Analyzing Social Roles using Enriched Social Network on On-Line Sub-Communities.

Mathilde Forestier, Julien Velcin, Djamel A. Zighed
Eric Laboratory, University of Lyon
Lyon, France
mathilde.forestier@univ-lyon2.fr,
julien.velcin@univ-lyon2.fr,
abdelkader.zighed@univ-lyon2.fr

***Abstract*—Analyzing the social roles inside on-line communities became a big challenge nowadays. The on-line communities formed around exchange platforms (e.g., forums) create an increasing source of data for analyzing user's behavior. This paper proposes an exploratory analysis of communities in news website based on its sub-communities. Actually, we assume that people who participate in forum debate in news websites focus their participation in one or a very few topics (also called context), i.e., they formed the sub-communities. These sub-communities, will help us to find the *contextual celebrity*: the pertinent users in the sub-communities. We based our analysis on a dataset composed by 11,143 users writing more than 35,000 posts on 57 different forums grouped in 3 topics, and on social networks enriched with relations extracted from the content of the users' posts.**

***Keywords*-Social role; Social network; On-line community.**

I. INTRODUCTION

During the Roman era, the forum was the public place of the city, i.e., the social, political and economic center. The forum allowed people to communicate, exchange, debate and socialize. Forums still exist nowadays in a different way: thanks to the forums on the Web 2.0, users communicate interactively on a common interest.

People who participate in these forums (also called users) form an *on-line community*. Schoberth et al. [1] use this term "to describe the communication and social interaction that is seen in the Internet and web-based list servers, bulletin boards, Usenet newsgroups and chats". We can complete the definition with the one given by Hymnes [2] about the speech community which represents "a group of people who share rules for the conduct and the interpretation of speech, and rules for the interpretation of at least one linguistic variety". So, people who participate in forums form an on-line speech community. People in this on-line speech community, as in the real life [3], play a social role, as define in [4] "beside having personalities, by being part of the social group, people occupy positions in the social structures of groups that allow them to do and say certain things, as well as constrain them from saying or doing other thing". Golder and Donath follow the Goffman's theory [3] through which a role represents the "rights and duties attached to a given status". Still, in a Goffman's position, Gleave et al. [5] specify that a social role can only be apprehended in the interaction, i.e., people play a role depending on others.

In this paper, we focus on the understanding of sub-communities (from a whole community) in order to find good clues to comprehend the *contextual celebrity* social role. We define a sub-community as a sub-part of a whole community depending on a topic (also called context). In other words, a sub-community represents all the users who participate in a specific topic in a news website (e.g., politic, media, etc.). We assume that users participate in one a very few topics in their interest. So, the *contextual celebrity* represents a user particularly interested in a specific kind of topic compared to the whole on-line speech community. This user is recognized as a pertinent one by the other members of his sub-community.

So, the contributions of this paper are to explore an on-line community based on the analysis of the sub-communities which belong to it. The general idea is to confirm that users participate depending on a context and find some clues to detect the *contextual celebrity* for each kind of sub-communities. Note that, in this paper, we use the term topic or context independently.

This paper is organized as follow: first, we explain some related work and we position our work according to the existent one. Then, we describe the dataset we use to make the analysis of the social role inside sub-communities. We continue by briefly presenting the construction of our enriched social network using the structure and the content of the data. Finally, we explore the on-line speech community with its sub-communities and the concept of *contextual celebrity* social role.

A. Related Work

The social role analysis was highlighted by Goffman in [3]. According to his theory, human being adopts a "pre-established pattern of action, which is unfolded during a performance and which may be presented or played through on other occasions". According to him, individuals play a role during the interaction. This notion had a great

 ISBN: 978-1-61208-176-2

repercussion through the apparition of Web 2.0 and the emergence of new media of exchange. Some researchers used database of email exchange and probabilistic model as blockmodel to define some social roles in firms [6][7][8]. Other researchers looked at predefined roles as the expert [9] (who is the most expert?) or the influencers in social network [10][11] (who gets the power to convince people in the social network). In an other perspective, computer scientists and sociologists found a great interest to analyze the social roles in forum debates. Their works aim to extract social roles as a predefined behavior in the on-line speech community using a social network analysis and the user participation behavior. This double analysis allows to capture the place of the user inside the community based on his implication and his reputation. Golder and Donath made an ethnological study and found out six kinds of social roles: the celebrity, the newbie, the lurker, the flamer, the troll and the ranter (refers to [4] for the definitions). These social roles can be positive (e.g., the celebrity) or negative (e.g., the flamer or the troll). This ethnological approach considers that a content analysis of the posts brings a lot of informations. In our work, we use a content approach to extract our social network with the aim to define the social roles. We will see in Section I-C how we enrich our social network with new relations extracted from the content of the discussion. Others social roles have been discovered in this on-line speech community as the answer people and the discussion people [12]. In a political discussion context, Himelboim et al. [13] looked for the discussion catalyst. This kind of users influences the information that enters in a newsgroup and affect the discussion evolution within it. Kelly et al. [14] found three social roles in this kind of discussion: the friends, the foes and the fringes. The authors highlighted that people prefer to speak to users who are in another political affiliation than themselves. The great majority of the users in political discussion looks for virulent debate on society and way of life. Furthermore, the authors found the fringe social role which refers to a marginal group of people that raises interesting questions for qualitative study. Fisher et al. [15] took more largely into account the context of participation to analyze social roles. According to them, the user's participation is different if he participates in help forums opposed to a flame forums. Their idea makes us think that in Usenet, there are some specialized forums for flame, for help etc. But in a news website the configuration of participation is quite different, there is no specialized forum as in Usenet, but there are some topics where users are more interested to debate in. Very close to our work, Angeletou et al. [16] and Chan et al. [17] explain some on-line sub-communities by their composition of users roles, but each sub-community represents one community: there is no overlapping, no confrontation between the sub-communities. In this paper, the context of participation is represented by the topic which the forum belongs to, e.g., politic, media, living, etc. So, we propose a new way to understand social role depending on the context in on-line sub-communities. Finally, we refer the reader to Gleave et al. [5] and Forestier et al. [18] in order to have a larger state of the art and analysis about social roles.

B. Dataset introspection

In this section, we present the data we used to analyze the sub-communities and the *contextual celebrity*. We based our analysis on the forums of the HuffingtonPost (www.huffingtonpost.com) news website. We extracted 57 forums dealing with three topics, i.e., context: politic, living and media. The dataset is composed of 19 forums of each topic. The whole dataset contains 11,443 users and 35,175 posts. Table I presents the basic statistics on each topic. The overlapping of the sub-communities implies that the sum of the users from the three sub-communities is upper than 11,443 users. Note that the on-line speech community represents all the users and we are looking to the *contextual celebrity* in sub-communities (communities depending on a context).

Table I
BASIC STATISTICS ABOUT THE PARTICIPATION ON THE THREE TOPICS

	Politic	Living	Media
# of users	4547	3667	5973
# of posts	12725	8274	14176
Average number of posts per user	2.8	2.3	2.4
% of users who exclusively participate in this kind of topic	58%	68%	65.5%
% of users having one post on all users who participate in this topic	50%	58%	54%
% of users having between]1,5] posts	39%	34%	38%
% of users having between [6,11]	7%	5.7%	5.5%
% of users having between [12,16]	2%	<2%	<2%
% of users having between [17,∞[	<2%	<1%	1%
Total	100%	100%	100%

Table I shows that it exists in all sub-communities a hard core of specific users, i.e., users who participate only in one topic. Furthermore, the ratio between the number of posts and the number of users is quite the same in the three sub-communities. Users in living topic (respectively media topic) post an average of 2.3 messages (respectively 2.4). In politic forum, the ratio is a little bigger, i.e. 2.8, posts per user. Most of the users concentrates their participation on one topic and for each sub-community, at least half of the users post just one post in one topic. This comportment seems really interesting and, even if this is not the object of this paper, and in a perspective way, the study the behavior of these users through the sub-communities and in a temporal way, can be really interesting.

 ISBN: 978-1-61208-176-2

The three sub-communities follow the same rule of participation: most of the users posts less than six messages. There is a real gap between people who write less than six messages and those who write more. For each sub-community, an average of 6% of the user post between five and ten messages. Finally, a very few users posts more than ten messages in a topic. The *contextual celebrity* is being more likely among them.

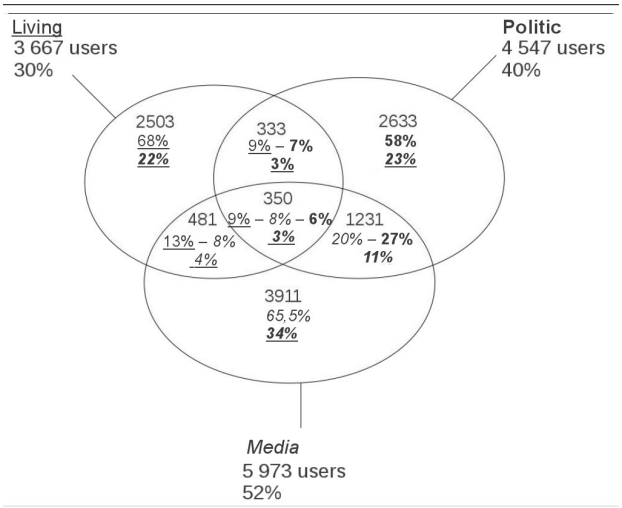

Figure 1. Overlapping of the tree topics

Figure 1 shows the overlapping of the sub-communities. The bold numbers represent the attributes of the politic sub-community, the underlined numbers represents the attributes of the living sub-communities and the ones in italic represent the attributes of the media sub-community. So, the numbers underlined, in bold and in italic represent the percentages for the whole community. The living (underlined on Figure 1) sub-community represents 30% of the community. Inside the living sub-community, 68% of the users participates only in this topic (it corresponds to 22% of the whole community). The statistics are quite the same in the two others sub-communities. A very few users participates in living topic and an other topic (9% of the users in living sub-community participates also in the politic topic (in bold), and 13% in both living and media topics (in italic). This minority represents only 3% and 4% of the whole community. Note that less than 30% of the media sub-community wrote at least one post in politic. These two sub-communities seem to be closer than the politic and living sub-communities. Finally, only 3% of the population participates in the three topics (it represents 9% of the users of living, 8% of politic and 6% of media sub-community). In conclusion, only 21% of the users participate in more than one topic.

C. Enriched social network

Web forums have the particularity that they structure the debate. Users who participate can, using this structure, reply to the post they want to reply to. This structure is used to extract social networks in existent works treating social roles [9][12][13][15]. But, reading the forums shows that people not only interact using the structure (reply to) but also through quotations. We find two kinds of quotations: the text quotation and the name quotation [19]. These two quotations allow an user to reply to several ones through one post; and people who read the forum automatically understand when an author is quoted (by the name, or by the quotation of a previous post). The idea is when a person quotes the name of another one, he adopts certain community codes and he considers himself to be entitled to refer to the person by his pseudonym, e.g., a newbie (i.e., new user) never feels the right to call other users by their pseudonym. So quoting the name implies the user's integration in the online speech community. The text quotation relation brings some important information during the analysis. Actually, more a user quotes another, more these two users are linked. Furthermore, the text quotation frequently implies a precise conversation between the two users, i.e., if I quote a part of your post, I really reply to you, and in most of case I argue your discourse with an opinion. To make a finer interaction analysis, we wanted that the analyze taking into account these quotations in a an automatically way. So, we created an enriched social network where users can be linked by three relations:

- The structural: a user replies to another one using the structure of the forum;
- The name quotation: when a user quotes the name of an other user in his post;
- The text quotation: when a user quotes a part of a previous post in his post.

Finally, we have three separates but complementary social networks, i.e., one social network for one kind of relation. Each of this social networks give some clues to understand the user behavior. The social network constructed with the name quotation relation gives some informations about the user's reputation: is he known by his sub-community? Is he often quoted? Is he often quotes? The social network constructed with the text quotation relation give some others clues: does the user like to debate? Does he bring some interesting informations to debate?

Our model reaches a quite good score in term of precision (ratio between number of quotations found by both evaluators and system compared to the number of quotations found by the system). We refer the reader to [19] for more information about the social network extraction.

Figure 2 shows the three separate social networks. We used the Jung Java toolkit for SNA to build the graphs. The social networks on Figure 2 are built only with users having written more than 15 posts in all the dataset. The gray scale of nodes represents the kind of forum a user participates in. Black nodes make the connection between subgraph of gray

 ISBN: 978-1-61208-176-2

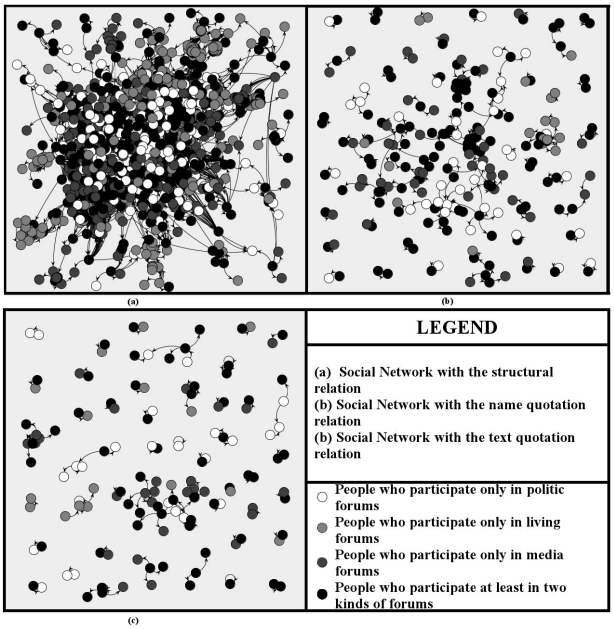

Figure 2. Social networks visualization

users (users who participate only in one kind of forum). We can also see that name quotation (b) is used more than the text quotation relation (c) by user who participate in the whole website. Referring to Table II, the name quotation represents the double of the text quotation (994 name quotations against 476 text quotations). The media sub-community uses more the name quotations compared to the two others sub-communities, we can interpret this result as a closer sub-community, i.e., the users of the media sub-community know better the others who participate in it. Surprisingly, it is not the users who post the most in a topic who quote or are quoted the most. This result is important concerning the detection of the *contextual celebrity*: users who are quoted by the name and the text and who post a lot of messages have more chance to be recognized by the others and to have a good reputation [20] in their sub-communities.

II. A NEW SYSTEM TO ANALYZE SUB-COMMUNITIES AND SOCIAL ROLES

As we saw in Section I-A, social role analysis became an important research study. Nowadays, it seems really important to understand who is who in the on-line speech communities. But, as we saw before in these works, most of the researchers use Usenet to extract social roles. The fact is that forums on news websites become increasingly generic while Usenet is quite specific. Furthermore, news websites allow users to treat several kinds of forums, e.g., politic, societal, etc. and the social role is dependent of the context[3][5]. The goal, here, is to retrieve social roles depending on the context, i.e., the kind of forum treated. Finally, the three relations between users (see Section I-C) allow a finer perception of the interaction. These relations will help us to a better extraction of the social roles.

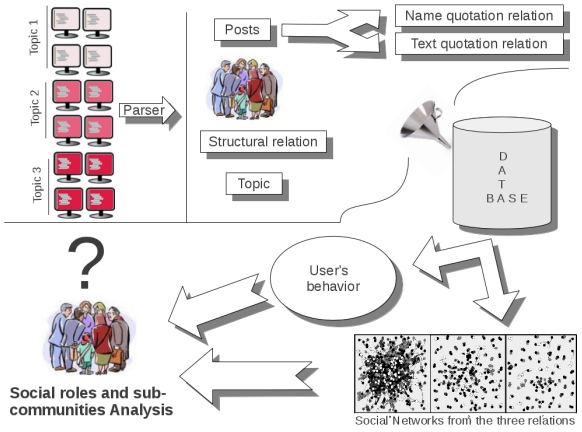

Figure 3. Presentation of the system

Figure 3 shows the process of our system from the website to the analysis. First of all, we collect the forums from the website using a parser. Note that the parser is specific for each website. The parser retrieves the forum topic, the users pseudonyms, the posts and the structural relation, i.e., which post replies to which one? Who replies to whom? Then, the system analyzes the content of the posts in order to extract the name and the text quotation relations. All the data is scored in a database. Finally, using the enriched social network (with the relations extracted from the content of posts) and the user's participation behavior, we analyze the social role based on the context, i.e., the topic of the forum.

We will present in the next section the way we choose to analyze the social roles taking into account the context and

Table II
BASIC STATISTICS ABOUT THE ENRICHED SOCIAL NETWORK ON THE THREE TOPICS. TQ : TEXT QUOTATION, NQ : NAME QUOTATION

	Politic	Living	Media
# of TQ	177	146	153
# of users who use TQ	151	119	118
# of NQ	350	183	461
# of users who use NQ	256	128	118
# of users having used TQ	151	119	118
# of users having more than 15 posts and are quoted by text	17	15	13
# of users having more than 15 posts and quoting by text	19	0	19
# of users having used NQ	256	128	375
# of users having more than 15 posts and are quoted by their name	33	15	23
# of users having more than 15 posts and quoting by the name	28	7	28

the kind of relations between users.

III. Sub-communities and social role Analysis

To analyze the sub-communities and to find the *contextual celebrity(ies)*, we decided to perform a principal component analysis (PCA)[21]. This unsupervised method aims to create a new description space of the data. We use Tanagra [22] to compute the PCA.

A. Criteria of analysis

We created several criteria to analyze the on-line community in a contextual perspective. These criteria are based on the individual's behavior and the analysis of the social network. We calculate for each individual who participate in the forums:

- Number of politic / living / media forums the user participates in;
- Number of posts in politic / living / media forums;
- In-degree with the structural relation function of the topic;
- Out-degree with the structural relation function of the topic.

So, each user is defined by 12 criteria measuring the user's interest in the topics and his place inside the sub-communities. Actually, the participation is comprehended by the user's participation behavior metrics; and his place inside the sub-community by his place in the social network using the in-degree and out-degree with the structural relation. These criteria allow us to create an unsupervised method to explore the on-line speech community.

B. Principal Component Analysis

The Principal Component Analysis (PCA) consists in transforming the criteria of analysis (see Section III-A) in new variables, each independent of each other. The aim of this method is to create a new space where the dimensions are not correlated one to the other. It also allows to reduce the information description to a limited number of components, less than the initial number of criteria of analysis. PCA is really interesting for several reasons. First of all, we want to explore the sub-communities in a unsupervised way. The social role of the users depends on the interest of the user for one topic and his place inside the sub-community. We expect that PCA finds three groups (one group for each topic) that are not correlated one to the others. Furthermore, this is an unsupervised method of analysis because our dataset does not allow the usage of supervised methods: we do not have labels to learn rather we have to discover and interpret the knowledge from the data. Finally, this old method (more than one century) made proof of its performance and it is still used today.

The first three axes found by the PCA resume 75% of the knowledge contained in the data. The fourth axis only adds 5% of supplementary information, so we keep the first three axes. Note that a resume of 75% of information is a quite good score for real data.

The first axis is described on the positive part by the politic topic: number of posts, in- and out-degree with the structural relation. On the negative side, the axis is described by the living topic. The second axis is constructed on the positive part with the politic forum. The third axis is constructed with all the criteria concerning the media topic. This construction proves that the on-line speech community is divided into sub-communities function of the topic. Nevertheless, the sub-communities are not completely separate (otherwise the PCA gives some correlations about one) and some users being part of several sub-communities.

Figure 4 represents the correlation scatter plot created with the two first axis of the PCA. The three forums are visibly separated. We have on the top left of the graphic the forums about living, on the bottom right the forums dealing with media and on the top right the forums dealing with the politic. This graphic proves that individuals have certain behavior depending on the kind of forum they participate in. Figure 4 shows that the angle between the living topic and the media topic is about 180° function of the gravity center (see the right line on figure 4 between the two groups). It means that it exists a negative correlation between the two groups. In other words, the more the users participate in forums dealing with media topic, the less they participate in living topic and vice versa. In another way, the politic topic is almost on a right angle compared to both media and living forums. There is a statistical independence between the politic topic compared to the media and the living topic. The PCA does not find a correlation between them. It seems that the participation in the politic topic does not influence the participation in media and/or living topic.

Finally, the PCA confirms that users mostly participate in one kind of topic (i.e., in a context). To find who are the *contextual celebrities* we propose to find users who maximize all the criteria on one topic and who have no or very few participation in the others. So, in a perspective way, we are thinking to use multicriteria aggregation so that we find not just one *contextual celebrity* per topic but a list

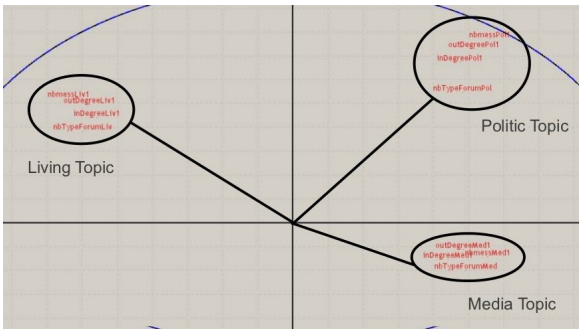

Figure 4. Correlation scatter plot

of *contextual celebrities* for each sub-community.

IV. Conclusion and future works

This paper presents a new exploratory approach to understand on-lines communities based on its sub-communities and give good clues to comprehend the *contextual celebrity* in these sub-communities. A lot of people interact on news websites, this media became increasingly widespread, the dimensionality of data makes it difficult to comprehend. We use the Principal Component Analysis (PCA) to understand how people interact starting from the hypothesis that people focus their participation in one or a very few topics (i.e., context) and not in the website as a whole. The PCA confirmed this hypothesis. This exploratory method finds three kinds of groups defined by the kind of context the users participates in.

The *contextual celebrity*, i.e., a user who participates in one kind of topic and be recognized by his sub-community as a pertinent user, needs to maximize the criteria in one topic. Furthermore, using an enriched social network allows a finer perception of the real interaction between users and brings interesting informations to characterize the community, the sub-communities and the *contextual celebrity* himself.

In perspective, we want to extract the *contextual celebrity* and evaluate the model using a temporal evaluation. We are also interested in the analysis of the users who participate a few (one post) in one topic Who are these people? Why do they participate so little?

References

[1] T. Schoberth, J. Preece, and A. Heinzl, "Online communities: a longitudinal analysis of communication activities," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.* IEEE, pp. 1–10.

[2] D. Hymes, *Foundations in sociolinguistics: An ethnographic approach*. Psychology Press, 2003.

[3] E. Goffman, *The presentation of self in everyday life*. Harmondsworth, 1978.

[4] S. Golder and J. Donath, "Social roles in electronic communities," *Internet Research*, vol. 5, pp. 19–22, 2004.

[5] E. Gleave, H. Welser, T. Lento, and M. Smith, "A conceptual and operational definition of'social role'in online community," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 2009, pp. 1–11.

[6] F. Lorrain and H. White, "Structural equivalence of individuals in social networks," *Social networks: a developing paradigm*, vol. 1, p. 67, 1977.

[7] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 249–272, 2007.

[8] A. Wolfe and D. Jensen, "Playing multiple roles: Discovering overlapping roles in social networks," in *Proceedings of the 21st International Conference on Machine Learning, Workshop on Statistical Relational Learning and its Connections to Other Fields.*, 2004.

[9] J. Zhang, M. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proceedings of the 16th International conference on World Wide Web*, 2007, pp. 221–230.

[10] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *WSDM '08: Proceedings of the international conference on Web search and web data mining*. New York, NY, USA: ACM, 2008, pp. 207–218.

[11] P. Domingos, "Mining social networks for viral marketing," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80–82, 2005.

[12] H. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the signatures of social roles in online discussion groups," *Journal of Social Structure*, vol. 8, no. 2, 2007.

[13] I. Himelboim, E. Gleave, and M. Smith, "Discussion catalysts in online political discussions: Content importers and conversation starters," *Journal of Computer-Mediated Communication*, vol. 14, no. 4, pp. 771–789, 2009.

[14] J. Kelly, D. Fisher, and M. Smith, "Friends, foes, and fringe: norms and structure in political discussion networks," in *Proceedings of the 2006 international conference on Digital government research, May*, 2006, pp. 21–24.

[15] D. Fisher, M. Smith, and H. Welser, "You are who you talk to: Detecting roles in usenet newsgroups," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, 2006, pp. 59b–59b.

[16] S. Angeletou, M. Rowe, and H. Alani, "Modelling and analysis of user behaviour in online communities," *The Semantic Web–ISWC 2011*, pp. 35–50, 2011.

[17] J. Chan, E. Daly, and C. Hayes, "Decomposing discussion forums and boards using user roles," in *AAAI Conference on Weblogs and Social Media*, 2010, pp. 215–218.

[18] M. Forestier, A. Stavrianou, J. Velcin, and D. A. Zighed, "Roles in social networks: Methodologies and research issues," *Journal of Web Intelligence and Agent Systems*, p. To appear, 2011.

[19] M. Forestier, J. Velcin, and D. Zighed, "Extracting social networks to understand interaction," *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2011)*, pp. 213–219, 2011.

[20] J. Donath, "Identity and deception in the virtual community," *Communities in cyberspace*, pp. 29–59, 1999.

[21] K. Pearson, "Principal components analysis," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 6, no. 2, p. 559, 1901.

[22] R. Rakotomalala, "Tanagra: un logiciel gratuit pour lenseignement et la recherche," *Actes de EGC*, vol. 2, no. 3, pp. 697–702, 2005.

 ISBN: 978-1-61208-176-2