Extracting social networks to understand interaction

Mathilde Forestier, Julien Velcin, Djamel Zighed
ERIC Laboratory, University of Lyon 2
5, avenue Pierre Mendès France, 69676 Bron Cedex, France
Email: firstname.name@univ-lyon2.fr
Telephone: +33 (0)4 78 77 31 54
Fax: +33 (0)4 78 77 23 75

Abstract—Web forums are a huge data source. They allow people to interact with unknown individuals. Studying forums shows that the interaction is not obvious only through the structure but also through the content of the post. Taking into account this observation, we extract a social network with different kinds of relationships i.e. the structural relation, the name and the text quotations relation. We present here the promising results we obtain, and the difficulties we face while extracting the quotations in this kind of textual content. These results are obtained from real data (from two information websites) which make the validation difficult. So, we create a validation protocol composed of two steps and based on human raters. Finally, we will see the objective of this work which is understanding interactions in order to extract the social roles of individuals.

I. INTRODUCTION

The interaction between people has experienced many changes: from carrier pigeon to email, the communication media has evolved a lot. However, it seems that people need to interact and communicate and the Web 2.0 has brought a certain number of dedicated media (e.g. blogs, forums, etc.). This kind of social media enable individuals to communicate freely and easily. This huge volume of data motivates the researchers to find new ways of data modeling and representation. For example, within our object of analysis, the forums on the web, can be represented as an opinion graph [1], or as a social network [2], [3]. In this paper, we present a new way to model interactions in forums thanks to a social network with different kinds of relationships between the individuals who participate.

Web forums are a very interesting source of data for several reasons. On the one hand, forums represent a wide variety of topics treated e.g. economy, politics. On the second hand, web forums are available from the Internet which points out a quite easy data repatriation. On the third hand, the content of forums is very interesting to analyze, although at the same time the amount of information is too big to be processed manually. And last but not least, this medium allows users to interact thanks to a predefined structure. Indeed, participation in web forums depends on personal conviction and involvement in the community [4]. They gain a reputation inside the community thanks to their posts, their ideas, their participation. In this context, we can cite an article from a french information website: a user writes an article with the title “I became Charles Mouloud, virtual being of flesh and blood”¹. This title expresses very well the notion of personal implication in the virtual community. Charles Mouloud is a virtual user, but he is recognized as a real person by the other users of the community. In fact, in discursive communities, the users have a social role i.e. a position taken by an individual while interacting with others.

In the aim to extract the social roles of individuals, we first need to model the interaction. People interact within forums thanks to the website HTML structure i.e. forums generally allow users to reply specifically to a post written by an other user through a physical link on the webpage. However, reading the content of forums reveals other kinds of relationships between users. For instance, they can quote each other in two ways. The first way is through a user name quotation i.e. a user mentions in his post the pseudonym of another user. The second one is the text quotation i.e. a user repeats in his post a part of another user’s post. These two kinds of relationships are really important to extract a good representation of the relationship between users. The difficulty of this work is to extract these two relationships considering the quality of texts. Individuals who participate in these forums do not always use typographic rules (e.g. a misuse of quotations marks), they make some semantic changes on the name quotation, etc.

This paper discusses two contributions of our work. The first one is considering interaction not only by the structural way but also by adding an analysis of the post content. It implies that individuals in a social network are linked with each other through different relationships: the structural relation, the name and text quotation. The major contribution is to identify the quotations in the forum posts. The second one is to explain the importance of using several validations concerning non-labeled data to obtain a trustful corpus to evaluate our model.

In this paper, we begin by presenting the related work concerning the social network extraction. We continue to explain the protocol of the validation and the system overview. In the fifth and sixth part we introduce the interaction by extracting the name and text quotations. Finally, we discuss a state of the art regarding social roles that can take benefit from the previous system.

¹http://www.rue89.com/les-moulouderies-de-charles/2010/11/28/je-suis-devenu-charles-mouloud-etre-virtuel-de-chair-et-de-sa
II. RELATED WORK

The state of the art about social network extraction is quite important. We present here a reduced part of the existing works which seems to us really interesting for our work objectives.

With the increasing volume of textual data, a lot of researchers aim to extract a social network in order to model them. The email databases consist a highly used source. Emails allow linking people thanks to the header containing one sender and at least one recipient. We can cite several works using these email databases [5], [6]. Mika imagines Flink [5], a complete system which runs in three phases: data acquisition (from the web, FOAF, emails and publications), data storage and aggregation and, finally, the social network visualization. Note that in this system, the nodes of the social network are given. Culotta et al. [6] extract a social network thanks to an email database and the content of the web. The system searches the sender, the recipient(s) and the names contained in the email. Then, the system searches on the web the personal web page for each name to extract key words about his working domain. Finally, thanks to a graph partitioning algorithm, the system groups researchers into communities.

Others researchers use only the web to extract a social network. Kautz et al. [7] create a system called ReferralWeb, that allows to extract a social network from several resources (links found on home page, co-authors and citations of papers, exchange between individuals in netnews archives and, organization charts). Matsuo et al. [8] create Polyphonet. This system allows to extract a social network from a community of researchers. This system recognizes four types of possible relationships between two actors: Co-author (when a paper is written by several authors), lab (when two people belong to the same laboratory), Proj (when two people are involved in a project or even a committee), and finally Conf (when two researchers participate in one conference). As in Flink, the nodes are given.

Jin et al. ([9], [10]) think that social network extraction concerns different functions of the studied population i.e. a social network extraction about researchers cannot be the same as a social network extraction about companies. In the light of this observation, they extract two social networks: one for the Japanese firms, the other for Japanese Artists. Note that they extract several relationships between actors e.g. alliance or litigation for firms.

Finally, some researchers are interested in social network extraction from Usenet forums [2], [3], [11] or from the Java Forum [12]. In these works, the social network is extracted: nodes are the individuals who participate in discussions and the relations are extracted thanks to the structure (who replies to whom). They do not use the data content to extract other kinds of relations between users. They also use sometimes a projection of the social network on one actor with his neighbors at a certain distance.

All the works presented here, show that models rely on the data structure and content for the email database source and just the structure for forums. Nevertheless, forums contain implicit relationships that are not taken into account in these works. In this way we create a new model that extracts several kinds of relationships between users thanks to the structure and the content of the data.

III. THEORETICAL FRAMEWORK AND SYSTEM OVERVIEW

We present in this section the theoretical framework for the extraction of our social network and the system overview. The social network has to integrate the several relationships we want to extract between users: the structural, text and name quotation relationships.

A. Theoretical framework

We define a theoretical framework based on three sets to model our social network:

- X: the set of authors \( X = \{x_1, \ldots, x_n\} \) where \( n \) is the number of authors in the forum.
- R: the set of relationships. \( R \) is a finite set of three relationships \( R = R_{str} \cup R_{text} \cup R_{name} \) respectively structural, text quotation and name quotation relationships.
- D: the set of documents (each post represents one document) with \( D = \{d_1, \ldots, d_m\} \) where \( m \) is the number of posts in the forum. Note that \( D \) is partitioned in \( T \) groups where \( T \) represents the number of threads in the forum. A thread is a part of the discussion where the posts reply to each other using the structural relation. Finally, a forum contains a time dimension so \( d' \prec d \) if \( d' \) is published before \( d \).

From these three sets, we can define the following mapping:

- **author**: \( D \rightarrow X \): the post \( d \) is written by the author \( x \)
  \[ d \mapsto x \]

Knowing that \( \delta \in \{str, text\} \), the following binary relationships stand:

- \( d_a R_{\delta} d_b \Leftrightarrow \) document \( d_a \) replies to \( d_b \) with the \( \delta \) relationship.
- \( d_a R_{name} x \Leftrightarrow \) document \( d_a \) quotes the pseudonym \( x \).

And the following relationships between authors:

- \( x_i R_{str} x_j \Leftrightarrow \exists d_a, d_b \in D \times D, d_a R_{str} d_b \) and \( \text{author}(d_a) = x_i \) and \( \text{author}(d_b) = x_j \)
- \( x_i R_{name} x_j \Leftrightarrow \exists d \in D, d R_{name} x_j \) and \( \text{author}(d) = x_i \)

We define a graph \( G = (X, A) \) where \( X \) represents the set of authors and \( A \) the set of directed edges. Each directed edge \( a_{ij} \) represents a reply from an actor \( x_i \) to \( x_j \) with the relation \( r \in \{str, text, name\} \).

B. System overview

Figure 1 shows the system overview from online data to the social network visualization. The first step is to retrieve data from the html page. To this extent we create two parsers (one for rue89², the other for the Huffington post³). These

²http://www.rue89.fr
³http://www.huffingtonpost.com/
parsers aim to recognize the actors names, the posts and the structural relationships thanks to the HTML structure. All this information is stored in a database. The second step consists in two modules:
- The extraction of the name quotation relationship: when an author is quoted in the body of a post (see section V).
- The extraction of the text quotation relationship: when a part of a post is repeated in another post (see section VI).

The last stage uses all actors and relationships to create the social network.

### IV. EXPERIMENTAL FRAMEWORK OF INTERACTION EXTRACTION

Forums allow people interaction thanks to the website HTML structure. This structure shows to the reader who replies to whom thanks to the gap between two posts (see in Figure 2: Alvarask replies to popart) and allows to well understand the conversation between users. However, reading forums reveals that there are several types of interaction that may exist between users. The first one has been explained beforehand: the structural relation. Another relation is the name quotation i.e. when a user quotes the name of another user. The third type of interaction is the text quotation appearing when a user quotes a part of a post written by another user. To well represent the data structure and content, we decide to extract a social network where users are the actors, and the relationships are the three kinds of interaction.

To measure the performance we calculate the recall (number of quotations found by both evaluators and system compared to the number of quotations found by the evaluators), the precision (number of quotations found by both evaluators and system compared to the number of quotations found by the system) and the F-measure (harmonical average of recall and precision) to have a performance overview.

As expected, the precision increases for each forum for all forums with the adjusted validation i.e. the system found some quotations that escaped the evaluators during the first reading see Figure 3. The precision increases because the evaluators did not see some quotations. It is possibly due to the forum length: the task of rating a forum was quite long. Looking at the increase of the precision for each forum, it is necessary to make this second evaluation to have a trustful corpus to evaluate our model. In fact, the adjusted validation modifies both precision and rappel. Sometimes just one evaluator found a quotation during the first evaluation. Thanks to the second evaluation, one or both of the two other evaluators found

### A. Analyzed forums

In order to deal with the lack of labeled data, we decide to analyze four forums from two sources: two forums are from a French information website and the two others come from an American information website. The first french forum talks about the foreign policy of the president Sarkozy and the second one talks about a file about Roma People. The first American forum deals with the faith and the second one is about the Diabet disease. All the forums contain more than 350 posts each.

### B. Validation protocol

The validation was a real problem: how to validate the results when you do not have some labeled data? How to measure the experiment performance? The best option was using human evaluators. For each analyzed forum, we asked three people to rate the results. They had to read the forum and note every text or name quotations they found out. We called this step, the non-adjusted validation. In this step, we keep all quotations found by at least two evaluators.

Every studied forum was quite long (more than 350 posts each) and the attention of the evaluators could sometimes be less efficient. In the light of this observation, we decide to make a second validation called adjusted validation: we create a new file with the quotations found by the system and those found by just one evaluator. We chose to present a second time to the evaluators the quotations found by at least one evaluator in order to be sure that it is not an oversight of the two other evaluators. We gave this new file to the evaluators and they had to answer whether they found a quotation or not. We consider that in that way, the chances that the three evaluators oversaw quotations are minor. In other words, this second evaluation allows to be sure that evaluators do not inadvertently miss some quotations. Finally, we keep the quotations found by at least two evaluators during the non-adjusted validation and those find by at least two evaluators during the adjusted validation.

To measure the performance we calculate the recall (number of quotations found by both evaluators and system compared to the number of quotations found by the evaluators), the precision (number of quotations found by both evaluators and system compared to the number of quotations found by the system) and the F-measure (harmonical average of recall and precision) to have a performance overview.

As expected, the precision increases for each forum for all forums with the adjusted validation i.e. the system found some quotations that escaped the evaluators during the first reading see Figure 3. The precision increases because the evaluators did not see some quotations. It is possibly due to the forum length: the task of rating a forum was quite long. Looking at the increase of the precision for each forum, it is necessary to make this second evaluation to have a trustful corpus to evaluate our model. In fact, the adjusted validation modifies both precision and rappel. Sometimes just one evaluator found a quotation during the first evaluation. Thanks to the second evaluation, one or both of the two other evaluators found quotations.
as right this quotation, even if the system did not find it. Thus, with the adjusted validation, the precision increases and the recall can sometimes decrease. Note that all future results shown in this work are evaluated based on the adjusted validation.

V. EXTRACTING NAME QUOTATION RELATIONS

Name quotations can be easily retrieved if the user name is correctly written. Our parser retrieves the list of all the participants in the forum thanks to the HTML structure. However reading forums shows that the name quotation is not that trivial for the following reasons:

• the name is badly spelt in the post;
• a single part of a compounded name is written in the post;
• the name is written by an abbreviation, a diminutive, a synonym.

Based on this observation, we propose an algorithm that takes into account the spelling errors in a user name, and the case when just a part of a compounded name is not mentioned. To explain the process, see Algorithm 1.

Algorithm 1 Name quotation Extraction

Require: D, X, α, dist, dictionary
1: for each d ∈ D do
2:    X′ ← {x ∈ X, ∃d′ ∈ D/author(d′) = x and d′ < d}
3:    x ← author(d)
4:    for each x′ ∈ X do
5:        x′ is (w₁, w₂, ..., wₖ)
6:        if ∃w ∈ x′/ w ∈ d then
7:            R ← R ∪ xRnameₓ'
8:        else if ∃wᵢ ∈ x', wᵢ ∈ d/dist(wᵢ, wⱼ) < α and dictionary(wⱼ) = 'unknown' then
9:            R ← R ∪ xRnameₓ'
10:       end if
11:    end for
12: end for

To catch spelling errors we use the Levenshtein Distance [13] (see algorithm 1 at line 8) in two ways. The Levenshtein distance allows to calculate the number of letters to add, modify, or delete when going from one sentence to another.

The first way is to calculate it “pure”. The second way is to normalize it with the length of the user name that is compared to. In the following experiments, the threshold for the Levenshtein distance result had to be inferior to 3 and inferior or equal to 0.3 for the normalized Levenshtein distance.

Furthermore, we use a dictionary to know whether a word exists or not. Reading forums shows that users usually use non-existent words as pseudonym. Thus, when a word is not recognized by the dictionary, there is a better chance that it can be a user name. In our experiments we use the dictionary contained in TreeTagger [14]. TreeTagger is a morphosyntactic tool. It allows to retrieve the kind of a word (verb, noun, adverb etc.) and when a word is not known by the dictionary the term “unknown” is associated to it.

Figure 4 shows the improvement of the results when using the normalized Levenshtein distance and the dictionary of TreeTagger. For all the four forums, the association of the three tasks (or at least the normalized Levenshtein distance and the dictionary) increases the performance of the system. Note that the results, with just the Levenshtein distance, are really bad.

Table I shows the performance overview. For all forums, the precision is better than the recall i.e. the system does not make a lot of mistakes but it does not retrieve all the quotations found by the evaluators. Furthermore, some name quotations are really difficult to retrieve in particular the synonymic changes. For example in the forum about the French president Sarkozy, a user is called “the.clam” and in a post, another user wrote “my dear gastropod” in reference to the first user. This

<table>
<thead>
<tr>
<th>Name Quotation</th>
<th>Sarkozy</th>
<th>Roma</th>
<th>Quiet Faith</th>
<th>Diabet</th>
</tr>
</thead>
<tbody>
<tr>
<td># of quotations</td>
<td>11</td>
<td>25</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>recall</td>
<td>0.688</td>
<td>0.64</td>
<td>0.8</td>
<td>0.329</td>
</tr>
<tr>
<td>precision</td>
<td>0.815</td>
<td>0.842</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.746</td>
<td>0.727</td>
<td>0.889</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table I shows the performance overview. For all forums, the precision is better than the recall i.e. the system does not make a lot of mistakes but it does not retrieve all the quotations found by the evaluators. Furthermore, some name quotations are really difficult to retrieve in particular the synonymic changes. For example in the forum about the French president Sarkozy, a user is called “the.clam” and in a post, another user wrote “my dear gastropod” in reference to the first user. This
kind of quotation cannot be easily found. The two options we thought about this case is for the first one to add an ontology, and the second one to add a morphosyntactic labeling e.g. in this case “my dear + name”. These two options seem very time consuming functions of the number of quotations they can retrieve. At this moment, we are still looking for another option that can be less time consuming.

VI. EXTRACTING TEXT QUOTATION RELATIONS

The text quotation between two posts brings the certitude that people really interact. In fact, when I quote a part of your post, it is to comment, to give my opinion, so I reply to you personally. To this extend, it seems really important to extract this relation which reinforces the structural relationship. First of all, we need to define what is a text quotation in a forum. A text quotation is when a part of a post is present in another post. There are different ways to find quotations: the easier way is to search the quotation marks in the post. But in this kind of text, people do not always respect the typographical rules and they sometimes do not use quotation marks.

The aim of the text quotation extraction is to find quotations between two posts, even if there is no quotation marks. In other words, the system has to find similar sequences of words between two posts. For this purpose, we had to compare each post with others. However, given the forum length, it seems very time consuming. Reading forums help us to find a way: text quotation often comes with structural relation. Based on this observation, we decide to compare a post with the one it replies to with the structural relation. We add to the algorithm which just searches quotation marks, the comparison of post. The pseudo-code of the text quotation extraction is described in the Algorithm 2.

Algorithm 2 Text quotation Extraction

Require: D, X, ω, Rstr, sim, ω
1: for each d ∈ D do
2: x ← author(d)
3: \{Quotation marks\}
4: if d contains quotation marks then
5: s ← sentence between quotations marks
6: for each d′ ∈ D / d′ < d do
7: x′ ← author(d′)
8: if s ∈ d′ then
9: R ← R ∪ xRtextx′
10: end if
11: end for
12: \{Comparison of posts\}
13: d′ ← dRstrd′
14: x′ ← author(d′)
15: if sim(d, d′) ≥ ω then
16: R ← R ∪ xRtextx′
17: end if
18: end for

The last question we have to answer is: how many following words in common do we need to extract a text quotation (see algorithm at line 16)? To answer to this question, we made some experiments, as shown in Figure 5. On the abscissa, we have the minimum number of words and in the ordinate, the results of the F-measure with the adjusted validation. From the figure 5, we can see that two posts need to have at least six words in common. The number eight can be also interesting, it increases the F-measure for the forum about the Roma File; it does not change the results for the forums about the French president Sarkozy and the Diabet disease. However, the decrease of the F-measure for the forum about the faith, convinces us to choose a minimum of six words in common between two posts.

Fig. 5. Evolution of F-measure functions of the numbers of words in common between two posts

The figure 6 shows the improvement of the F-measure by adding comparison of posts in the algorithm (see line 13 to 18 in Algorithm 2). For each forum, the F-measure increases.

Fig. 6. Improvement of the F-measure thanks to the comparison of posts.

Table II shows the system performance overview (with ω = 6). For each forum, the recall is quite good (between 0.85 to 1) i.e. when the system finds a text quotation, it is mostly a good one. However, the precision is worse. Actually, the evaluators find some quotations less than six words, sometimes even just one word. We call this kind of quotation a concept of strong sense i.e. it makes reference to a sharing of knowledge between users. For example in the beginning of the forum about the Roma file, a user tackles the Godwin theory (theory according to which the bigger the forum is, the more the probability to approach the Second World War and/or Adolf Hitler aims towards one) and this theory is getting back in the

Table II: System performance overview (with ω = 6).

<table>
<thead>
<tr>
<th>Forum</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roma File</td>
<td>0.88</td>
<td>0.69</td>
</tr>
<tr>
<td>Quiet Faith</td>
<td>0.89</td>
<td>0.67</td>
</tr>
<tr>
<td>Diabet</td>
<td>0.87</td>
<td>0.66</td>
</tr>
</tbody>
</table>

The figure 6 shows the improvement of the F-measure by adding comparison of posts in the algorithm (see line 13 to 18 in Algorithm 2). For each forum, the F-measure increases.

Fig. 6. Improvement of the F-measure thanks to the comparison of posts.

Table II shows the system performance overview (with ω = 6). For each forum, the recall is quite good (between 0.85 to 1) i.e. when the system finds a text quotation, it is mostly a good one. However, the precision is worse. Actually, the evaluators find some quotations less than six words, sometimes even just one word. We call this kind of quotation a concept of strong sense i.e. it makes reference to a sharing of knowledge between users. For example in the beginning of the forum about the Roma file, a user tackles the Godwin theory (theory according to which the bigger the forum is, the more the probability to approach the Second World War and/or Adolf Hitler aims towards one) and this theory is getting back in the
discussion thread. But at this time, we do not find a good way to extract this quotation. Perhaps by using a data dictionary with a TF-IDF measure, we will be able to extract them. Furthermore, some quotations without quotation marks contain less than six words, but we had to make a choice. In fact, if we reduce that number, the recall can increase but the precision can drastically reduce. Some other quotations contain several parts of a post e.g. in the forum on the diabet, the actor Lawson Meadows quotes the actor Singermuse: ‘food-like substances, fast food feedlots and feeding troughs’ (Lawson Meadows) quotes “‘food-like substances’, fast food feedlots and feeding troughs’ (Singermuse). The additional quotation mark after “substances” in the text of Singermuse prevents the detection of the quotation by the system. To increase the performance we can add a threshold to measure the similarity between two sentences for the quotations with quotations marks.

TABLE II
NUMBER OF QUOTATIONS FOUND BY AT LEAST TWO EVALUATORS(§), RECALL, PRECISION AND F-MEASURE FOR THE FOUR FORUMS

<table>
<thead>
<tr>
<th>Text Quotation</th>
<th>Sarkozy</th>
<th>Roma</th>
<th>Quiet Faith</th>
<th>Diabet</th>
</tr>
</thead>
<tbody>
<tr>
<td># of quotations</td>
<td>32</td>
<td>46</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>recall</td>
<td>0.688</td>
<td>0.87</td>
<td>0.375</td>
<td>0.455</td>
</tr>
<tr>
<td>precision</td>
<td>0.846</td>
<td>0.976</td>
<td>0.857</td>
<td>1</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.759</td>
<td>0.92</td>
<td>0.522</td>
<td>0.625</td>
</tr>
</tbody>
</table>

VII. MODELING INTERACTION TO EXTRACT SOCIAL ROLES

In this section, we analyze the importance of extracting several relationships between forum users and the objective of this work: extracting social roles from forum users. Extracting several kinds of relations (as we saw above) allows to take into account a part of the data content. In this way, we understand better the relations between the participants with the purpose of having a better extraction of social roles. The notion of a social role appears in sociology and represents the position taken by an individual in relation to others. In fact, forum users on a website are a community. Each user of a website community has a social place which functions of his participation, his reputation, the way he brings important ideas or not. This social role can be apprehended only in the interaction between individuals.

Golder and Donath [15] made an anthropological study of social roles in Usenet community. They found out six kinds of social roles: the celebrity, the newbie, the lurker, the flamer, the troll and the rantor. All these social roles have a certain kind of participation, of reputation functions of the others. In the same way, researchers are interested in social roles in one kind of forums. In political newsgroups, Himelboim et Al. [16] define the Discussion Catalyst and Kelly et Al. [3] search three types of social roles: the fighter, the friendly and the fringe. Some researchers work on finding experts in technical forums as Zhang et al in Java forums [12] or Cozzi et al. in an email conversation [17]. Finally other researchers are interested in finding the influencers i.e. individuals who influence others inside the social network as in [18], [19]. Finally, Mc Callum et al [20] propose a model to find social roles in a social network. This work is quite different from those cited above from his definition of the social role i.e. it is more a social position in a predefined structure (e.g. a firm) than a social position acquired by an individual through his participation, his interest. But all those works are based on the interaction between individuals to understand their social roles.

Regarding all works about social roles, it seems very important to know who is who in the discursive community. The several relationships extracted by the system can perform the research: when I quote you by your pseudonym, I know you, I refer to you personally i.e. it reinforces the structural relationship. Furthermore, we evaluate our model on four forums, but the next step is to repatriate a lot of forums with the section they belong to: world, politic, economy... The aim is to identify social roles considering the context in which people participate. The social roles analysis allows for example to recognize interesting users in a huge mass of data.

VIII. CONCLUSION

Extracting a social network by taking account the structure and the content of the data is not a simple task as we could see in this work. The kind of data treated (forum posts) makes the quotation extraction difficult. However, thanks to several measures (e.g. Levenshtein distance), and tools (e.g. TreeTagger) we can obtain satisfactory results. To improve the results, we can for example add a similarity measure for text quotations with quotations marks. This improvement allows to reduce the cut and paste mistakes. We are also thinking to try others measures than the Levenshtein distance in the name quotation extraction.

Furthermore, the lack of labeled data was a real problem and the evaluation with human raters can bring some bias. The second evaluation, that we called adjusted validation, allows us to have a trustworthy corpus. Indeed, the complementarity of the two steps of the validation allows to reduce human raters oversights.

The different kinds of relations extracted in this work allows to have a better understanding of what happens in the forum. The name and text quotations extraction strengthen the structural relation and bring a finer modeling of the interaction between individuals.

Finally, this interaction modeling brings us to the last part of our work which deals with the social role extraction thanks to this new model. This part seems very interesting: knowing who is who in a discursive community can help the forum reader and the user in several ways. On the one hand, it allows to understand the discursive community itself. On the other hand, identifying who is whom allows to select the posts from people that are considered to be interesting by the community. Finally, it can help new users to qualify the individual who they are talking to. For all these reasons we are very interested in the analysis of the social roles.
REFERENCES


[17] I. Eiron, A. C. Byron Dom, and Y. Zhang, “Graph-based ranking algorithms for e-mail expertise analysis,” in SIGMOD workshop on Research issues in data mining and knowledge discovery (DMKD), 2003.

