International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:7, 2013

A Comparison of Fuzzy Clustering Algorithms to Cluster Web Messages

Sara El Manar El Bouanani and Ismail Kassou

***Abstract*—**Our objective in this paper is to propose an approach capable of clustering web messages. The clustering is carried out by assigning, with a certain probability, texts written by the same web user to the same cluster based on Stylometric features and using fuzzy clustering algorithms. Focus in the present work is on comparing the most popular algorithms in fuzzy clustering theory namely, Fuzzy C-means, Possibilistic C-means and Fuzzy Possibilistic C-Means.

***Keywords*—**Authorship detection, fuzzy clustering, profiling, stylometric features.

I. Introduction

THE last two decades have witnessed an exponential growth of the Internet. This is mainly due to the great number of web applications available online and to the increasing number of their users. This has generated huge quantities of data related to the users' interactions with websites. This valuable information is stored by servers in user access log files or in web sites pages. In this respect, a number of research studies using web mining techniques have been carried out to analyze the interests and the profiles of web users so as to identify and recommend appropriate services [7], [17], [18]. This identification, also termed "profiling" is applied in several areas such as criminology, e-commerce, education, etc. In criminology, for example, detecting terrorists and racist groups is of utmost importance. In [4] and [5], two approaches based on social networks and on the exploration of the web in general were proposed to identify terrorists and racist groups and to analyze their behavior and profiles.

Another computer profiling application concerns the identification of the personal interests of website users. A tool set that exploits neural networks and self-organizing maps (SOM) to identify customers' Internet browsing patterns is described in [19]. For their parts, [18] identified the potential customers of an online bookstore through web content mining whereas [2] provided a methodology combining a fuzzy K-means algorithm and neural networks to study Chilean bank clients' behavior.

As far as is education is concerned, the possibility of tracking users' behavior in e-learning environments creates new possibilities for system architects, pedagogical and instructional designers to create and deliver learning contents [3]. Examples of such studies are described in [1], [3] and [16].

Taking into account previous studies on profiling web users, it can be noticed that most of these works are based on analyzing access log files stored by servers and user's transaction records. However, access to these files is not always possible in all sites and as such cannot be easily extracted. To make up for this problem, the identification of profiles is based on texts available in web forums, blogs or social networks. To identify a user profile, we have to extract web messages, analyze them and detect texts written by our specific profile.

In the present paper, focus is on presenting part of our work, which is to cluster web messages written by the same author into the same cluster.

The rest of the paper is structured as follows. Section II defines the problem. Section III presents the concept of authorship detection and details Stylometric features. Section IV introduces the concept of fuzzy logic and fuzzy clustering algorithms. In Section V, we present our method and experimental results. Finally, we conclude in Section VI and point out some future research lines.

II. Research Questions

Today, the internet has become the most useful platform for communicating and sharing ideas. Anyone can easily access the web and make comments, publish thoughts, ideas or express opinions. Generally, most Internet users are interested in a specific topic and usually voice the same opinion when moving from one page to another. The main objective of our research study is to propose an approach that can identify a specific profile based on its writings on the web. To detect a profile, we have to extract texts from the web, analyze the writing style and the vocabulary used. Based on studies conducted on authorship detection, we developed a tool capable of clustering web messages with a certain probability using fuzzy clustering algorithms. An attempt will be made to answer the following questions: Can we detect messages written by the same person? What kind of features that would allow us to distinguish a profile from another? Given the specific characteristics of web messages, can authorship identification techniques be applied to these messages? What types of features are effective for identifying the authors of online messages? What classification techniques are efficient for clustering web messages? Can specific vocabulary be used to identify a profile's text? Finally, how could we specify a unique write print for each web user?

S. El Manar El Bouanani is with ENSIAS, Mohammed V Souissi University, Rabat, Morocco (corresponding author e-mail: sara.bouanani@um5s.net.ma).

I. Kassou is with ENSIAS, Mohammed V Souissi University, Rabat, Morocco (e-mail: i.kassou@um5s.net.ma).

III. Stylometric Features

Stylometry studies have shown that every individual can have a print related to his writing style. This is called "writeprint" [11]. The writing style of an individual is defined in terms of word usage, selection of special characters, composition of sentences and paragraphs, organizing sentences into paragraphs and paragraphs in documents [11]. Studies carried out on stylometry ([8], [11] and [20]) have defined types of features:

- ***Lexical features*** are used to learn about the special characters and words that an individual prefers to use. This is especially related to the frequency of different alphabets, the total number of uppercase letters, average number of characters per word, average number of characters per sentence.
- ***Syntactic features***, also called style markers, are function words such as "good," "where," "as," "your" and punctuation marks. '
- ***Structural features*** used to determine how an individual organizes the presentation and structure of texts (subsections, paragraphs, sentences ...)
- ***Content-specific features*** are used to characterize certain discussion forums by a few keywords or terms.

Some examples of these features are summarized in the following table (Table I).

TABLE I
Stylometric Features

Lexical Features
Character Count
Digits Count
Uppercase Letters Count
Spaces Count
Tabs Count
Occurrence Of Alphabets
Occurrence Of Special Characters
(<,>,%, |, {,}, [,], \, @, #, +, -, *, $, ^, &, ~, ÷, /)
Tokens Count
Average Sentence Length In Terms Of Characters
Average Token Length
Short Words Count (1-3 Characters)
Yule's Measurement
Hapax Legomena (Words Repeated Once)
Dislegomena Hapax (Words Repeated Twice)
Syntactic Features
Occurrence Of Punctuations (,.?!:; "')
Occurrence Of Functions Word Such As (Well, Where, As, Your, Our)
Structural Features
Rows Count
Sentences Count
Paragraphs Count
Average Paragraphs Length In Terms Of Sentences
Average Paragraphs Length In Terms Of Words
Max Length Of Sentences In Terms Of Words
Min Length Of Sentences In Terms Of Words
Presence / Absence Of A Salutation
Content-Specific Features
All the Words in the lexical field of thematic within a given text

Stylometric features are used to find the potential author of a text. To identify the authors of posts in the web, [20] presented a framework by using the four abovementioned features. Experience has shown that this framework detects anonymous authors of Web messages with a probability of 70-90% and the SVM classifier gives better results than neural networks or decision trees. Reference [10] describes another approach to analyze emails, extract the writing style, and track text messages written by criminals. Similarly, [8] presented an approach for detecting the authors of messages using the SVM algorithm. Reference [14] proposed a methodology to detect blogs' authors based on the features of texts and the LIWC software that introduces the concept of feelings from the authors' vocabulary. Another proposal was made in [15] to identify potential authors of instant web messages using three classifiers implemented in WEKA (J48, IBk, and Naive Bayes).

In our previous work [9], we demonstrated that Stylometric features are able to detect web messages written by the same person. Due to the specificity of web messages, Stylometric features tested in [9] are specific to web messages. The clustering algorithms used are C-means and Expectation Maximization (EM). Experience shows that both approaches can classify texts written by the same person in the same cluster with a high accuracy. However, we noted that this kind of classification named "hard clustering" is not really the best solution to classify authors' messages. Each text actually belongs to one and only cluster and this methodology does not allow us to explain why some texts are not correctly classified. In this context, we thought to use a fuzzy clustering approach that will assign texts to clusters with a certain probability.

Unfortunately, no studies about authorship detection using the fuzzy clustering theory are available in the literature.

IV. Fuzzy Clustering Algorithm

Cluster analysis is the formal study of algorithms and methods for grouping data. It is also a tool for exploring data structure. Therefore, it may reveal relations and structure in data. Cluster analysis has been used in a variety of disciplines such as pattern recognition, image processing, information retrieval, marketing and many more [13].

Most traditional cluster analysis algorithm is crisp partitioning which means each pattern belongs to one and only cluster. However, most objects have ambiguous attributes and thus method for soft partitioning is required [13].

Fuzzy Set theory was originally proposed by Lotfi A Zadeh [20]. It differs from classical notion of set in that it provides the gradual assessment of the membership function, which is ranged within the interval [0,1]. This function represents the degree of the statement in a fuzzy way [19]. Fuzzy Logic is used in problems where the results can be approximate rather than exact.

A. Fuzzy C-Means (FCM)

The FCM algorithm assigns membership values which are inversely related to the relative distance of a point to the prototypes (cluster centers in the FCM model) [6]. In FCM, the closeness of each data x_k, to the center a cluster v_i, is defined as the membership (u_{ki}) of x_k to the i cluster of X

minimizing the following objective function:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}^{m}||x_k - v_i||^2 \quad (1)$$

where X_k a given set of unlabeled N data; V_i are the cluster centers and m = [1, ∞] is the weighting exponent which determines the fuzziness of the resulting clusters, U = [μ_{ik}] matrix $c \times n$, where μ_{ik} is membership of x_k to the i cluster $\sum \mu_{ik}$ =1, $k = 1, 2, \ldots n$.

The cluster centers and the memberships are computed as follow:

$$Vi = \frac{\sum_{k=1}^{n}\mu_{ik}^{m}x_k}{\sum_{k=1}^{n}\mu_{ik}^{m}} \quad (2)$$

$$u_k = 1/\sum_{j=1}^{c}\left(\frac{||x_k - v_i||}{||x_k - v_j||}\right)^{2/(m-1)} \quad (3)$$

B. Possibilistic C-Means (PCM)

The PCM algorithm considers the clustering problem from the viewpoint of possibility theory [6]. The approach adopted in PCM differs from the FCM algorithm because the resulting membership values can be interpreted as degrees of possibility (or compatibility) of the points belonging to the classes. The PCM algorithm simultaneously produces both membership and typicality values. Outliers have low typicality values and automatically eliminated by the algorithm. The objective function for PCM is:

$$P_m(T,V;X,\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{c}t_{ik}^{m}d_{ki}^{2} + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-t_{ki})^{m} \quad (4)$$

where t_{ki} is the typicality of x_k to the cluster i, vi, T is the typicality matrix, defined as T = [t_{ki}]$_{NC}$, d_{ki} is a distance measure between x_k and c_i, and y_i denotes a user-defined constant, $y_i > 0$ and $1 < i < c$. By using an approximate optimization (AO) of P_m, PCM-AO algorithm, additional conditions are necessary for (1), $1 \le i \le c$, $1 \le k \le n$, as:

$$t_{ki} = 1/\left(1 + \frac{d_{ik}}{\gamma_i}\right)^{1/m-1}, \forall i,k \quad (5)$$

$$v_i = \frac{\sum_{k=1}^{n}t_{ki}^{m}x_k}{\sum_{k=1}^{n}t_{ki}^{m}}, \ \forall i \quad (6)$$

PCM algorithm solves (4) with (6) and adds the next condition on $\{y_i\}$;

$$\gamma_i = K\frac{\sum_{k=1}^{n}u_{ki}^{m}d_{ki}^{2}}{\sum_{k=1}^{n}u_{ki}^{m}}, K > 0 \quad (7)$$

where u_{ki} are membership values obtained in FCM and K=1 is mostly used.

C. Fuzzy Possibilistic C-Means (FPCM)

Using the same notation as in FCM, μ_{ik} is the membership values of the data point x_k in cluster i, while t_{ik} is the typicality value of x_k in cluster i. The objective of FPCM model is to find the partition of X into c fuzzy subset by minimizing the equation as follow [14]:

$$J_{m,n}(U,T,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik}^{m} + t_{ik}^{n})||x_k - v_i||^2 \quad (8)$$

where subject to the constraints $m > 1$, $\eta > 1$, $0 < u_{ik}, t_{ik} < 1$, $\sum\mu_{ik} = 1$ and $\sum t_{ik} = 1$ where m and η are both weighting exponents.

Under the constraints above and conditions established on c-means optimization problems, we will have the first order necessary conditions for extreme of $J_{m,\eta}(U,T,V)$ in terms of Lagrange multiplier theorem as follows.

$$t_{ik} = 1/\sum_{j=1}^{n}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/m-1}, \forall i,k \quad (9)$$

$$v_i = \frac{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{n})x_k}{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{n})x_k}, \ \forall i,k \quad (10)$$

where d_{ik} is the distance of the data point x_k to the prototype v_i, computed as:

$$d_{ik} = ||x_k - v_i|| = (x_k - v_i)^T A(x_k - v_i) \quad (11)$$

where A is symmetric positive definite matrix. When A is identity matrix, d_{ik} represents Euclidean distance which represents the similarity between data points and cluster center.

V. Our Method using Fuzzy Clustering Algorithm and Experimental Results

As a part of our work, we have to bring together texts written by the same author in the same cluster. We have N messages written by M authors extracted from the Web. What is then needed is to develop a tool that identifies the M clusters grouping with a certain probability texts written by the same person.

As Fig. 1 clearly shows, the proposed method goes through four phases. In the first phase, web messages are extracted from the Web. In the second stage, Stylometric features are computed. In the third phase, all web messages are converted into vectors containing Stylometric features. Finally, in the fourth stage, the fuzzy clustering algorithm is made use of to detect the clusters and calculate the membership function.

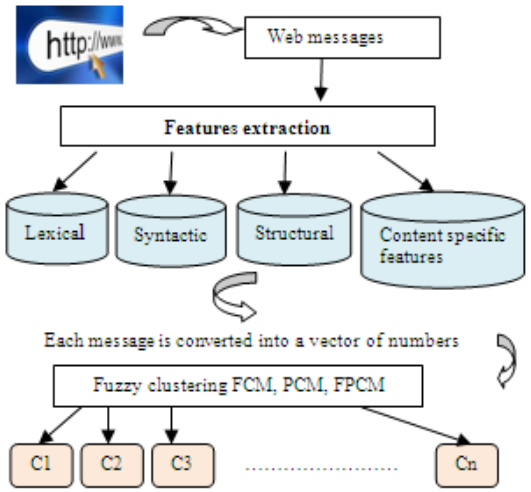

Fig. 1 Clustering web messages using fuzzy logic

Based on the studies effected on the stylometry domain and authorship detection ([8], [10]-[12], [20]), we developed an experimental prototype using JAVA language by taking as input different web messages and by calculating different Stylometric features. In other words, every text is converted into a vector whose components are the calculated Stylometric features. As an output we have a file .txt which contains N vectors (N is the number of messages). Note that the features used are specific to web messages and are those tested in [9].

The approaches used to cluster web messages are the fuzzy clustering algorithms listed in Section IV (*FCM*, *PCM* and *FPCM*). Every algorithm has as input the file containing the Stylometric features generated by our tool. These algorithms were implemented using MATLAB software program.

To validate our experimental results, we used both *Recall* and *Precision* parameters as in [9]. Both parameters calculate the *harmonic average* (denoted *F-measure*), which demonstrate our system's performance.

TABLE II
PARAMETER SYSTEM PERFORMANCE

P (i, j) = number of author's i texts in cluster j / number of author's i texts

R (i, j) = number of author's i texts in cluster j / number of texts in cluster j

F (i, j) = (2 * P (i, j) * R (i, j)) / (P (i, j) + R (i, j))

As an experiment, we considered a set of 30 web messages extracted from some websites (3 authors, 10 messages for each author). We generated the vectors (Stylometric features) using the proposed tool and then applied the fuzzy C-means (FCM) to these vectors. As a result, the three clusters generated by the algorithm can be clearly displayed. In Fig. 2 below, we can obviously see the texts assigned to each cluster together with their membership function (between 0 and 1).

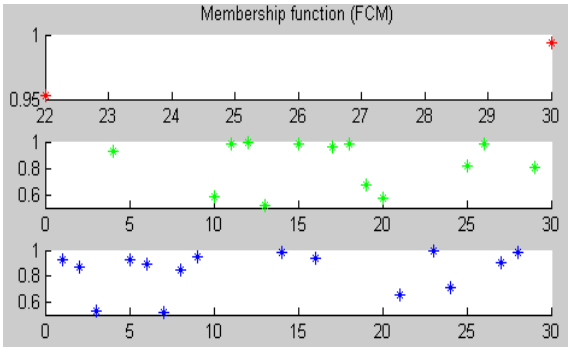

Fig. 2 FCM algorithm applied to a set of 30 messages (3 authors)

The different membership function is calculated as follows in Fig. 3.

	E1	E2	E3	K1	K2	K3	T1	T2	T3
C1	0.0676	0.1083	0.444	0.9885	0.9897	0.0124	0.3332	0.0314	0.0022
C2	0.0061	0.0213	0.0237	0.0029	0.002	0.0013	0.018	0.9531	0.0003
C3	0.9263	0.8705	0.5323	0.0086	0.0083	0.9863	0.6488	0.0155	0.9975

Fig. 3 Membership function

For example, message E1 (message 1written by author E) is assigned to cluster C1 with a probability of 0.0676, to cluster C2 with a probability of 0.0061 and to cluster C3 with a probability of 0.9263. It seems quite logical that message E1 will be assigned to cluster C3 and so will be messages E2 and E3 that are written by the same person. Note that in this case, the F-measure is equal to 0, 87.

The Possibilistic C-means (PCM) applied to the same set of texts yields the following results in Figs. 4 and 5. In this case, the F-measure is equal to 0, 8.

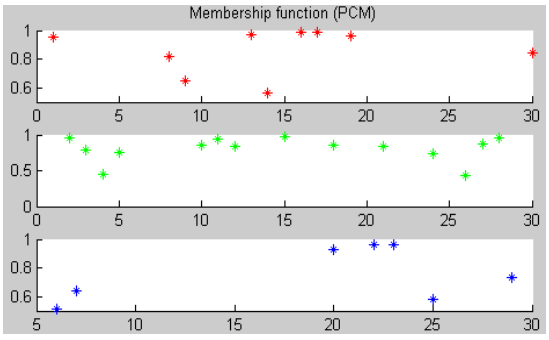

Fig. 4 PCM algorithm applied to a set of 30 messages (3 authors)

	E1	E2	E3	K1	K2	K3	T1	T2	T3
C1	0.9502	0.0276	0.0913	0.0374	0.0701	0.9665	0.0972	0.0084	0.0084
C2	0.0383	0.9562	0.7831	0.9335	0.8336	0.0265	0.8402	0.0336	0.0336
C3	0.0116	0.0162	0.1256	0.0291	0.0963	0.007	0.0626	0.958	0.958

Fig. 5 Membership function

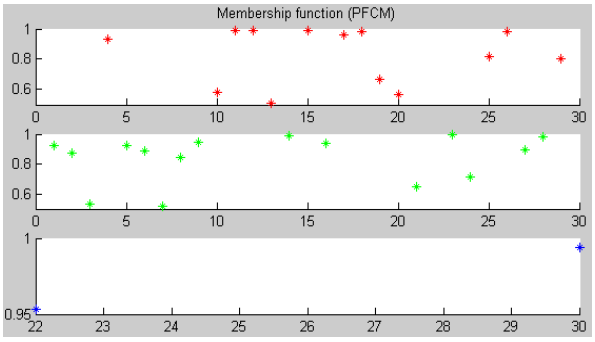

Fig. 6 FPCM algorithm applied to a set of 30 messages (3 authors)

	E1	E2	E3	K1	K2	K3	T1	T2	T3
C1	0.0709	0.0887	0.4234	0.9793	0.9587	0.6554	0.3126	0.0087	0.0004
C2	0.0059	0.0156	0.0212	0.004	0.0062	0.2236	0.0163	0.9873	0
C3	0.9232	0.8956	0.5554	0.0167	0.0351	0.121	0.6711	0.004	0.9996

Fig. 7 Membership function

The Fuzzy Possibilistic C-Means (FPCM) applied to our set gives the following results in Figs. 6 and 7. F-measure calculated for this third experience in equal to 0, 85.

The obtained results for a set of 30 messages (3 authors) are as follows in Fig. 8.

	FCM	PCM	FPCM
F-measure	0,87	0,8	0,85

Fig. 8 F-measure calculated for each algorithm

As Fig. 8 clearly shows the value of F-measure spans from 0.8 to 0.87. These values prove that that the results are good for the first experiment.

To evaluate the performance of this system, we generated other sets of web messages and carried out some other experiments by varying the number of authors or the number of texts per author.

	FCM	PCM	FPCM
10 messages	0,75	0,7	0,7
20 messages	0,8	0,75	0,78
30 messages	0,8	0,8	0,85

Fig. 9 F-measure - 3 authors

	FCM	PCM	FPCM
10 messages	0,8	0,75	0,75
20 messages	0,85	0,8	0,78
30 messages	0,87	0,8	0,8

Fig. 10 F-measure - 4 authors

	FCM	PCM	FPCM
10 messages	0,8	0,85	0,85
20 messages	0,85	0,87	0,88
30 messages	0,9	0,88	0,9

Fig. 11 F-measure - 5 authors

The results of these experiments demonstrate that the three algorithms give a high value for F-measure. As is clearly seen, the highest value is obtained by FCM and FPCM in cases in which the number of messages is 30 per author. We can also conclude that F-measure rises when the number of texts per author increases.

VI. Conclusion and Future Work

In this paper, we have presented an approach to cluster web messages based on fuzzy clustering algorithms. Results have proven that these algorithms seem to be a good way to collect messages written by the same person and group them into clusters with a certain probability; in our future work, an attempt will be made to find other tools and use them in combination with this first classification so as to define the unique write print for each user, and, hence, identify profiles.

References

[1] J. Ai, J. Laffey « Web Mining as a Tool for Understanding Online Learning », MERLOT Journal of Online Learning and Teaching, Vol. 3, No. 2, June 2007.
[2] S. Arayaa, M. Silvab, R. Weberc « A methodology for web usage mining and its application to target group identification » Fuzzy Sets and Systems 148 (2004) 139–152.
[3] J. M. Carbo, J. Minguillon , E. Mort , "User navigational behavior in e-learning virtual environments". IEEE/WIC/ACM International Conference on Web Intelligence, 2005
[4] M. Chau, J. Wu , "Mining communities and their relationships in blogs: a study of online hate group". Int. J. Human-Computer Studies, pp.57-70, 2007
[5] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, G. Weimann, Uncovering the Dark Web: A Case Study of Jihad on the Web. Journal of the American Society for Information Science and Technology, Vol.(59), Issue 8, pp: 1347–1359, 2008
[6] C. Correa, P. Barreiro, M. P. Diago, J. Tard_aguila C. Valero "A Comparison of Fuzzy Clustering Algorithms Applied to Feature Extraction on Vineyard"
[7] K. K. Chen , P. H. Chou, P. H. Li, M. J. Wu, "Integrating web mining and neural network for personalized e-commerce automatic service", Expert System with applications, Vol.(37): 2898-2910, 2010
[8] O. De Vel, "Mining e-mail authorship". In: Proc. of the Workshop on text mining in ACM international conference on knowledge discovery and data mining (KDD).
[9] S. El Manar El Bouanani, I. Kassou "Vers une méthodologie de modélisation d'une signature unique des profils Web : Module de détection des auteurs des forums web", *JADT 2012*
[10] I. Farkhund, B. C. M. Fung, H. Binsalleeh, "Mining writeprints from anonymous e-mails for forensic investigation". digital investigation, Vol.(7): 56-64, 2010
[11] Iqbal F, et al. (2010). Mining writeprints from anonymous e-mails for forensic investigation. Digit. Investig, doi:10.1016/j.diin.2010.03.003.
[12] J. Li, H. Chen, R. Zheng «From fingerprint to writeprint". Communications of the ACM - Supporting exploratory search. Vol.(49), Issue 4, pp: 76-82, 2006
[13] K. L. Lo, M. H. Sohod, Z. Zakaria "Determination of Consumers' Load Profiles based on Two-stage Fuzzy C-Means", *Proceedings of the 5th WSEAS Int. Conf. on Power Systems and Electromagnetic Compatibility, Corfu, Greece, August 23-25, 2005 (pp 212-217)*
[14] H. Mohtasseb, A. Ahmed, "Mining Online Diaries for Blogger Identification". Proceedings of the World Congress on Engineering (WCE). London, U.K.
[15] A. Orebaugh, J. Allnutt, "Classification of Instant Messaging Communications for Forensics Analysis". The International Journal of Forensic Computer Science, Vol.(1): 22-28.
[16] D. Xu, H. Wang, Su K. "Intelligent Student Profiling with Fuzzy Models". Proceedings of the 35th Hawaii International Conference on System Sciences, 2002

[17] Y. C. Yang."Web user behavioral profiling for user identification". Decision Support Systems, Vol.(49): 261–271.
[18] I. C. Yeh, C. H. Lien, T. M. Ting, C. H. Liu, " Applications of web mining for marketing of online bookstore". Expert System with applications, Vol.(36) :11249-11256, 2009
[19] X. Zhang, J. Edwards, J. Harding , " Personalised online sales using web usage data mining". Computers in Industry, 2007, Vol.(58): 772–782.
[20] R. Zheng, J. Li, H. Chen, Z. Huang, "A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques". Journal of The American Society For Information Science And Technology, 2006, pp: 378-393.

Ismail Kassou was born in 1965. He received his PhD degree from Rouen University in 1992. In 1993, he joined the Engineering School of Computer Sciences (ENSIAS), Mohammed V Souissi University, Rabat, Morocco, where he has worked as Assistant Professor (1993), Associate Professor (2002) and Professor (2006). He has been Vice director of the engineering school of computer science since 2004. Since 2008, he has been director of the doctoral studies center in engineering sciences. His research interests are in the area of knowledge management, web and text mining and artificial intelligence.

Sara El Manar El Bouanani was born in 1982. She is an engineer in computer sciences graduated from the National School of Applied Sciences (ENSA Agadir-Morocco) in 2006. Since 2010, she is a PhD student in the Engineering School of Computer Sciences (ENSIAS), Mohammed V Souissi University, Rabat, Morocco. Her research topic is about profiling web users based on their writing on the Web.