International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 578 - 582

Discovery and Analysis of Usage Patterns for Web Personalization

Yacine Slimani
Laboratory of Intelligent Systems
University Ferhat Abbas Setif 1
Setif, Algeria
e-mail: slimani_y09@univ-setif.dz

Abdelouahab Moussaoui
Laboratory of Intelligent Systems
University Ferhat Abbas Setif 1
Setif, Algeria
e-mail: moussaoui.abdel@gmail.com

Ahlem Drif
Lab. of Network and Distributed System
University Ferhat Abbas Setif 1
Setif, Algeria
e-mail: adrif.univsetif@gmail.com

***Abstract*—**In this paper, we present a community discovery method based on information extraction from user sessions in order to find usage patterns. We have characterized the overall access from each page in the web site and have detected the pertinent users within communities using the potentially useful information available in different user sessions. Then we compare the proposed method with the random walk method which is one of the most popular graph clustering methods that gives a good partition of weighted and overlapped graphs. As a result, the proposed method helps us understand the behavior of the user in the web and improve access modes to information.

Keywords- *Community detection, Social networks, Quality function, Web usage mining*

*****

I. INTRODUCTION

WEB Usage Mining is the application of data mining techniques to discover usage patterns from Web data in order to better understand and serve the needs of Web applications [1]. In a general process of Web Usage Mining, there are three main phases: preprocessing, pattern discovery and model analysis [2]. In this paper, we propose a community discovery method based on the occurrence of resources during user sessions in order to extract the communities that describe user behavior. The proposed method analyzes the pretreated data of the session base and represents them via a functional graph, so that the resources will be represented by nodes, and the browsing sequences of users during each session will be represented by edges. After obtaining this graph, we proceed to the identification of the user communities based on the relevant coefficient that reflects how tightly linked a node is with a given community. Then, we redefine the modularity to assess the quality of the results. The paper is organized as follows: section 2 describes the preprocessing results, section 3 presents the proposed community discovery method and its basic idea, and then we describe the random walk method that we have used for comparison. Section 4 presents a comparative study between the two discovery methods and illustrates the identified communities. A general conclusion is presented in section 5.

II. PREPROCESSING RESULTS

Data preparation and filtering steps can take considerable amounts of processing time. Data preprocessing includes transformation, cleaning, structuring, and filtering. The data transformation module transforms the log file from its textual form to a structured form. A lexical and syntactic analysis is executed on each line of the Log file (UCLF). The data cleaning module is used to remove useless records in order to maintain only user data which can be accurately exploited to identify the browsing behavior of users. We have presented a complete methodology for preprocessing the Web logs in [3]. In this work, our preprocessing method has been tested on log files stored on the Web site Server of Ferhat Abbas University of Setif (Algeria) available at the URL http://www.univ-setif.dz.

The treated files cover the site activities during the period from 18-12-2011 at 04:04:57 to 19-01-2012 at 08:47:04. After structuration, we obtained 1,708,385 requests.

Then, we applied a data cleaning step to maintain only actions related to users' browsing behavior by eliminating the requests that a method is different from "GET" or a status is different from "200", requests to style files, multimedia files or scripts. After the cleaning phase, we obtained 314,115 (i.e. 18,38%) valid requests for the 27,520 users who accessed 23,872 pages. Table I recapitulates the different removed requests of the database.

In the second step of preprocessing, we applied a structuration algorithm [4] to determine the sessions taking account of maximum elapsed time ΔMax_t = 30min between two consecutive accesses [1,5], the result was 50,131 sessions for 314,115 navigations. After the identification of user sessions, we identified web robots' requests of search engines and web crawlers [6]. The minimum elapsed time between consecutive accesses is also fixed at ΔMin_t = 10 seconds. Thus, the result was 31,017 sessions for 175,681 navigations. Table 1 summarizes the results of identification sessions and robots.

TABLE I. SUMMARY OF IDENTIFICATION OF SESSIONS AND ROBOTS.

Category	*User*	*Page*	*Session*	*Navigation*
Before identification	27 520	23 872	50 131	314 115
Robots	1 438	20 266	19 114	138 434
After identification	26 082	7 259	31 017	175 681

After the identification of user sessions, we perform a data filtering step to remove less requested resources and retain only the most requested ones. For each resource r_i, we consider the number of sessions NS_i that required the resource r_i. Then, we remove all requests with $NS_i < \varepsilon$ such that ε is a given threshold (100). We thus obtained a significant 136 pages, to be used in the following phases of the web usage mining.

 ISSN: 2321-8169
578 - 582

III. COMMUNITY DISCOVERY METHOD

Several studies on the physical meaning and mathematical properties of complex networks have found that these networks share macroscopic properties. Among these properties, studies analyze prototype properties such as small-world effect [7], and scale-free [8], dynamic properties such as diffusion [9], [10], and structural properties such as community structure [11], [12], which appears to be common to many networks and allows us to understand the relationship between a single node, at the microscopic level, and a group at the macroscopic level. Communities (or clusters or modules) are groups of vertices that probably share common properties and/or play similar roles within the graph [13].

In this paper, the identification of communities is then used to find the center of interest of web users, and thus to restructure the site web. In this section, we describe the proposed method and clarify the concepts and definitions that we have used, then we briefly present the random walk method.

A. Relevance coefficient Algorithm

The community discovery methods based on unweighted graphs are often unable to detect a very significant aspect of community structure, since they ignore the information related to the weight of the link. To this end, we define a relevance coefficient that holds potentially useful information for different user sessions. The community discovery algorithm is defined in terms of sessions that have been identified in the preprocessing phase. A session is a contiguous set of resources that the same user has requested during his or her visit to the web site.

1) Session creation

We define:

- $R=\{r_1,r_2,\ldots,r_{nr}\}$ as a set of all distinct resources of the site web.
- $U=\{u_1,u_2,\ldots,u_{nu}\}$ as a set of all users that have accessed the website.
- $S=\{s^{(1)},s^{(2)},\ldots,s^{(ns)}\}$ as a set of session navigations, such that each user session is defined as: $s^{(i)}=(u^{(i)}, r^{(i)})$, where:
  - $u^{(i)} \in U$: is the user identification;
  - $r^{(i)}$: is the set of all resources requested during the i^{th} session (with corresponding access time) : $r^{(i)}=((t_1^{(i)},r_1^{(i)}),(t_2^{(i)},r_2^{(i)}),\ldots,(t_{ni}^{(i)},r_{ni}^{(i)}))$.

2) Graph creation

Let G=(V,E) be a graph describing pretreated data of the session base with V the set of nodes and E the set of edges.

$A_{i,j}$ is the adjacency matrix of the network. It is defined as 1 if resources r_i and r_j are connected, otherwise it is 0.

Once user sessions have been identified, we have to use them to extract the Web graph that represents the analytic network using the restructuring algorithm defined in [4]. A session base can be regarded as a sequence of pairs defining transition between pages. Each resource r_i corresponds to a node of the graph.

Each couple (r_i,r_j) of visited pages in chronological order of time during a session corresponds to an edge in the graph between these two nodes.

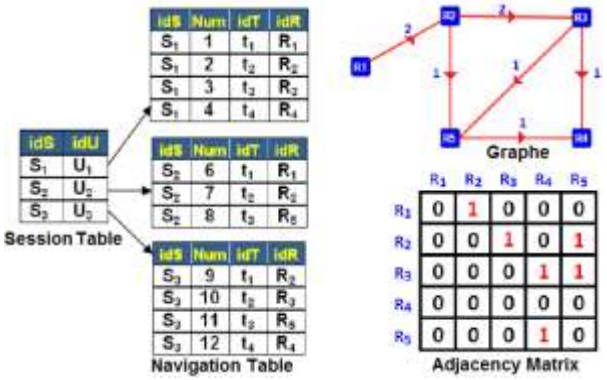

Figure 1. Creation of the graph from sessions.

3) Relevance coefficient

In order to identify the number of times the couples (r_i,r_j) appear in the sessions, we compute the user transaction.

Let $\beta=\{T^{(1)},T^{(2)},\ldots,T^{(ns)}\}$ be the set of transactions, where each transaction $T^{(i)}$ is a set of resources $r_j^{(i)}$ (item) required for a given session *i*.

k−Itemset: is an itemset containing k items. We associate a support for each item. The support gives the number of times the item appears in the database. The support of a pair *(i,j)* represents the support of k−item of second order (k = 2), it is the occurrence number of *(i,j)* in sessions divided by the total number of all sessions.

$$\text{Supp}(i, j) = \frac{\left|\{t \in \beta / (i, j) \subseteq \beta\}\right|}{|\beta|} \quad (1)$$

We take into consideration the number of times the couples *(i,j)* appear in sessions by assigning an appropriate weight to a graph as is shown in Fig. 1.

We define a matrix S that represents the probability of relationships between nodes in different communities. So the matrix S_{ij} can be described as :

$$S_{i,j} = A_{i,j} * \text{Supp}(i,j) \quad (2)$$

This matrix expresses the behavior of web users during the various transactions which allow us to represent the tendency of nodes to form communities. Then, we define a factor that refers to the occurrence frequency of the link *(i,j)* for each k-item sequence. Indeed, we can define:

$$f_{i,j} = \frac{S_{i,j}}{d_i} \quad (3)$$

Where d_i is the degree of node *i*.

Our idea revolves around the basic principle of community structure that a node has a high tendency to be a member of a community if it represents a meaningful or pertinent connection with a community. Thus we define a relevance coefficient, described as :

$$\lambda(i, C) = \frac{\sum_{j \in C} S_{i,j}}{\sum_{j \in V} S_{i,j}} = \sum_{j \in C} f_{i,j} \quad (4)$$

The detection of inter-community links is here based on the fact that such links are in less clustered areas. The relevance coefficient is defined as the pertinence value of node *i* in community *C*, divided by the total number of such possible relevance values for all communities.

4) Community detection Algorithm

Modularity measures when the division is a good one, in the sense that there are many edges within communities and only a few between them [14]. Modularity is used as a quality function and criterion to specify the best partition that can be found. The modularity is defined as :

$$Q = \sum_i e_{i,j} - a_i^2 = Tr_e - \left\| e^2 \right\| \quad (5)$$

where e_{ii} is the fraction of weights of edges belonging to community *i* in the total weights of all edges and a_i is the fraction of weights of edges connecting community *i* with other communities in the total weights of all edges. If the network is unweighted, the weight of each edge equals 1. Several studies [15] proposed an extended measure of modularity to account for other types of graphs. In this work, we have proposed to evaluate quality communities in weighted, and overlapped, networks. Thus the relevance coefficient reflects how much node i belongs to community c. We rewrite the modularity as :

$$Q = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in V} \left[\frac{s_{i,j} - (d_i d_j)}{2m} \right] \lambda(i,c)\, \lambda(i,c) \quad (6)$$

Where $0 \leqslant \lambda(i,c) \leqslant 1$.

The nodes are divided into communities such that if node i belongs to only one community c, λ(i,c) is equal to 1, if node i does not belong to community c, λ(i,c) = 0. We exploit the relevance coefficient to reveal the community structure of a network using an agglomerative algorithm that we have defined in Algorithm. 1.

Data: V
initially $P^{(0)} = (C_1^{(0)}, C_2^{(0)}, \ldots, C_n^{(0)})$
Repeat
- Compute relevance coefficient between each pair of adjacent communities
- Select the two communities $C_i^{(k)}$ and $C_j^{(k)}$ of $P^{(k)}$ maximize modularity
- Create the new partition $P^{(k+1)}$
- Update $S_{i,j}$ for all adjacent communities to the new community.

Until we have only one community;
Select the best partition in communities of the resulting dendrogram which maximizes the quality function.

Algorithm 1: Communities discovery.

B. Random walk Algorithm

Pons et al [16] have proposed a hierarchical agglomerative clustering method based on random walks. The walk-trap algorithm uses 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 stage. A random walk in graph *G* is a process in discrete time on the unit of the nodes *V*. Its transition probability *P* is defined:

$$P_{i,j} = \frac{A_{i,j}}{w_i} \quad (7)$$

Where $P_{i,j}^t$ is the probability of going from node *i* to node *j* by a random walk of a length *t*. Time is discretized (t=0,1,2…) and a walker is localized at every moment *t* on a node of the graph *G*.

The walker moves at every moment randomly and uniformly towards one of its neighbor nodes. The transition matrix can be defined from the adjacency matrix A of graph and diagonal matrix D of the weight of the nodes according to the following formula:

$$P = D^{-1} A \quad (8)$$

Let ρ(t) be the probability distribution vector of position of the walker after t step where :

- $\forall i \in V, \rho_i^{(t)} \geqslant 0$,
- $\sum_{i \in V} \rho\, i(t) = 1$.

Thus a walker has great chances to remain for a short walk in it origin community. The distribution $\rho^{(0)}$ gives the initial position of the walker, for example if the walker start from a single node *i*, $\rho_i^{(0)} = 1$, $\forall j \neq i$, $\rho_j^{(0)} = 0$. The walker may also start from several nodes with both different probabilities for each node. The probability to achieve a node *j* at time t+1 is directly related to the probabilities of position at time *t* and the transition probabilities $P_{i,j}$:

$$\forall j \in V, \rho_j^{(t+1)} = \sum_{i \in V} \rho_i^{(t)} P_{i,j} \quad (9)$$

This equation is written as :

$$\rho^{(t+1)} = P^T \rho^{(t)} \quad (10)$$

Note that P^T is the transpose matrix of P. Thus, $\rho^{(t)}$ can be computed as :

$$\forall j \in V, \rho_j^{(t)} = \sum_{i \in V} (P^t)_{i,j}\, \rho^{(0)} \quad (11)$$

Which allows interpreting the $P^t_{i,j}$ value as the probability of going from node *i* to node *j* in *t* steps. The reversibility property of the walk steps indicates that the probabilities $P^t_{i,j}$ and $P^t_{j,i}$ directly dependent. All the information of the random steps concerning a given node $i \in V$ is contained in the probabilities $P^t_{k,j}$, $k \in V$. These probabilities correspond to the ith line of the matrix P^t, and are noted by a vector column $P^t_{.,i}$. To compare two nodes i and j and to define a distance between them, Pons et al [16] have defined the following remarks:

- Two nodes *i* and *j* of the same community tend to see the other nodes in the same way $\forall k, P^t_{k,i} \approx P^t_{k,j}$
- If two nodes i and j are in the same community, the probability $P^t_{i,j}$ will be surely high, however an important probability $P^t_{i,j}$ does not mean inevitably that i and j are in the same community.
- The probability $P^t_{i,j}$ is influenced by the degree of the arrival node d_j since it is easier to reach the nodes of strong degree by a random walk.

The distance $r_{i,j}$ between the nodes of the graph is defined as follow:

$$r_{i,j} = \sqrt{\sum_{k=1}^{n} \frac{\left(P^t_{i,k} - P^t_{j,k}\right)^2}{w_k}} \quad (12)$$

Where D is the diagonal degrees matrix, *r* is a Euclidean distance on R^n.

The generalization of this distance is a distance between communities defines as follows:

$$P^t_{C_{i,j}} = \frac{1}{|C|} \sqrt{\sum_{i \in C} P^t_{i.}} \quad (13)$$

Where C ⊂ V is a community $P^t_{C.}$ is the probability vector which corresponds to a walk of length t starting uniformly from the node of the community C. This allows defining the particular case of a distance between a node *i* and a community C:

$$r_{iC} = \left\| D^{1/2} P^t_{C.} - D^{1/2} P^t_{i.} \right\| \qquad (14)$$

Pons et al [16] have defined distance metric which is 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 which discovered communities separately on connected sub-graphs.

IV. COMMUNITY DETECTION RESULTS

In pattern discovery step, we intend to identify community structure and detect the browsing behavior of users in order to exploit it in the process of Web personalization. The community discovery methods which we have used to detect clusters are agglomerative algorithms that start with each node in its own singleton community or another set of small initial communities, iteratively merging these communities into larger ones.

A. Graph description

We can obtain some insight into the network structure by throwing out information about the network degree distribution. The degree distribution gives important clues into structure of a network as show in Fig. 2. The degree of a node is its most basic structural property; most nodes have a medium node degree.

We remark that the degrees of all nodes are distributed around the average equal to 2.5 and most pages tend to connect to each other if they share similar topic.

Table 2 presents the results of the analysis of the studied network and synthesizes the degree information.

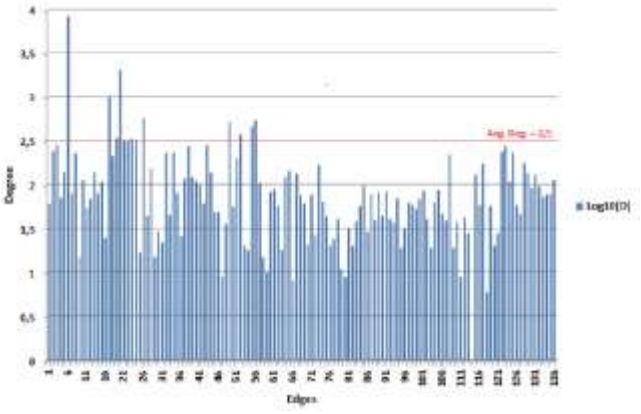

Figure 2. Degree distribution

TABLE II. SUMMARY OF GRAPH DESCRIPTION.

Category	*Value*
Number of Edges (n)	136
Number of Vertex (m)	2 456
Average Degree	377
Density	0,134

B) Quality function

The quality of partition in two discovery algorithms communities is illustrated by the graph modularity (Fig. 3). The relevance coefficient algorithm finds 12 communities, with a maximum value of modularity of 0.347. The random walk algorithm finds 13 communities structure with a larger value of Qmax = 0.247.

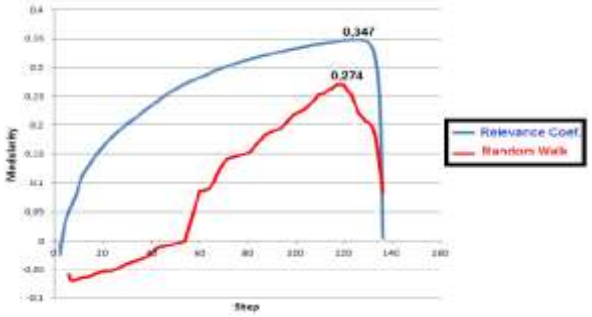

Figure 3. Comparison of the variation in modularity between relevance algorithm and random walk algorithm.

So, both methods show a significant community structure. The proposed method shows a significantly stronger increase in modularity than expected if we use the random walk algorithm. In the case of random walk algorithm, there is a large distance gap between an intercommunity edge and an intracommunity edge; on the other hand the relevance coefficient algorithm allows the starting node to detect its community locally. The random walk method that has defined in [16] takes advantage of spectral properties without resorting to an explicit computing of eigenvalues and eigenvectors. It takes advantage of the spectral properties while maintaining a reasonable complexity O(nmlog(n)) through calculations of random walks. However, when the steps increase, the quality of the results decreases. The proposed method is based on the concept of relevance coefficient which allows it to be much more accurate in the presence of nodes with high occurrences in a given community. The information contained in the edge weights can be efficient to detecting communities and preserving the data. So, it well describes the behavior of users. We note that if the node degree is low, the classification ability of the algorithm decreases. Both methods are able to cover the hierarchy organization of the hierarchical structure of the studied network and have the most similar size in some communities.

We have use weighted networks (Figure 4)which presents the degree of resource as strictly related to the frequency of accesses to that resource and we have determined the whole analytic network.

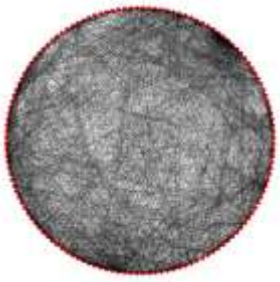

Figure 4. Graph of browsing behavior.

The proposed detection community algorithm based on relevance coefficient discriminate communities in O(mn). By applying the pattern discovery method, we have obtained 12 classes. Communities are identified as subsets of nodes within which connections are denser while between which are much sparser as in shown in figure 5. For example, we have identified the visitor interest to the web pages of Post graduation and accesses to the pages of competitive examination. The partition graph show this community presented by 12 intense mauve nodes. Another community regroups the accesses to the pages of the Faculty of Social sciences, it contains 09 red nodes. Clear blue nodes community illustrates the visits to the web pages of cooperation and those of the vice-chancellorship of the university. In the figure the 48 nodes labeled by brown represent the accesses to the pages of various teaching activities for faculties and department. The yellow community detects the accesses to the reviews and electronic resources in particular the reviews of Direct Science. The found structure well identifies the users' session and all the sessions according to the users' behavior.

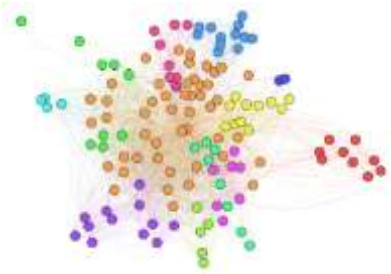

Figure 5. The identified community structure.

V. CONCLUSION

Web Usage Mining consists of three phases; preprocessing, pattern discovery and pattern analysis. In this work, the used extraction process creates more efficient session reconstructions and reduced the number of relevant requested resources which gives more accurate usage patterns. The contribution of the paper is to define the occurrence access of resources during different user sessions in order to extract the pertinent communities. The obtained results meet the needs of analysts to extract useful knowledge in order to restructure and customize the design of the study web site.

REFERENCES

[1] R. W. Cooley, "Web usage mining: discovery and application of interesting patterns from web data," Ph.D. dissertation, University of Minnesota, 2000.
[2] D. Tanasa, "Web usage mining: Contributions to intersites logs preprocessing and sequential pattern extraction with low support," Ph.D. dissertation, University of Nice Sophia Antipolis, 2005.
[3] Y. Slimani, A. Moussaoui, Y. Lechevallier, and A. Drif, "A community detection algorithm for web usage mining systems," in Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on, 2011, pp. 112-117.
[4] O. Nasraoui, "World wide web personalization," Encyclopedia of Data Mining and DataWarehousing, Idea Group, 2005.
[5] G. Paliouras, C. Papatheodorou, V. Karkaletsis, P. Tzitziras, and C. D. Spyropoulos, "Large scale mining of usage data on web sites," in AAAI 2000 Spring Symposium on Adaptive User Interfaces, 2000.
[6] P.-N. Tan and V. Kumar, "Discovery of web robot sessions based on their navigational patterns," Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 9-35, Jan. 2002.
[7] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," nature, vol. 393, no. 6684, pp. 440-442, 1998.
[8] A.-L. Barabsi and R. Albert, "Emergence of scaling in random networks," science, vol. 286, no. 5439, pp. 509-512, 1999.
[9] S. Bilke and C. Peterson, "Topological properties of citation and metabolic networks," Physical Review E, vol. 64, no. 3, p. 036106, 2001.
[10] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, "Modularity and extreme edges of the internet," Physical review letters, vol. 90, no. 14, p. 148701, 2003.
[11] J. Moody, "Race, school integration, and friendship segregation in america1," American Journal of Sociology, vol. 107, no. 3, pp. 679-716, 2001.
[12] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identication of web communities," Computer, vol. 35, no. 3, pp. 66-70, 2002.
[13] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Physics reports, vol. 424, no. 4, pp. 175-308, 2006.
[14] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, no. 2, p. 026113, 2004.
[15] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the de_nition of modularity to directed graphs with overlapping communities," Journal of Statistical Mechanics: Theory and Experiment, vol. 2009, no. 03, p. P03-024, 2009.
[16] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in Computer and Information Sciences-ISCIS 2005. Springer, 2005, p. 284-293.

Authors' Profiles

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 & 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 the UFAS1. He has plus then 10 papers in international conferences. His areas of interests include data mining, web mining and Artificial Intelligence.

Moussaoui Abdelouahab 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 & IJSC Referee. His researches are in the areas pattern recognition's algorithm, complex data mining and medical image analysis.

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 Ph.D. student and 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, dynamic topology, ad hoc networks.