Discovering Communities for Web Usage Mining Systems

Yacine Slimani*, Abdelouahab Moussaoui, Yves Lechevallier and Ahlem Drif

University Ferhat Abbas Setif 1,
Department of Computer Science,
Laboratory of Intelligent Systems,
Setif 19000, Algeria
E-mail: slimani_y09@univ-setif.dz
E-mail: moussaoui.abdel@gmail.com
E-mail: yves.lechevallier@inria.fr
E-mail: adrif.univsetif@gmail.com
*Corresponding author

Abstract: Discovering the community structure in the context of web usage mining has been addressed in many different ways. In this paper, we present a new method for detecting communities using Markov chains based on the set of frequent motifs. The basic idea is to analyze the occurrence probability of different frequent sequences during different user sessions in order to extract the communities that describe the users behavior. The proposed method is successfully applied on the web site of Setif university.

Keywords: Web usage mining; Community detection; Complex networks; Markov chains; Quality function.

Reference to this paper should be made as follows: Slimani, Y., Moussaoui, A., Lechevallier, Y. and Drif, A. (xxxx) 'Discovering communitie's for Web Usage Mining Systems', *Int. J. Big Data Intelligence*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Yacine SLIMANI received his computer science engineer degree in 1997 from Ferhat Abbas University, Algeria. He also received a Magister degree in Computer Science in 2006 from Ferhat Abbas University. He is a Researcher and Assistant professor at the Department of technologies since 2006. He is also a Ph.D. student and a member at the Laboratory of Intelligent Systems (LIS) at Ferhat Abbas University Setif 1. His areas of interests include data mining, web mining and Artificial Intelligence.

Abdelouahab MOUSSAOUI is Professor at Ferhat Abbas University. He received his BSc in Computer Science in 1990 from the Department of Computer Science from the University of Science and Technology of Houari Boumedienne Algeria. He also received and MSc in Space Engineering in 1991 from University of Science and Technology of Oran (USTO). He received also an MSc degree in Machine Learning from Reims University (France) since 1992 and Master's degree in Computer Science in 1995 from University of Sidi Bel-abbes, Algeria and PhD degree in Computer Science from Ferhat Abbas University, Algeria where he obtains a status of full-professor in Computer Science. He is IEEE Member and AJIT, IJMMIA and IJSC Referee. His researches are in of pattern recognitions algorithm, complex data mining and medical image analysis.

Yves LECHEVALLIER joined the INRIA in 1976 where he was engaged in the project of Clustering and Pattern Recognition. Since 1988 he has been teaching Clustering, Neural Network and Data Mining at the University of PARIS IX CNAM and ENSAE. He specializes in Mathematical Statistics, Applied Statistics, Data Analysis and Classification. His current research interests include Clustering Algorithms (Dynamic Clustering Method, Kohonen Maps and Divisive Clustering Method); Discrimination Problems and Decision Tree Methods and Build an efficient Neural Network by Classification Tree.

Ahlem DRIF received an engineering degree in computer science from University of Setif in Algeria (UFAS) in 2002 and magister degree in computer science in 2006. Currently, she is Assistant Professor and has 9 years teaching experience. She is a member at the Laboratory of Network and Distributed System (LRSD) at the UFAS and has 10+ papers in international conferences. Her main research interests lie in the areas of social networks, community discovery methods, mobile social networking , dynamic topology, ad hoc networks.

Introduction

Web Usage Mining is a process of extraction and analysis of data allowing the detection of the users' navigational behavior on a web site [35]. This task is based on data mining, in which many steps are necessaries for the realization of the entire process [6]. Our work is divided in we divide in three main steps: preprocessing, pattern discovery and pattern analysis. In fact, various works have described the techniques of data preprocessing and user modeling in order to extract the information [26]. Data preprocessing includes data cleaning, normalization, transformation, filtering and summarization. The discovery and analysis phases use a knowledge method to identify groups of users with similar behaviour for which personalized versions of the web site may be created. Therefore, the main contributions of the proposed work lies in the development of a novel Markov Chain-based method, and detecting the communities based on sequences of users, instead of the hyperlink connections among webpages. To this end, we have proposed a five-phase method, in which a stochastic model based on Markov chain is devised to measure the transition probability starting from and arriving at a page. And the derived stochastic process is used to compute the distance between nodes so that the hidden community among webpages can be distinguished. Experiments conducted on a small-scale web system, Ferhat Abbas Setif university website, show the reasonable results based on modularity and inner-outer community edges.

The remaining of the paper is organized as follows : section 1 explains our motivation behind the proposed method for the improvement of the web design. Section 2 describes the intended web usage data preprocessing, which help to increase the quality of the data obtained at the end of the preprocessing step. Our method is presented is section 3. The experiments results and evaluation are described in section 4; and finally, a general conclusion is presented in section 5.

1 Context and motivations

The objectives of a web usage mining are to improve the web site architecture, system performance analysis, understanding the reactions and motivations of the users, and adaptive construction of the web sites [21, 32, 36]. Several methods have been proposed for the personalization of a web site to discover interesting usage patterns from web data and better serve the needs of web-based applications. An overview can be found in [10]. Many methods are based on individual's past behavior or on the items of interest. Thus, these approaches are represented as a recommender systems [16, 20, 37], which make recommendations to the users. Collaborative filtering has also been known to be the most successful recommendation techniques that apply a quality of prediction [3]. In addition, some approaches adapt the content of the web site to visitors in order to deliver personalized information, which develops the potential of an intelligent web [5].

The web usage mining process is divided in three main steps: preprocessing, pattern discovery and pattern analysis. In preprocessing step, many kinds of data mining algorithms can be performed on the preprocessed data such as association rules mining [29], clustering [4] and K-Means clustering [30], sequential patterns [8, 24]. In pattern discovery step, several approaches inspired from community detection methods were applied in order to deal with the incapacity of numerous mining methods and to provide a significant community structure [23]. In fact, the communities are groups of nodes, which share probably a common proprieties and/or similar functions. The communities may be corresponding, for example, two groups of web pages accessible over the Internet that have a same subject [13].

Motivated by relevance of the extracted communities in web mining, we present a new community discovery method adapted to the identification of web user's navigational behavior from the log files of the web server. In the literature, there are several existing studies on detecting communities using sequence data. In the work [39], the authors have proposed a fast algorithm based on community detection from a graph, and it is used to transforms the (l, d) motif search in sequences under the more biologically-relevant ZOMOPS constraint (zero, one or multiple occurrences of the motif instances per sequence), where l is the length of a motif and d is the maximum number of mutations between a motif instance and the motif itself, to focus on the discovery of dense subgraphs within a graph. Lu et al [42] proposed a community detection algorithm based on the similarity sequence whereas the similarities of nodes are sorted in descending order to get a sequence. Then, pairs of nodes are merged according to the sequence to construct a preliminary community structure and the agglomerative clustering process is carried out to get the optimal community structure. With our work, we search for frequents motifs observing all transition from page v_i to page w_i during a session for each user and estimate the state transition probabilities using a stochastic process to discover dense subgraph from the generated graph of browsing behavior.

The proposed method create a directed and weighted graph and applied a hybrid method that combine the simplicity of frequent sequences, the efficiency of Markov chains and the robustness of the community detection methods using a similarity metric.

2 Data preprocessing

The aims of the preprocessing step in a WUM process are roughly to convert the raw log file into a set of transactions to reduce the quantity of data being analyzed and, at the same time, enhance its quality [38]. First of all, we give the main terms used in this work [35] :

- A user is a person using a Web browser.
- A Web resource is a resource accessible through any version of the HTTP protocol.
- A Web page is the set of data constituting one or several Web resources that can be identified by an URL.
- A user session consists in a delimited number of a user's explicit Web requests across one or more Web servers.
- A page view occurs at a specific moment in time, when a Web browser displays a Web page.
- A visit or a navigation activity (browsing behaviour) represents a subset of consecutive page views from a user session occurring close enough (measured by means of a time threshold or a semantical distance between pages).

In our work, we have performed a preprocessing step on the Web logs using the methodology defined in [31]. In the first step, the preprocessing includes the data transformation from log files followed by a data cleaning module, which is used to remove the irrelevant items stored in the log files that may not be useful for our analysis. Then, the data structuration module is used to regroup the requests of the log file in user sessions and split all the pages accessed by a user into different sessions which allow us to perform a data filtering step to remove the less requested pages and retain only the most requested ones. In the last step, a summary data is generated based on the data structure in order to select only the interesting information.

In this section we will discuss the preprocessing results in brief. Our preprocessing method has been tested on log files stored in the server of the studied web site (Algeria) available at the URL *http://www.univ-setif.dz*. The treated file covers the site activities during the period from 18-12-2011 at 04:04:57 to 19-01-2012 at 08:47:04. Figure 1 shows the summary of the preliminary analysis of log files. After data transformation, we have obtained 1 708 385 requests. Then, we have applied a data cleaning step to maintain only actions related to users' browsing behavior and to eliminate the following requests:

1. Method different from "GET": in general, the requests containing a value different from "GET" are not explicit requests of the users, but they often relate to accesses with CGI, of the visits of robots, etc.
2. Failed and corrupted requests: these requests are represented by records containing a HTTP error code. A status with value different from 200 represents a failed request (e.g. a status of 404 indicates that the requested file was not found at the expected location).
3. Requests for multimedia objects: in the HTTP protocol, an access request is carried out for every file, image, multimedia object embedded in a requested Web page. As a

consequence, a single request for a Web page may often produces several entries in the log file that corresponds to files automatically downloaded without an explicit request of the same user. The requests of this type of files can be easily identified since they contain a particular URL name suffix, such as gif, jpeg, jpg, and so on.

4. Requests originated by Web robots: log files contain some number of records corresponding to requests originated by Web robots. Web robots are programs that automatically download complete Web sites by following every hyperlink on every page within the site in order to update the index of search engine.

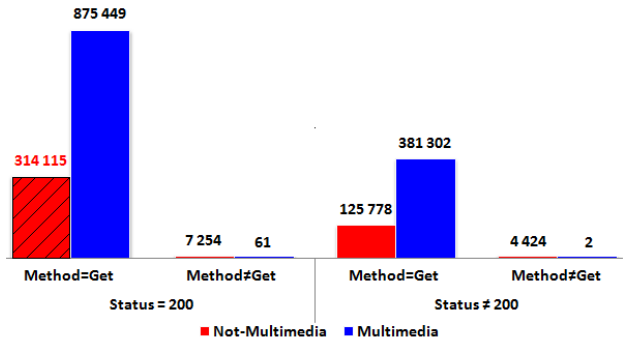

Figure 1 Summary of log files pre-processing results.

After data cleaning step, we have obtained 314 115 (i.e. 18,38%) valid requests for the 27 520 users who access to 23 872 pages.

Table 1 shows the results of sessions and robot identification. In this step, we have applied the structuration algorithm [22] to determine the sessions taking account of maximum elapsed time $\Delta Max_t = 30min$ between two consecutive accesses [6, 25], the results is 50 131 sessions for 314 115 navigations. After the identification of user sessions, we have identified web robots' requests of search engines and web crawler [34]. The minimum elapsed time between consecutive accesses is also fixed at $\Delta Min_t = 10seconds$. So, the result is 12 297 sessions through 147 711 navigations.

Table 1 Summary of identification of sessions and robots.

Category		User	Page	Session	Navigation
Total before identification robots		27 520	23 872	50 131	**314 115**
Identification method of Robots	Ip	303	8 362	5 083	80 484
	URL	783	19 128	14 543	124 519
	Agent	1 369	20 223	18 496	137 073
Total robots		1 438	20 266	19 114	138 434
Total after identification robots		26 082	7 259	31 017	175 681
Total after removal of ΔMin_t Req.		10 899	5 753	12 297	**147 711**

3 Community detection for Web Usage Mining Systems

3.1 Problem statement

The web usage mining aims at analyzing and discovering user's interesting patterns that are necessary for web administrator and recommendation system. In this paper, we are interested in detecting the communities of users using their Web browsing behaviors instead of the hyperlink graph structure of web pages which is more useful to find out the common interest of a group of users. For this reason, we establish the Markov chain model that best represents the similar sequences of navigation of users. So, the derived stochastic process is used to further compute the similarity between different users navigation session and define quality function to compare partitions. In addition, we propose a Discovery Community algorithm based on Markov Chain (DCMMC) which induces the extraction of the communities patterns. This allow us to understand what relevant information that users would expect to find through the website and find the most effective logical structure for the Web site.

Let S be a set of n_s sessions, in which users access at n web pages. We define:

- $S = s_1, s_2, ..., s_{n_s}$: a set of sessions navigations ;
- $V = v_1, v_2, ..., v_n$: a set of visited pages ;
- $S_i = (v_i^{(1)}, v_i^{(2)}, ..., v_i^{(n_i)})$: the sequences of visited pages during session S_i

The total number of visited pages n_v is :

$$n_v = \sum_{i=1}^{n_s} n_i \tag{1}$$

Where n_i is the number of visited pages during the session i. Assume, for example, that many users have accessed at pages $V = \{v_1, v_2, v_3, v_4, v_5\}$ during three sessions, then we obtain the following set of sessions:

$$S_1 = (v_1, v_2, v_3, v_4, v_2, v_3, v_5, v_1, v_4)$$

$$S_2 = (v_1, v_4, v_2, v_3, v_5, v_1, v_3, v_4)$$

$$S_3 = (v_2, v_3, v_4, v_2, v_3, v_5)$$

The method we are describing is based on a stochastic model which rely on the relevant information contained in the frequent pairs mining. In what follows, we describe its principal steps.

3.2 Step 1 : search for frequent motifs

In this phase, we model the user's access during sessions using a formal and comprehensive systematization for identification of frequent item sets. In general, a session base can be regarded as a sequence of pairs defining transition between pages [1].

Thus, we introduce the distinct sets of pairs that provide an easier time working in order to help in the search of frequents motifs and the creation of graph of browsing behaviour.

Formally, we define the pair (v_i, w_i) as a transition from page v_i to page w_i during a session $i \in [1..n_s]$.
Let M_i be the sequence of all pairs' transitions during the session i:

$$M_i = \Big((v_i^{(t)}, w_i^{(t)}) | w_i^{(t)} = v_i^{(t+1)}; \quad t = [1..n_i - 1]\Big) \tag{2}$$

It follows that the complete transitions which occur in all the sessions used to describe all the observed set of pairs sequences will be denoted :

$$M = \Big((v, w) | (v, w) \in U_{i=1}^{n_s} M_i\Big) \tag{3}$$

Therefore M has a cardinality $|M| = n_v - n_s$. Suppose that each pair (v, w) in sequence M has an occurrence number $n_v(v, w)$, so it indicates the potential information occurring during these n_s sessions and keeps any pair of pages that have been visited more than once, but not consecutively, it is given by :

$$n_v(v, w) = \sum_{(x,y) \in M} \delta_v\Big((v, w), (x, y)\Big) \tag{4}$$

Where δ_v is the Kronecker delta function, defined as :

$$\delta_v((v, w), (x, y)) = \begin{cases} 1 \text{ if } v = x \text{ and } w = y \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

To characterize the overall access from each page in the web site, we define the occurrence number of the page v, when it is a starter page as:

$$n_{out}(v) = \sum_{w \in V} n_v(v, w) \tag{6}$$

We can also define the occurrence number of the page w, when it is an arrival page as :

$$n_{in}(w) = \sum_{v \in V} n_v(v, w) \tag{7}$$

Our idea revolves around the basic principle of mining all frequent sets using the number of occurrence of pairs. We calculate the support of all pages for all pairs (starting page, arrival page) in session base. Therefore, we keep pages that have a support with respect to a minimal support threshold ($minsup$). The support of page v is its occurrence number in sessions divided by the total number of all sessions. Thus

$$Supp(v) = \frac{n_{out}(v) + n_{in}(v)}{2n_s} \tag{8}$$

The support of a pair (v, w) represents the support of k-item of second order ($k = 2$), it is the occurrence number of (v, w) in sessions divided by the total number of all sessions, so that

$$Supp(v, w) = \frac{n_v(v, w)}{n_s} \tag{9}$$

3.3 Step 2 : Graph creation

Let $G = (V, E)$ be a weighted and directed graph that represents the users navigation session, such :

V is the set of nodes

E is the set of edges.

E is the distinct pairs of M ($|E| = m \leq |M|$), defined as:

$$E = \Big\{(v, w)|(v, w) \in M\Big\} \tag{10}$$

We define the adjacency matrix A from the frequent pair of pages (v, w) :

$$A_{i,j} = n_v(v_i, v_j) \tag{11}$$

See Fig. 2 for a toy example in which we have computed the degree of nodes as strictly related to the frequency of accesses to pair of pages and we have determined the whole graph.

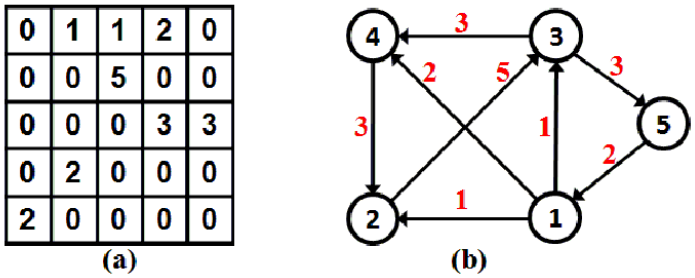

Figure 2 (a) adjacency Matrix - (b) Graph

3.4 Step 3 : Creation of stochastic model

Consider a random variable X observable on a Markov chain process. X represents the users' navigation from starting page $v^{(t)}$ to an ending page $w^{(t+1)}$. The observation of random sessions can be represented by the Markov process [11, 12], such that:

$$S_i = \Big\{(X^{(i,t)}, X^{(i,t+1)})|t = 1..n_i - 1 \quad , \quad (X^{(i,t)}, X^{(i,t+1)}) \in V\Big\} \tag{12}$$

The sample for all sessions is defined as:

$$\begin{aligned} E_x = \Big(& (X^{(1,1)}, X^{(1,2)}), (X^{(1,2)}, X^{(1,3)}), ..., (X^{(1,n_1-1)}, X^{(1,n_1)}), \\ & (X^{(2,1)}, X^{(2,2)}), (X^{(2,2)}, X^{(2,3)}), ..., (X^{(2,n_2-1)}, X^{(2,n_2)}), ... \\ & (X^{(n_s,1)}, X^{(n_s,2)}), (X^{(n_s,2)}, X^{(n_s,3)}), ..., (X^{(n_s,n_n-1)}, X^{(n_s,n_n)})\Big) \end{aligned} \tag{13}$$

The size of the sample E_x is :

$$n_E = \sum_{i=1}^{n_s} (n_i - 1) = \sum_{i=1}^{n_s} n_i - n_s = n_v - n_s \tag{14}$$

n_E represents the total number of transition during all the sessions, taking into account that there is no transition between session S_i and the next session S_{i+1}. It can be also calculated as follows:

$$n_E = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \tag{15}$$

The estimation of the transition from page v to page w is:

$$E_x = \left\{ (X^{(t)}, X^{(t+1)}) | t = 1..n_E \quad , \quad X^{(t)}, X^{(t+1)} \in V \right\} \tag{16}$$

The probability of an occurrence of a transition during all sessions is defined as:

$$P_E = \frac{1}{n_E} \tag{17}$$

In the case of first order Markov chain, the state transition probabilities don't depend on the all history of the process, such the preceding state is taken into account. So, the sate transition probabilities can be described as:

$$P[X^{(t)} = v \quad \text{and} \quad X^{(t+1)} = w] = \frac{A_{vw}}{n_E} \tag{18}$$

Such that :

$$P[X^{(t)} = v] = \frac{\sum\limits_{w=1}^{n} A_{v,w}}{n_E} \tag{19}$$

and

$$P[X^{(t+1)} = w] = \frac{\sum\limits_{v=1}^{n} A_{v,w}}{n_E} \tag{20}$$

Observe then that the Markov property is defined as:

$$P[X^{(t+1)} = w | X^{(t)} = v] = \frac{P[X^{(t)} = v \quad \text{and} \quad X^{(t+1)} = w]}{P[X^{(t)} = v]} = \frac{A_{v,w}}{\sum\limits_{w=1}^{n} A_{v,w}} \tag{21}$$

Then, we define the transition probability matrix on page w knowing that we are in the page v,$P = [P_{v,w}]$, so that:

$$P_{v,w} = \frac{A_{v,w}}{\sum\limits_{w=1}^{n} A_{v,w}} \quad \text{with} \quad \forall v : \sum_{w=1}^{n} P_{v,w} = 1 \tag{22}$$

We define also the probability vector that v is a starting page $P_{out} = [Pout_v]$, so that :

$$Pout_v = \frac{\sum_{w=1}^{n} A_{v,w}}{\sum_{v=1}^{n}\sum_{w=1}^{n} A_{v,w}} \tag{23}$$

The probability vector that w is an arrival page $P_{in} = [Pin_v]$ is given by :

$$Pin_w = \frac{\sum_{v=1}^{n} A_{v,w}}{\sum_{v=1}^{n}\sum_{w=1}^{n} A_{v,w}} \tag{24}$$

3.5 Step 4 : Search the steady state probability

Let $\pi^{(1)}$ be an initial probability vector of the navigational state of users in the time $t = 1$, we write it as:

$$\pi^{(1)} = \{\pi_i^{(1)}\}_{i=1..n} \tag{25}$$

We get :

$$\pi_i^{(1)} = P[X^{(1)} = i] \quad \text{where } \forall i \pi_i \geq 0 \quad , \quad \sum_{i=1}^{n} \pi_i = 1 \tag{26}$$

So, we have the probability vector of the state of navigation in the time $t + 1$:

$$\pi_i^{(t+1)} = P * \pi_i^{(t)} \tag{27}$$

There is a unique steady-state probability vector π^* that is the principal left eigenvector of P [17],[2]:

$$\lim_{t \to +\infty} \pi^t = \pi^* \tag{28}$$

Where π^* is the eigenvector of the eigenvalue ($\lambda = 1$). Then, we obtain a convergence to a stationary probability regardless of the initial state, we find that:

$$\pi_i^* = P * \pi_i^* \tag{29}$$

3.6 Step 5: Communities discovery

In this section, we define communities of webpages and discover similarities (or dissimilarities) between communities based on the distance between nodes that we have defined using the proposed Markov chain-based mechanism. Further, we have rewritten the modularity in terms of the transition probability.
A community of web page, denoted C_i, $\forall i \in 1..K$, is a subgraph of web pages that have the same usage (users that have accessed several times these web pages), in other word, we are looking for dense graphs whose internal links (inner edges) represent users who browse similar proximity web pages and external links (outer edges) represent the tendency of users to navigate through the web site to access other web pages with close similarities.

Many algorithms have been proposed to find a measure to extract similarities between communities in complex networks. [27] have used a measure of distance between nodes to identify structure similarities . An Euclidean distance between nodes have already been used based on the spectral properties of the Laplacien matrix of the graph. [7] have defined this measure and used it in a hierarchical clustering algorithm. [19] have observed the organization of complex systems at each level and defined a proximity concept in the hierarchy between all pairs of nodes. Here, we compute the distance between two pages v and w based on the probability to visit the others pages k, as the square root of the sum of the square of differences between probabilities of the two pages. We obtained a symmetric matrix D defined as:

$$d_{vw} = \sqrt{(\pi^*_v - \pi^*_w)^2} \tag{30}$$

Let the symmetric matrix $S = [s_{vw}]$, where s_{vw} measures the normalized similarity between pages v and w of the set V. Where $d = \max_{(v,w)\in V^2} d_{vw}$, we define S as:

$$s_{vw} = 1 - \frac{d_{vw}}{d} \tag{31}$$

To illustrate our idea, we present the results obtained from the graph (Fig. 2). We obtained the probability matrix P, and the two probability vectors of in-going P_{in} and out-going edges P_{out} of each node, the steady state probability vector $\pi^{(*)}$is also presented and finally the normalized similarity matrix S(Fig. 3).

P					P_{in}	$\pi^{(*)}$	S				
0	¼	¼	½	0	0.20	0.14	1.00	0.25	0.00	0.50	1.00
0	0	1	0	0	0.25	0.24	0.25	1.00	0.75	0.75	0.25
0	0	0	½	½	0.30	0.27	0.00	0.75	1.00	0.50	0.00
0	1	0	0	0	0.15	0.21	0.50	0.75	0.50	1.00	0.50
1	0	0	0	0	0.10	0.14	1.00	0.25	0.00	0.50	1.00

P_{out}	0.10	0.20	0.30	0.25	0.15

Figure 3 Example of stochastic and symetric matrix.

Modularity are used as a quality function and criterion to specify the best partition may be found. Then, the modularity can be written:

$$Q = \frac{1}{2 * m} \sum_{v,w} [P_{vw} - (n.\pi_v^*.\pi_w^*)]\delta(C_v, C_w) \tag{32}$$

such as :

$$\delta(C_v, C_w) = \begin{cases} 1 \text{ if } \{v, w\} \text{ are in the same community} \\ 0 \text{ else.} \end{cases}$$

We exploit S to reveal the community structure of a network using the algorithm defined in [7]. Thus, the proposed Discovery Community Method based on Markov Chain (DCMMC) optimizing an adequacy criterion J that measures the fitting between a partition of communities $P = (C_1, \ldots, C_K)$ and its prototypes $G = (G_1, \ldots, G_K)$, we write it as:

$$J(P, G) = \sum_{k=1}^{K} \sum_{v_i \in C_k} s_{(v_i, G_k)}, \tag{33}$$

Such that $s_{(v_i, G_k)}$ is the similarity between v_i and the prototype of his community G_k.

We call the DCMMC algorithm to find the best partition in communities that produce the optimal modularity as described in algorithm 1.

Data: S
$QMax \leftarrow 0$
$KMax \leftarrow 1$
for *k=1 to n* **do**
  Call $[P_k, G_k]$=DCMMC(S, k)
  Compute Q with formule (32)
  if $QMax < Q$ **then**
    $QMax \leftarrow Q$
    $KMax \leftarrow k$
  end
end

Algorithm 1: Finding the best communities

The DCMMC algorithm selects the best partition in the communities that maximizes the objective criterion J. The DCMMC algorithm is structured in three principal steps like representing in algorithm 2.

Data: S, k
$t \leftarrow 0$
Select k prototypes from centers: $G^{(0)} = (G_1^{(0)}, \ldots, G_k^{(0)})$
repeat
  Step 1: *Build the best partition*
  begin
    $t \leftarrow t$+1
    $test \leftarrow 0$
    $P^{(t)} \leftarrow P^{(t-1)}$
    for *m=1 to k* **do**
      forall $v_i \in C_m^{(t)}$ **do**
        find the new winning cluster $C_l^{(t)}$ such that :
        $l = \arg\max_{1 \leq h \leq k} s(v_i, G_h)$
        if $l \neq m$ **then**
          $test \leftarrow 1$
          $C_l^{(t)} \leftarrow C_l^{(t)} \cup \{v_i\}$
          $C_m^{(t)} \leftarrow C_m^{(t)} \setminus \{v_i\}$
        end
      end
    end
  end
  Step 2: *Find the best prototypes*
  begin
    for *m=1 to k* **do**
      Compute the prototype $G_m^{(t)} \in V$ of cluster $C_m^{(t)}$ according to:
      $G_m^{(t)} = \arg\max_{G \in C_m^{(t)}} \sum_{v_i \in C_m^{(t)}} s(v_i, G)$
    end
  end
until $test = 0$;
return P_k, G_k

Algorithm 2: Algorithm of communities discovery (DCMMC)

For the sake of argument let's use the graph presented in figure 2, the result of applying DCMMC on this graph is shown in figure 4. The algorithm correctly identifies three quite cohesive communities and the dendrogram provides the levels of partition of the graph nodes.

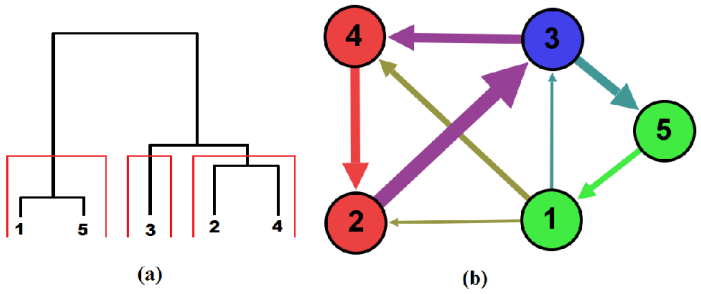

Figure 4 (a) Dendrogram partition - (b) Graph partition

4 Community detection results

In pattern discovery step, we intend to identify community structure and detect the browsing behavior of users which can be exploited in the process of web personalization. We carried out experiments on two web systems:

- the large-scale datasets of Web pages that describe the page visits of users who visited msnbc.com on September 28, 1999. The data comes from from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September, 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail, that is, at the level of URL, but rather, they are recorded at the level of page category. The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports", such as the number of users is 989 818, the average number of visits per user is 5,7 [41]. After creating the graph of browsing behavior, we have obtained a graph with 17 nodes and edges 289 edges, for a total weight of the graph 3 708 976.

- the Ferhat Abbas Setif university website. The created web graph is a directed graph which represents the user sessions of a web site and contains 136 nodes and 2 456 edges, for a total weight of graph 25 676 (See figure 5). To obtain a first description of connection between nodes and degree distribution, a data analysis is carried out. The degree distribution is the probability distribution of degrees over the whole network. Thus the knowledge of the degree distribution reveals interesting information but does not tell us the complete structure of the network. We remark that most of nodes have low degree. The highest degree node in the network has degree 12 685. The plot cuts off at degree 260, knowing that a total of 2 456 nodes in the network. The most highly connected page is connected to 1 720 other pages. The density of this graph is 0,134.

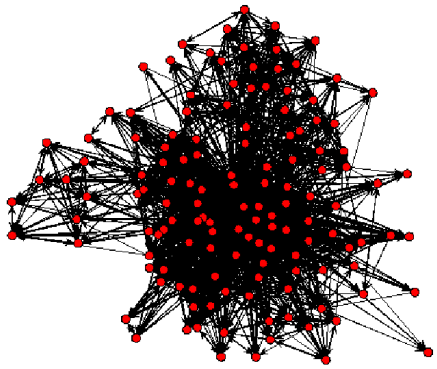

Figure 5 Graph of browsing behavior.

4.1 Modularity Evaluation

To test the performance of DCMMC algorithm, we analyze the value of modularity and compare its results with those obtained using PAM algorithm (Partitioning Around Medoids algorithm)[18], Ward's algorithm, walktrap algorithm and k-means algorithm.

PAM algorithm:

[28] compute k representative objects, called medoids. A medoid is an object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. The partitioning around mediod algorithm (PAM) chooses iteratively mediod object i and another non-medoid object j, and computes the quality of the new partition if the roles of i and j are reversed, then substitutes one for another if the switched medoids produce the best partition. The k representative objects should minimize the objective function, which is the sum of the dissimilarities of all objects to their nearest medoid.

Ward's algorithm:

Ward's method is an agglomerative hierarchical clustering method, such as the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. In [33] the author has proposed an extention of Ward's method based on Euclidean distance rather than squared Euclidean distance. An objective function and cluster distance in terms of any power α of Euclidean distance in the interval $(0, 2]$ have been proposed in order to achieve a best ability to identify clusters with nearly equal centers.

Walktrap algorithm:

The walktrap algorithm is based on the intuitive report that if a walker is in a community, it has a strong probability to remaining in the same community at the following step. [27] have defined distance metric related to the spectral approaches which are based on the fact that two nodes belonging to the same community have similar components on the principal eigenvectors. The algorithm computes the connected components, and applies then an agglomerative algorithm in order to discovered communities separately on connected sub-graphs.

k-means algorithm:

k-means clustering method is widely used and studied. It is developed by Mac Queen in 1967. K-means separates data into k mutually excessive groups. This iterative partitioning minimizes the sum of distance from each data to its clusters.

In this section, we compare the quality of optimization achieved by DCMMC method to the chosen approaches, and we evaluate the results according to the variation of modularities using the dissimilarity between nearest neighbors of a network.

Figure 6 illustrates the modularity over the course of algorithms. When applying the algorithms on the large scale data, the maximum modularity is $0,489$ using DCMMC algorithm. K-means algorithm also finds a good community partition when $Q = 0,46$.

Therefore, these results show the capabilities of each algorithm to identify the community structure. DCMMC algorithm identify three communities from the graph of browsing behavior (see figure 7).

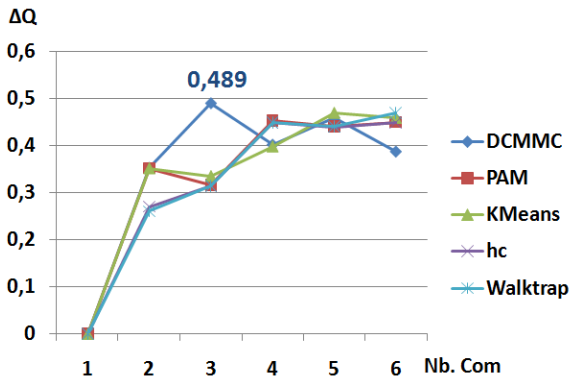

Figure 6 Comparison of the variation in modularities based on dissimilarity measure for the network partitions conducted on Graph of browsing behavior of msnbc.com data set

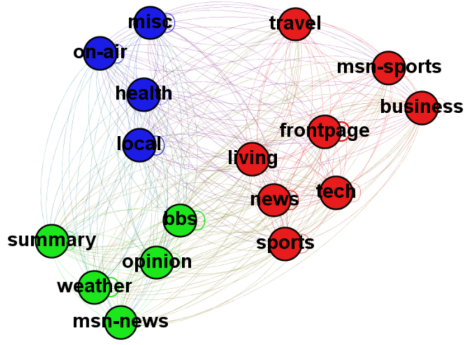

Figure 7 Extracting communities from the Graph of browsing behavior related to msnbc.com data set

The results of comparison for the modularity, when applying algorithms on Ferhat Abbas Setif university graph, are depicted in figure 8. Applying the proposed algorithm, the communities are well detected until values of $Q = 0,247$. Ward's algorithm finds a significant community structure with a larger value of $Q = 0,236$. PAM algorithm finds a good community partition when $Q = 0,213$. Using Walktrap algorithm, the maximize modularity is $Q = 0,160$. The modularity of k-means algorithm is $Q = 0,186$. We remark that Walktrap algorithm gives the lowest modularity because the quality of the results change according to the choice of length steps. The results of k-means algorithm presents a non relevant community structure du to the fact that the centroid computation influences the final result of the partition. We remark that PAM maximize the modularity and offers a reduction in search space exploration. It appears also that Ward's algorithm reveals a significant community structure. The DCMMC algorithm indicates a clear modular structure because it employ the useful potentially information contained in the sequence of sessions, therefore it ensures a better revealing of a community structure.

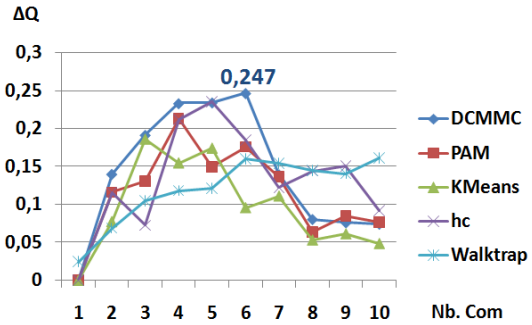

Figure 8 Comparison of the variation in modularities based on dissimilarity measure for the network partitions conducted on the Graph of browsing behavior of Ferhat Abbas Setif university website.

4.2 Community structure

Network community is typically thought of as a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network [14]. We have analyzed the identification of communities using DCMMC, Ward, PAM, k-means and walktrap to find whether the network can be divided into two or more communities. Table 2 shows results and gives the total inner community edges and the total outer community edges for each community partition. The partitions represent a good community structure when inner community edges exceed the outer community edges. For example, the partition in eight communities gives 28% total inner community edges and 72% total outer community edges using Walktrap algorithm, so it means that this partition identify a very weak community structure. We can also remark, however, the presence of a relevant community structure when comparing the different values of modularity.

Table 2 The results of detecting communities using DCMMC, Ward, PAM, k-means and walktrap. Comparaison of inner community edges and the inter community edges for different network partitions.

Nb Comm.	DCMMC		PAM		KMeans		HC(Ward)		Walktrap	
	outer	inner	outer	inner	outer	inner	outer	inner	outer	inner
2	15 101	10 575	12 991	12 685	12 991	12 685	12 991	12 685	10 545	5 305
	59%	41%	51%	49%	51%	49%	51%	49%	67%	33%
3	9 388	16 288	8 633	17 043	9 574	16 102	9 694	15 982	1 980	1 945
	37%	63%	34%	66%	37%	63%	38%	62%	50%	50%
4	6 683	18 993	6 705	18 971	6 554	19 122	7 210	18 466	1 966	2 546
	26%	74%	26%	74%	26%	74%	28%	72%	44%	56%
5	5 360	20 316	5 976	19 700	6 216	19 460	6 930	18 746	709	1 208
	21%	79%	23%	77%	24%	76%	27%	73%	37%	63%
6	4 873	20 803	5 651	20 025	5 061	20 615	5240	20 436	3 195	2 101
	19%	81%	22%	78%	20%	80%	20%	80%	60%	40%
7	4 949	20 727	4 999	20 677	5 146	20 530	5 019	20 657	2 006	2007
	19%	81%	19%	81%	20%	80%	20%	80%	50%	50%
8	4 518	21 158	4 359	21 317	4 323	21 353	4 628	21 048	15 854	6 306
	18%	82%	17%	83%	17%	83%	18%	82%	72%	28%
9	4 306	21 370	4 121	21 555	4 266	21 410	3 982	21 694	17 766	6 541
	17%	83%	16%	84%	17%	83%	16%	84%	73%	27%
10	4 215	21 461	3 118	22 558	4 183	21 493	3 420	22 256	13 559	8 741
	16%	84%	12%	88%	16%	84%	13%	87%	61%	39%

According to the statistics that qualify the quality of a division of the network given in table 3, we conclude that DCMMC algorithm maximize the modularity whith the partition in six communities. Therefore, we study the division of networks in six communities and we focus on the links that interconnect nodes inside a community in order to illustrate the presence of a detected relevant community structure.

Table 3 statistics quantifying the quality of a division of the network found by the methods: DCMMC, Ward, PAM, k-means and walktrap. Note that DCMMC algorithm maximizes the modularity in the partition of 6 communities.

Nb Comm.	DCMMC	PAM	KMeans	HC(Ward)	Walktrap
2	0,140	0,116	0,078	0,116	0,069
3	0,192	0,131	0,186	0,073	0,105
4	0,234	0,213	0,155	0,212	0,118
5	0,235	0,149	**0,174**	**0,236**	0,121
6	**0,247**	**0,176**	0,096	0,185	**0,160**
7	0,140	0,137	0,111	0,123	0,154
8	0,080	0,064	0,053	0,145	0,145
9	0,076	0,085	0,061	0,151	0,140
10	0,074	0,077	0,049	0,092	0,161

Results shown in table 4 compare the communities connections in the partition of 6 communities detected by DCMMC, Ward, PAM, k-means and Walktrap. We compare the ratio between the sum of weights of intra-links (links connecting nodes in the same community) and the sum of weights of inter-links (links connecting nodes from different communities)in order to illustrate if community exhibits Radicchi strong or weak community property.

Table 4 Analysis of the communities connections in the partition of 6 communities detected by DCMMC, Ward, PAM, k-means and walktrap.

#Comm.	DCMMC		PAM		KMeans		HC(Ward)		Walktrap	
	outer	inner	outer	inner	outer	inner	outer	inner	outer	inner
1	8 345	0	1 415	2721	731	1 798	1 616	2 369	1335	242
	100%	0%	34%	66%	29%	71%	41%	59%	85%	15%
2	1 230	2 127	2 541	3138	2 141	3 039	2 164	3 391	669	343
	37%	63%	45%	55%	41%	59%	39%	61%	66%	34%
3	218	472	8 345	0	853	2 630	128	1 899	264	64
	32%	68%	100%	0%	24%	76%	6%	94%	80%	20%
4	454	2 405	359	1 018	1 336	2 748	8 345	0	20 094	1 559
	16%	84%	26%	74%	33%	67%	100%	0%	93%	7%
5	1 359	3 302	1 336	2 748	8 345	0	1 332	2 377	781	227
	29%	71%	33%	67%	100%	0%	36%	64%	77%	23%
6	1 612	4 152	2 055	0	2 055	0	2 055	0	74	24
	28%	72%	100%	0%	100%	0%	100%	0%	76%	24%
Total	25676									

[9] have defined two quantitative community definitions; Community in strong sense and Community in weak sense. Then, a subgraph $C \subset G$ will be a community in the strong sense if each of its nodes has more links connecting it with nodes in C than those that connect it with other nodes not belonging to C. In a similar way, $C \subset G$ will be a community in the weak sense if the sum of the number of links that interconnect nodes inside C is larger than the sum of all links that connect nodes in C with nodes not belonging to C. For example, in the fourth community, the weight of intra-links is equal to 84% while the weight of inter-links equal to 16% when using DCMMC, and when applying PAM,the weight of intra-links is equal to 74% while the weight of inter-links equal to 26%, the k-means also gives a significant results. This means that these algorithms allow a cohesive community structure because of the existence of more intra links, and fewer inter link. However, for this example, Walktrap algorithm gives a non cohesive community structure.

Table 5 The communities size detected by DCMMC algorithm and the other community descovery algorithms for the studied web site .

#Comm.	DCMMC	PAM	KMeans	HC(Ward)	Walktrap
1	1	54	58	80	8
2	52	29	24	25	8
3	29	1	1	22	4
4	27	43	8	1	100
5	19	8	1	7	14
6	8	1	1	1	2
Total	136				

Table 5 focus on the size of communities when applying the discovery methods on the extracted graph. The DCMMC algorithm has detected a cluster composed of single node which represents the starter page and is classified in the fourth community. Ward algorithm have also detected this root page which corresponds to the site's home page, it is detected in third community. This result is obtained because Ward's algorithm creates a hierarchical structure that reflects the order in which groups are merged and DCMMC is based on the occurrence of resources. The first community detected by DCMMC contain the users interesting to scientific manifestation, its size is 19 nodes, this community is belonging to the sixth community identified when using PAM method, k-Means method, and Ward algorithm, but community size depends on the clustering manner of of each method, for example, PAM method regroups the visitors consulting scientific manifestation, scientific activities, and electronic resources in one community of 29 nodes. We remark that it is more significant to make comparison according to the best division of each algorithm (PAM algorithm has a best division in three communities, k-means provides an optimal modularity with two partitions, and Ward's method detects a good five communities). However, Walktrap algorithm generates misclassification in some nodes, that is way a accurate choice of the length of steps is crucial. DCMMC method extracts third community identifying nodes that access to the pages of competitive examination and Post graduation, it contains 27 nodes which corresponds to the fifth community detected by Ward's algorithm. In this case it reveals better patterns by identifying two groups (the fifth community that reflecting post graduation access with 22 node and the fourth community that contains competitive examination with 7 nodes). We recognize that Ward's method, PAM and DCMMC methods reveal some interesting patterns.

As we can see from figure (9), this network contains six communities that identify groups of users with similar behavior for which a personalized versions of the Web site of Ferhat Abbas University (Algeria) may be created. The structure identifies the users' session and all the sessions (the nodes represent resources and edges represent the browsing sequences of users during each session).

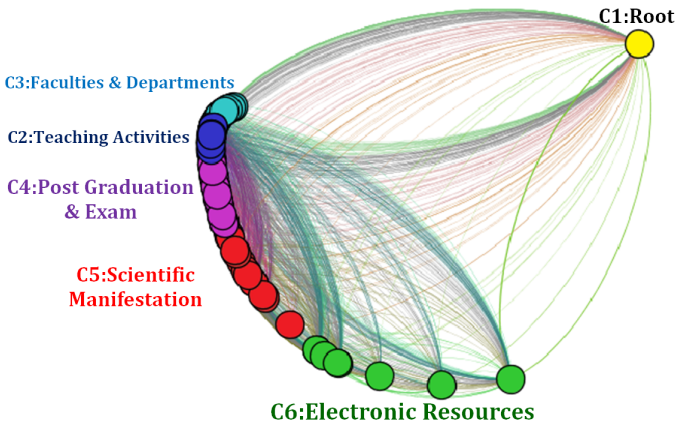

Figure 9 The community structure of the studied web site is detected by the proposed DCMMC method

- Yellow node illustrates the community of the root web page.
- The second community has produced the sixth cluster reflecting groups that access to teaching activities (blue node).
- In figure, the nodes labeled by clear blue represent the third community that regroups the accesses to the pages of various faculties and departments. (29 nodes)
- The fourth community illustrates the visits to the web pages of Post graduation and accesses to the pages of competitive examination . Mauve nodes are belonging to this community.
- The fifth community identifies the accesses to scientific manifestation. The partition graph shows this community presented by 19 red nodes.
- The last community identifies the visitors consulting the scientific activities of Ferhat Abbas University and the pages of electronic resources. Figure (9) shows that it contains 8 green nodes.

A good choice of the optimal value of quality function provided better quality of results. We have obtained the network structure that identifies the users' session and describes the pertinent communities in the pattern of the website.

5 Conclusion

In summary, our work aims to use the web user mining process to extract web access users' behavior and to improve the web site design. That for we have proposed an efficient technique in the discovery phase in which we use a dissimilarity measure based on a stochastic process that converges to a steady probability distribution. The results show an accurate identification of community structure and give a meaningful description of existing communities allowing best personalization of the information presented in the users' behavior.

References

[1] Agrawal, R., Srikant, R. (1994) 'Fast algorithms for mining association rules', Proceedings 20th Int. Conf. Very Large Data Bases, VLDB. Vol. 1215, pp. 487–499.

[2] Arasu, A., Novak, J., Tomkins, A., Tomlin, J. (2002) Pagerank computation and the structure of the web: Experiments and algorithms, Proceedings of the Eleventh International World Wide Web Conference, Poster Track. pp. 107–117 .

[3] Baraglia, R., Lucchese, C., Orlando, S., Serrano, M., Silvestri, F. (2006) 'A privacy preserving web recommender system', Proceedings of the 2006 ACM symposium on Applied computing. pp. 559–563. SAC'06, ACM, New York, NY, USA .

[4] Benedek, A., Trousse, B. (2002) 'Adaptation of self-organizing maps for CBR case indexing', Proceedings of the Forth International Workshop on Symbolic and Numeric Algorithms for Scientific Computing. pp. 31–45 .

[5] Bontognali, U. (2008) 'AWI, Applicazione Web Intelligente', SUPSI.

[6] Cooley, R.W. (2000) 'Web usage mining: discovery and application of interesting patterns from web data', Ph.D. thesis, University of Minnesota.

[7] Donetti, L., Munoz, M.A. (2004) 'Detecting network communities: a new systematic and efficient algorithm', Statistical Mechanics: Theory and Experiment.

[8] Ezeife, C.I., Lu, Y. (2005) 'Mining web log sequential patterns with position coded pre-order linked WAP-Tree', Data Min Knowl Disc, Vol. 10, No. 1, pp. 5–38.

[9] F. Radicchi, C. Castellano, F.C.V.L., Parisi, D. (2004) 'Defining and identifying communities in networks', Proceedings Nat. Acad. Sci. USA 101, pp. 2658–2663.

[10] Facca, F.M., Lanzi, P.L. (2005) 'Mining interesting knowledge from weblogs: a survey', Data and Knowledge Engineering, Vol. 53, No. 3, pp. 225–241.

[11] Feller, W. (2008) 'An Introduction to Probability : Theory and Its Application', Wiley India Pvt. Limited, 3rd edn., Vol. 1. Aug 2008.

[12] Feller, W. (2008) 'An Introduction to Probability : Theory and Its Application', Wiley India Pvt. Limited, 2nd edn., Vol. 2. Aug 2008.

[13] Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.M. (2002) 'Self-organization and identification of web communities', Computer , Vol. 35, No. 3, pp. 66–70.

[14] Girvan, M., Newman, M. (2002) 'Community structure in social and biological networks', Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 12, pp. 7821–7826.

[15] Girvan, M., Newman, M. (2006) 'Computing communities in large networks using random walks', Journal of Graph Algorithms and Applications, Vol. 10, No. 2, pp. 191–218.

[16] Jaczynski, M., Trousse, B. (1998) 'WWW assisted browsing by reusing past navigations of a group of users', Advances in Case-Based Reasoning, Springer, pp. 160–171.

[17] Kamvar, S., Haveliwala, T., Manning, C., Golub, G. (2003) 'Exploiting the block structure of the web for computing pagerank', Stanford University Technical Report.

[18] Kaufman, L., Rousseeuw, P.J. (2009) 'Finding groups in data: an introduction to cluster analysis', Vol. 344. Wiley. com.

[19] Sales-Pardo, M., Guimera, R., Moreira, A.A.M., Amaral., L.A.N. (2007) 'Extracting the hierarchical organization of complex systems', Proceedings of the National Academy of Sciences, Vol. 104, No. 39, pp.15224–15229.

[20] McSherry, F., Mironov, I. (2009) 'Differentially private recommender systems: building privacy into the net', Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 627–636.

[21] Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (2004) 'Advances in Web Mining and Web Usage Analysis' Proceedings 6th International Workshop on Knowledge Discovery on the Web, WEBKDD 2004, Seattle, WA, USA, August 2004 , Lecture Notes in Artificial Intelligence. Springer (Nov 2006).

[22] Nasraoui, O. (2005) 'World wide web personalization', Encyclopedia of Data Mining and Data Warehousing, Idea Group (2005).

[23] Newman, M.(2009) 'Networks: an introduction', Oxford University Press, Oxford (2009).

[24] Ning, L., Yang, G., Guifeng, T., Shifu, C. (2004) 'Mining web sequential patterns using reinforcement learning', Proceedings Advanced Web Technologies and Applications, No. 3007 in Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 920–923.

[25] Paliouras, G., Papatheodorou, C., Karkaletsis, V., Tzitziras, P., Spyropoulos, C.D., 'Large-scale mining of usage data on web sites', AAAI 2000 Spring Symposium on Adaptive User Interfaces.

[26] Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D. (2003) 'Web usage mining as a tool for personalization: A survey'. User Modeling and User-Adapted Interaction, Vol. 13, No. 4, pp. 311–372.

[27] Pons, P., Latapy, M. (2006) 'Computing communities in large networks using random walks', Journal of Graph Algorithms and Applications, Vol. 10, No. 2, pp. 191âŁ"-218.

[28] Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J. (2006) 'Clustering rules: a comparison of partitioning and hierarchical clustering algorithms', Journal of Mathematical Modelling and Algorithms, Vol. 5, No. 4, pp. 475–504.

[29] Sandvig, J.J., Mobasher, B., Burke, R. (2007) 'Robustness of collaborative recommendation based on association rule mining', Proceedings of the 2007 ACM conference on Recommender systems. RecSys'07, ACM, New York, NY, USA , pp. 105–112.

[30] Sandvig, J.J., Mobasher, B., Burke, R. (2008) 'A survey of collaborative recommendation and the robustness of model-based algorithms', IEEE Data Engineering Bulletin, Vol. 31, No. 2, pp. 3–13 .

[31] Slimani, Y., Moussaoui, A., Lechevalier, Y., Drif, A. (2011) 'A community detection algorithm for web usage mining systems', The Fourth IEEE International Symposium on Innovation in Information & Communication Technology (ISIICT2011) , pp. 112–117.

[32] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (Jan 2000) 'Web usage mining: discovery and applications of usage patterns from web data', SIGKDD Explor. Newsl., Vol. 1, No. 2, pp. 12–23.

[33] Szekely, G.J., Rizzo, M.L. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method, Journal of Classification, Vol. 22, No. 2, pp. 151–183,2005.

[34] Tan, P.N., Kumar, V. (Jan 2002) 'Discovery of web robot sessions based on their navigational patterns', Data Mining and Knowledge Discovery, Vol. 6, No. 1, pp. 9–35.

[35] Tanasa, D. (2005) 'Web usage mining: Contributions to intersites logs preprocessing and sequential pattern extraction with low support', Ph.D. thesis, University of Nice Sophia Antipolis.

[36] Trousse, B., Jaczynski, M., Kanawati, R. (1999) 'Using user behaviour similarity for recommendation computation: the broadway approach', Proceedings of the HCI International, Vol. 99, pp. 85–89.

[37] Zhu, T., Greiner, R., Hubl, G. (2003) 'An effective complete-web recommender system', The Twelfth International World Wide Web Conference (WWW2003).

[38] Bamshad Mobasher and Olfa Nasraoui, CHAPTER 12: Web Usage Mining, invited book chapter in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)âŁœ, Second Edition, July 2011, by Bing Liu.

[39] Jia, C., Carson, M. B., Yu, J. (2013). A fast weak motif-finding algorithm based on community detection in graphs. BMC bioinformatics, 14(1), 1.

[40] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment P1000,(10),2008.

[41] I. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White, Visualization of navigation patterns on a Web site using model-based clustering, Journal of Data Mining and Knowledge Discovery.

[42] Lu, H., Zhao, Q., Gan, Z, A Community Detection Algorithm Based on the Similarity Sequence. In International Conference on Web Information Systems Engineering (pp. 63-78). Springer International Publishing, October 2014.