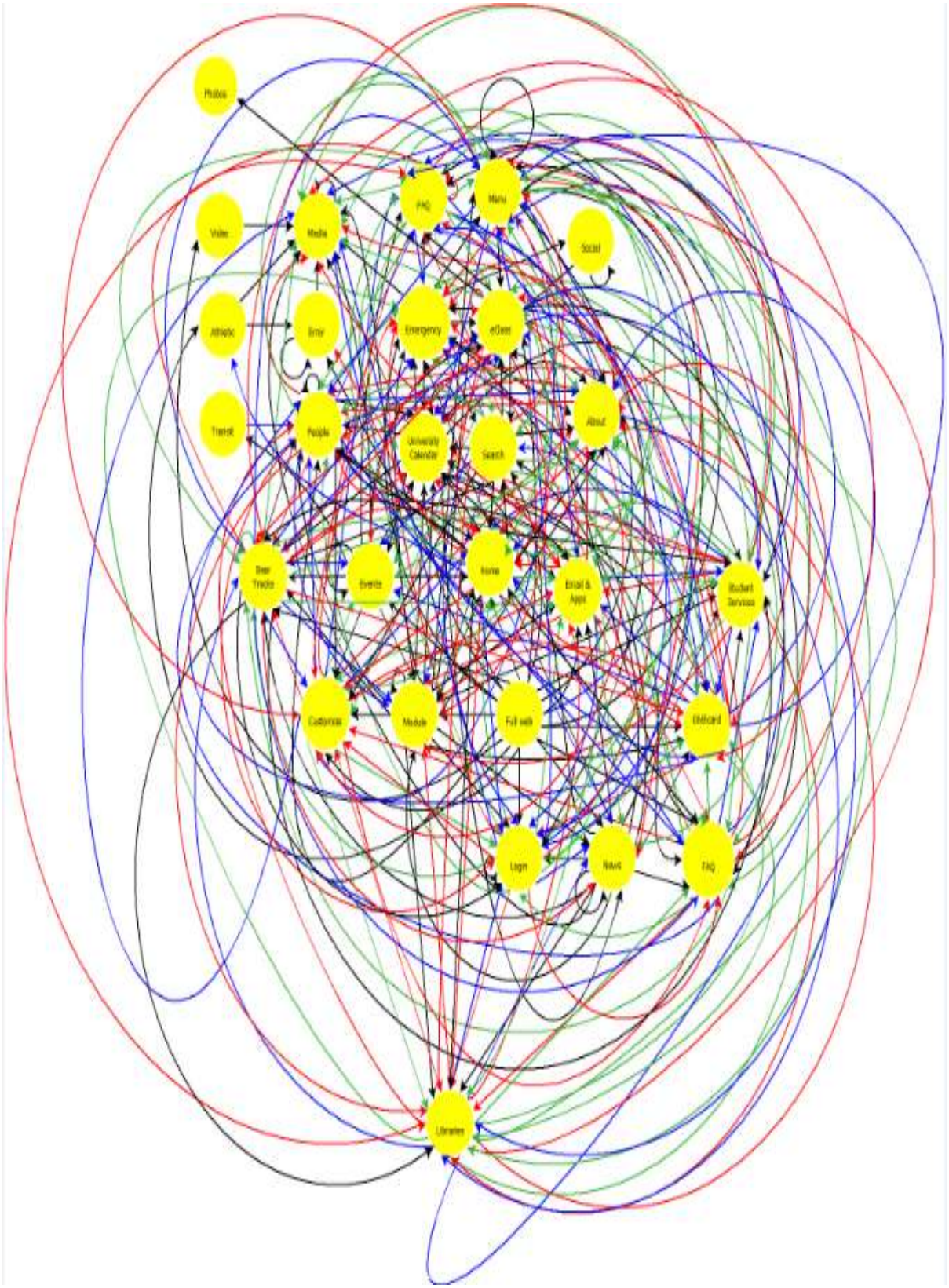
Photos
Video
Media
FAQ
Menu
Social
Athletix
Error
Emergency
eClass
Travel
People
University Calendar
Search
About
Bear Tracks
Events
Home
Email & Apps
Student Services
Calendar
Module
Full web
Offices
Login
News
FAQ
Libraries

Figure. 2 DTMC of University of Alberta Website

Figure. 3 Evaluation of clustering results

Table. 4 Basic information of each cluster

Cluster number	Number of nodes in each cluster	Number of Arcs in each cluster
2	11 and 17	56 and 300
3	11, 7 and 10	56, 114 and 186
4	1, 10, 7 and 10	1, 55, 114 and 186
5	1, 10, 5, 6 and 6	1, 55, 79, 107 and 114
6	1, 5, 5, 5, 6 and 6	1, 8, 47, 79, 107 and 114
7	1, 3, 3, 4, 5, 6 and 6	1, 4, 6, 45, 79, 107 and 114
8	1, 3, 3, 4, 2, 5, 4, and 6	1, 4, 6, 45, 30, 84, 72 and 114
9	1, 3, 3, 4, 2, 3, 4, 3 and 5	1, 4, 6, 45, 30, 49, 70, 55 and 96

In Table 5 we represent the activity summary of each cluster. In general, the links such as “Home”, “Search”, and “Menu” usually have the majority share of visits as they help users navigate through the site.

Table. 5 Description on each cluster

Cluster label	Proportion of dataset	Number of links	Summary of activity in the cluster
1	0.27	20	Users have interest on the photos related to the University such as photos of classrooms, Seminars, Workshops etc.
2	1.10	80	Users have interest on some basic information’s of the transit to and from the university or in the athletic facilities at the university. They are also interested in videos of the seminars, workshops etc.
3	3.98	290	Users are interested on social networking at the University. They are also interested on some financial services such as student loans, Graduate student’s association bursary etc.
4	9.62	700	Most of the users of this cluster browse the University website via mobile devices and are interested on the current news and events of the University.

5	8.78	640	Most of the users of this cluster are students of the University. They make their schedule of group project meetings, presentations, tutorials, labs etc.
6	13.72	1000	Most of the users of this cluster are students, staff or faculty members of the University. They customize their experience such as residence experience, profiles etc. and they have interested on University library.
7	19.20	1400	Most of the users of this cluster have a campus community ID (CCID). They are interested on different types of Student Services and use the University web mail for communication regularly.
8	15.23	1110	Users have interest in the University, Faculties, Departments and people of the University. They have also check the frequently asked questions. They are also interested on admission processes. So, they check the faculty list, try to contact with the faculty members, ask questions about their admission process.
9	28.12	2050	Most of the users of this cluster have a campus community ID (CCID). Most of the users are involved with course work. So, they visit the eClass regularly. Some of the users are the employee of the University. So, they visit the employee links frequently.

This is true in case of the current clustering system considered. Cluster 9 consists of these 5 links and it has 2050 visits which is about 28% of the total visits. Therefore, this table helps in understanding the user behaviour clearly in the web system considered. Lower percentages of visits indicate that the links in those clusters don't serve a purpose of majority of users. As an example, the links that require the usage of CCID, are restricted to a particular group of users, they have a lower visitation count when compared to other links that have no restrictions.

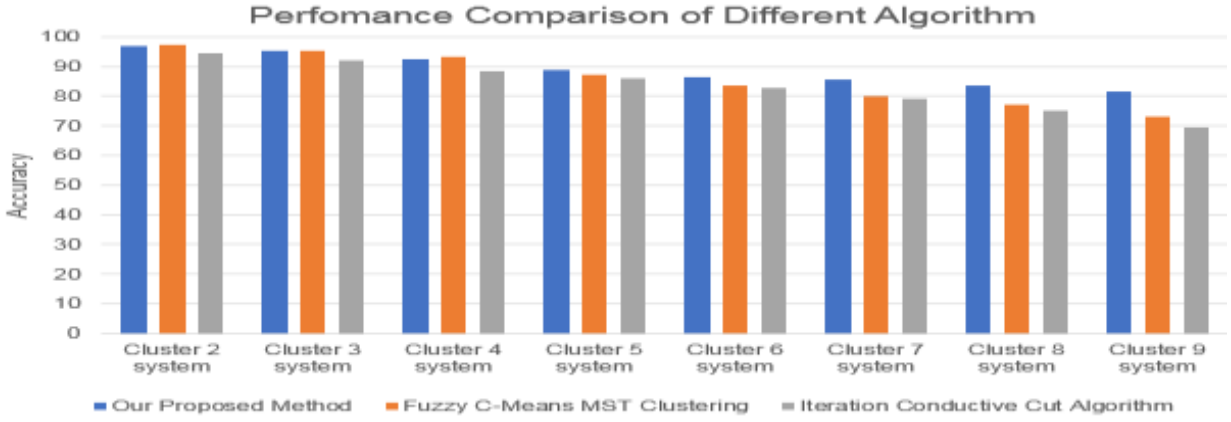

Figure. 4 Performance comparison of different clustering algorithm

For evaluating our clustering, we compare it with two renowned graph based clustering algorithms: The fuzzy C-Means MST Clustering Algorithm and The Iterative Conductive Cut Algorithm. The Fuzzy C-means MST clustering was improved by Foggia et al. [2007]. The Fuzzy C-Means MST Clustering algorithm starts with construction of a complete graph. Fuzzy Clustering can belong to more than one cluster with each data. Cluster analysis includes transferring data points to clusters. Clusters can be recognized with similarity measures like connectivity, intensity and distance. Different similarity measures can be chosen based on the data of the application. On the other hand, the Iterative conductive cut algorithm, proposed by Kannan et al. [2000] works in an ordered way. It initiates with one cluster consisting of the whole graph and it tries to divide the cluster into two. Cluster conductance is used as the measurement to grade the opportunity of the split. A clustering is considered superior if the conductance is lower. There is exponential time complexity when it comes to the search split of the minimizing the conductance. In figure

4, we present a performance comparison between three algorithms, Fuzzy C-Means MST Clustering algorithm, and The Iteration Conductive Cut Method and our proposed method. For comparing the algorithms, we use the accuracy of the derived clustering. We start with the cluster 2 system and end with the cluster 9 system. We do a simple statistical analysis for finding which method has the highest accuracy. Accuracy data of 8 samples are calculated using the 3 methods, and Mann-Whitney U test is performed to compare their performances.

Mann-Whitney U test, also called rank sum test, is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less or greater than a randomly selected value from a second sample. The 3 methods were compared in pairs, and the hypothesis was that they have equal accuracy (equal median for this test) and set 0.05 as the significance level as it is mostly used value in this analysis. If this hypothesis is true, the probability we have the sample results like what we get for our proposed method and the Fuzzy C-Means MST Clustering is 0.5992, much larger than 0.05. So, we cannot decline the statement that our proposed model is superior to Fuzzy C-Means MST Clustering. Similarly, if the hypothesis is true, the probability we get results like what we get for our proposed method and The Iteration Conductance Cutting Algorithm is 0.1949, also larger than 0.05. So, we cannot reject the hypothesis either. The p-value of the test for Fuzzy C-Means MST Clustering and The Iteration Conductance Cutting Algorithm is 0.5737, which was still much greater than 0.05. Thus, we cannot decide which one method has better accuracy than another one at 0.05 significance level for the data collected using Mann-Whitney U test. Therefore, from Mann-Whitney U test, we cannot detect significant differences of accuracy between any pairs among the 3 methods.

4.5 User behaviour pattern results

Due to lack of space we have just represented the unique navigation paths of the user. The other paths are the subset of these unique paths. We represent AND by "+" sign and OR by "/" sign at our automatic generated navigation patterns. Table 6 shows the clustering of navigation paths. The table indicates that the cluster 1 has the lowest number of user sequences while cluster 9 has the highest.

Table. 6 Clustering of navigation paths

Cluster label	Unique sequences of that cluster
1	Home (+ /) Photos+ Home
2	Home+ Video+ Media (+/) Media Home+ Transit+ People Home+ Athletic+ Media/ Error (+/) Media/ Error
3	Home+ Social (+/) Social+ eClass Home+ Error (+/) Error+ Media Emergency+ University Calendar (+/) Search (+/) eClass (+/) FAQ (+/) Menu (+/) Media (+/) Home (+/) Social (+/) About
4	Events+ Module (+/) Customize (+/) University Calendar (/+) eClass (+/) Events (+/) FAQ (+/) Media (+/) Transit (+/) Bear Tracks (+/) Home + Media Events+ Module (+/) Customize (+/) Email & Apps (+/) University Calendar (/+) Emergency (+/) Search (+/) ONEcard (+/) EClass (+/) Events (+/) Login (+/) FAQ (+/) News (+/) Menu (+/) Media (+/) Libraries (+/) Transit (+/) Bear Tracks (+/) Maps (+/) Home (+/) Student Services + News Media+ Module (+/) Customize (+/) University Calendar (+/) eClass (+/) Events (+/) FAQ (+/) Media (+/) Transit (+/) Bear Tracks (+/) Home (+/) + Events Media (+/) Media+ News News+ Module (+/) Email & Apps (+/) University Calendar (/+) Emergency (+/) Search (+/) ONEcard (+/) eClass (+/) Login (+/) News (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Maps (+/) Home (+/) Student Services+ Media
5	Module+ Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Home (+/) About (+/) Student Services+ University Calendar

	University Calendar+ Email & Apps (+/) University Calendar (+/) Emergency (+/) ONEcard (+/) Events (+/) News (+/) People (+/) Media (+/) Bear Tracks (+/) Maps (+/) About (+/) Student Services (+/) University Calendar
6	Libraries+ Module (+/) Customize (+/) Email & Apps (+/) Emergency (+/) Full web (+/) eClass (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Home (+/) About (+/ Student Services +Maps Libraries+ Module (+/) Email & Apps (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Maps (+/) Home (+/) About (+/) Student Services + Customize
7	Student Services+ Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) Search (+/) eClass + ONEcard Student Services+ Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) EClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services + Login Student Services +Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ Email & Apps Email & Apps + Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) Search (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ ONEcard Email & Apps+ Customize (+/) Email & Apps (+/) University Calendar (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home + Student Services Email & Apps + University Calendar (+/) Emergency (+/) Search (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Bear Tracks (+/) Maps (+/) Home (+/) Error (+/) About (+/) Student Services + Login
8	FAQ+ Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Error (+/) About (+/) Student Services+ About FAQ + Module (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ People About + Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ FAQ About + eClass (+/) ONEcard (+/) Events (+/) Login (+/) Menu (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ people People + Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) News (+/) Menu (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) About (+/) Student Services+ FAQ
9	Search + Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services+ Menu/ Menu/ Student Services/ Bear Tracks/ Home eClass + Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Menu (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) Home (+/) Social (+/) Error (+/) About (+/) Student Services

	+ Search/ Menu/ Bear Tracks/ Home/ Students services Menu+ Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) eClass (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Bear Tracks (+/) Maps (+/) About (+/) Student Services+ Search/ eClass/ Bear Tracks/ Home Bear Tracks (+/) Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Media (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) (+/) About (+/) Student Services+ Search/ Menu/ eClass/ Home Home (+/) Module (+/) Customize (+/) Email & Apps (+/) University Calendar (+/) Emergency (+/) ONEcard (+/) Events (+/) Login (+/) FAQ (+/) News (+/) People (+/) Libraries (+/) Video (+/) Transit (+/) Athletic (+/) Social (+/) Error (+/) About (+/) Student Services + Menu/ eClass/ Bear Tracks/ Search

Table 7 also shows nine clusters we obtain, this time indexed by which users exist in which cluster. We grant a unique number (value) to all 600 users in the system. By using this number, we can represent the members of each cluster. All of the users are not assigned to these clusters because they don't belong to any of the below. Navigation patterns of this clusters are automatically generated in our proposed system. By performing an analysis of the navigation patterns of the 600 users, we can find that 490 are members of any clusters so our automatically generated patterns cross with their navigation patterns. We manually labelled the all population of the log file in cluster 9 system. Then we conclude that 490 of the user are a member of a cluster. The remaining 110 users doesn't match with any cluster. So, this is the proof that the clustering accuracy of the algorithm is 81.67.

Table. 7 Clustering users of the website

Cluster label	Member	Common links of that cluster
1	132, 291	Photos
2	66, 99, 120, 122, 153, 292	Video, Transit, Athletic
3	72-73, 98, 105, 113, 118-119, 150, 169-170, 205, 226, 411-415, 437-438, 440	Social, Error, Emergency
4	21, 44, 67-68, 74, 97, 104, 114, 121, 151-152, 158-159, 160, 171, 206-208, 222-223, 293- 296, 311-320, 421-428, 441, 448, 453, 545, 460	Events, Full web, Media, News
5	20, 23-26, 69, 72, 75, 101-103, 115-116, 143-145, 161-163, 182, 225, 297-298, 305-310, 391-392, 416- 420, 429- 436	University Calendar, Module
6	18, 19, 22, 45- 47, 70- 71, 81- 82, 100, 117, 123-126, 146-149, 164-165, 183-185, 224, 321- 330, 393- 400	Customize, Maps, Libraries
7	15-16, 29-30 43, 84, 86, 91-92, 107-108, 154, 186-189, 196-198, 201, 212, 221, 227, 230-232, 251, 257, 258-260, 266-267, 341-360, 521, 523, 525-548, 586-595	ONEcard, Student Services, Login, Email & Apps.
8	17, 27, 28, 31, 41, 85, 106, 190-192, 202-204, 209-211, 219-220, 228-229, 252-256, 261-262, 265, 361-370, 549-560, 571- 585	FAQ, About, people
9	1-4, 6, 8-10, 13-14, 48-49, 61-62, 65, 93-96, 109-112, 127-130, 141-142, 155-157, 166-168, 193-195, 213-218, 268-290, 371-373, 375-377,380-384, 386- 390, 461-520, 596, 598-600	Menu, Search, eClass, Bear Tracks, Home

5. CONCLUSION

Web usage mining is a growing field of interest for professionals from literally every field. Large number of research projects is conducted in this domain and the focus is mainly in analyzing the user behavior pattern using clickstream datasets. Commercial enterprises will benefit a lot from the user behavior patterns by developing a selective approach to each customer without any manual supervision. This promises to raise business intelligence to a new level and adds competitive advantages to the business of the firm. In this research, we have constructed a group

of user behavior patterns generated from the anonymous clickstream data. The main advantage of our pattern discovery process is that, the patterns do not depend on any personal data about the site users, but instead, it depends on their aggregated browsing patterns. This eliminates the assumptions regarding the user categories. In this study, we use an economically available source of information log files. The principal findings of this research include:

- We name the unique links of the website as the atomic proposition at our study. Our system is dynamic and when the website performs an update, we can also derive the new unique links.
- We create a directed graph model of the website with the nodes as the unique website links and edges as the transition probability of going from one link to another. Our automated system can generate this model for any website using the log server file of that website.
- We cluster the links based on the user frequency. This will help in developing a new website or modifying an existing website in such a way that the most frequently visited links are made easily accessible for the users. Our proposed system does this for any given website.
- We use user behaviour profiling to find out the percentage of visits for every unique link in the website. In other words, this distributes the total user interests in a more appropriate manner and facilitates in decision making.

However, the principal achievement is that we derive clusters of navigation patterns from user interaction. This can play a very important role in the research of website personalization. The previous work related to this research can only find highly sought out links and the lowest sought out links of the website. But our new proposed system can find out the most used navigation patterns which is beneficial for both users and developers in web personalization. The clustering of navigation patterns can improve the quality of personalized web recommendations. This helps to predict which unique links are most likely to be visited next by the current users. We generate the best navigation patterns of the current users and as the recommendation, the links of these pages will then be inserted into the currently requested page dynamically for display. This will assist users to access their favourite and required information efficiently. Besides this, the clustering of navigation patterns is a very strong guide to organize the contents of the sites and webmasters can update a website in terms of the desires of users by using this. As an example, the necessary links of the website can be "added" together, which do not share the same topics at first sight, but they were visited one after another by a large number of users. Moreover, the pages or links that "collected" a large number of clicks can be highlighted, while pages which were not visited for a period of time can be moved or discarded. Therefore, website management becomes dynamic and proactive. As a result, the visitors of the website will be attracted to become consumers or regular users of the website.

REFERENCES

1. Haibin Liu and Vlado Keselj. 2007. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data and Knowledge Engineering.61 (2007), 304-330.
2. Miao Wan, Arne Jonnson, Cong Wang, Lixiang Li and Yixiang Yang. 2012. A Random Indexing Approach for Web User Clustering and Web Prefetching. Springer-Verlag Berlin Heidelberg. LNAI 7104. pp 40-52.
3. Istvan K. Nagy and Csaba Gaspar-Papanek. 2009. User Behaviour Analysis Based on Time Spent on Web Pages. Springer-Verlag Berlin Heidelberg. SCI 172. pp 117-135.
4. Arindam Banerjee and JoydeepGhosh. 2009. Clickstream Clustering using Weighted Longest Common Subsequences. !6th European Conference and Artificial Intelligence. August 22-27.
5. Matthias Schur, Andreas Roth and Andreas Zeller. 2013. Mining Behaviour Models from Enterprise Web Applications. ESEC/FSC (2007), 422-432.
6. James Pitkow and Peter Pirolli. 1999. Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. Proceedings of USITS' 99: The 2nd USENIX Symposium on Internet Technologies & Systems.
7. Yunjuan Xie and Vir V. Phoha. 2001. Web User Clustering from Access Log using Belief Function. k-CAP 2001, pp 202-208.
8. Miao Wan, Lixiang Li and Jinghua Xiao. 2010. CAS Based Clustering Algorithm for Web User Non-Linear Dyn .61 (2010), 347-361.
9. Jianbo Shi and Jitendra Malik. 2007. Normalized Cuts and Image Segmentation. Computer Vision and Pattern Recognition. 731-737.

10. Chris Giannella and Eric Bloedorn. 2015. Spectral Malware Behavior Clustering. International Conference on Intelligence and Security Informatics.2015, 7-12.
11. B. Zhou, S.C. Hui and K. Chang. An intelligent recommender system using sequential web access patterns. International Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems, Singapore, 2004, pp. 1–3.
12. D. S. Phatak and R. Mulvaney. Clustering for personalized mobile web usage. International Proceedings of the IEEE FUZZ '02, Hawaii, USA,2002. pp. 705–710.
13. H. Dai and B. Mobasher. A road mapMaps to more effective web personalization: Integrating domain knowledge with the web usage mining. International Conference on Internet Computing, 2003, pp. 58–64.
14. B. Mobasher, H. Dai, T. Luo, Y. Sun and J. Zhu. Integrating web usage and content mining for more effective personalization. Proceedings of the First International Conference on Electronic Commerce and Web Technologies, Springer-Verlag, 2000. Pp. 165-170.
15. V. Keselj, F. Peng, N. Cercone and C. Thomas. N-gram-based author profiles for authorship attribution. Proceedings of the Conference Pacific Association for Computational Linguistics, Nova Scotia, Canada, 2003.
16. R. Cooley, B. Mobasher and J. Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems 1 (1) (1999) 5–32.
17. P.K. Chan. A non-invasive learning approach to building web user profiles, in: Workshop on Web usage analysis and user profiling. Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, 1999.
18. Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, Ben Y. Zhao. Unsupervised Clickstream Clustering for User Behavior Analysis. Mining Human Behaviors, CHI 2016, San Jose, CA, USA, page 225-236.
19. Gang Wang, F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2009. Characterizing User Behavior in Online Social Networks. In Proc. of IMC.
20. L. Lu, M. Dunham, and Y. Meng. 2005. Mining significant usage patterns from clickstream data. In Proc. of WebKDD.
21. J. Y. Park, N. O'Hare, R. Schifanella, A. Jaimes, and C. Chung. 2015. A Large-Scale Study of User Image Search Behavior on the Web. In Proc. of CHI
22. N. Sadagopan and J. Li. 2008. Characterizing Typical and Atypical User Sessions in Clickstreams. In Proc. of WWW.
23. Q. Su and L. Chen. 2015. A method for discovering clusters of e-commerce interest patterns using click-stream data. ECRA 14, 1 (2015), 1–13..
24. I. Ting, C. Kimble, and D. Kudenko. 2005. UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a Web Site's Design. In Proc. of ICWI.
25. Jan-Willem van Dam, Michel van de Velden, Online profiling and clustering of Facebook users, Decision Support Systems 70 (2015) 60–72.
26. S. Ho, D. Bodoff, K. Tam, Timing of adaptive web personalization and its effects on online consumer behavior, Information Systems Research 22 (2011) 660–679.
27. K. P. Wiedmann, H. Buxel, G. Walsh, Customer profiling in e-commerce: methodological aspects and challenges, Journal of Database Marketing 9 (2) (2002) 170–184.
28. R. Rishika, A. Kumar, R. Janakiraman, R. Bezawada, The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation, Information Systems Research 24 (2013).
29. N. Park, S. Lee, J.H. Kim, Individuals' personal network characteristics and patterns of Facebook use: a social network approach, Computers in Human Behavior 28 (2012) 1700–1707.
30. C. Xu, C. Du, G.F. Zhao, S. Yu. A novel model for user clicks identification based on hidden semi-Markov. Journal of Network and Computer Applications 36 (2013) 791–798.
31. Arumugam G, Suguna S. Optimal algorithms for generation of user session sequences using server side web user logs. In: Proceedings of the N2S '09, Paris, France, 24–26 June 2009. P. 1–6.
32. Chapelle O, Zhang Y. A dynamic Bayesian network click model for web search ranking. In: Proceedings of the 18th international world wide web conference; 2009. P. 1–10.

33. Chen Zhixiang, Fu Ada Wai-Chee, Tong Chi-Hung. Optimal algorithms for finding user web access session from very large web logs. Journal of World Wide Web: Internet and Information Systems 2003; 6:259–79 [Springer].
34. T. M. Kodinariya, P. R, Makwana. Review on determining number of clusters in K-means clustering. International Journal of Advanced Research in Computer Science and Management Studies, 2013, 1(6), pp. 90-95.
35. P. Foggia, G. Percannella, C. Sansone, M. Vento. Assessing the performance of a graph-based clustering algorithm. in: F. Escolano, M. Vento (Eds.),. Lecture Notes in Computer Science. vol. 4538, Springer-Verlag. Berlin, 2007. pp. 215– 227.
36. R. Kannan, S. Vampala, A. Vetta. On clustering: good, bad and spectral. Foundations of Computer Science (2000) 367–378.

AUTHORS PROFILE