Collaborative Content-Based Method for Estimating User Reputation in Online Forums

Amine Abdaoui[1], Jérôme Azé[1], Sandra Bringay[1] and Pascal Poncelet[1]

[1] LIRMM B5 UM CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier, France
{amin.abdaoui, jerome.aze, sandra.bringay, pascal.poncelet}@lirmm.fr

Abstract. Collaborative ratings of forum posts have been successfully applied in order to infer the reputations of forum users. Famous websites such as *Slashdot* or *Stack Exchange* allow their users to score messages in order to evaluate their content. These scores can be aggregated for each user in order to compute a reputation value in the forum. However, explicit rating functionalities are rarely used in many online communities such as health forums. At the same time, the textual content of the messages can reveal a lot of information regarding the trust that users have in the posted information. In this work, we propose to use these hidden expressions of trust in order to estimate user reputation in online forums.

Keywords: Trust, reputation, online forums, social networks.

1 Introduction

Online forums are areas of exchange generated by their own users. Therefore, the veracity and the quality of the posted information vary wildly according to their author. With the massive and rapid growth of these conversational social spaces, it becomes very difficult for human moderators to separate good posts from bad ones. Consequently, more and more forums are implementing automated trust and reputation metrics to infer the trustworthiness of posts and the reputation of their authors. These metrics vary from ranks based on a simple post count to more elaborated reputation systems based on collaborative ratings. If the first category of metrics tries simply to reward users according to the number of their posts, the second category uses collaborative intelligence to rate a user's posts and then aggregate these ratings to give him a reputation value [1]. This idea has been successfully applied in many online forums such as news groups (*Slashdot*[1]*)*, question-answering websites (*Stack Exchange*[2]*)*, etc. However, collaborative rating is not so popular in other communities such as health forums, where users prefer to post a new message in order to thank each other rather than clicking the 'like' or 'vote up' button. The objective of this

[1] http://www.slashdot.org/

[2] http://www.stackexchange.com/

work is to use this implicit collaborative intelligence hidden in the textual content of the replies in order to infer user reputations.

Many definitions of trust and computational trust exist in the literature [2], [3]. Here we define the trust that a user *A* has in another user *B* as: "*the belief of A in the veracity of the information posted by B*", and the reputation of a user *A* as "*the aggregation of trust values given to user A*". To infer such trust from textual replies and aggregate user reputations, we need to know both the recipient of each forum message and the trust expressed in it. However, the forum structure does not always provide explicit quoting or direct answering functionalities. Besides, when these functionalities are provided, many users prefer posting a message answering the whole thread rather than a one answering or quoting another specific message. In order to deal with this issue, we propose a rule based heuristic to extract an interaction network where the nodes are the users and the edges are the replying posts. Regarding the semantic evaluation of each post's content, the features that we are looking for are agreement and valorization for trust, and disagreement and depreciation for distrust. The rest of posts are considered as neutral. Finally, we propose a metric to aggregate trust and distrust replies that a user receives and infer his reputation in the forum. The proposed reputation metric considers propagation aspects by giving more weight to the replies posted by trusted users and less to the replies posted by untrusted ones.

The rest of the paper is organized as follows: section 2 presents a summary of related work that match our methods. Section 3 gives the theoretical framework, presents the corpus of our study and describes the proposed approach. Section 4 presents and discusses the obtained results using manual annotations. Finally, section 5 gives our main perspectives.

2 Related work

Most of the methods found in the literature in order to extract interaction networks from online communities use the HTML structure of the web page [4]–[7]. They try to identify explicit message quoting. However, explicit quoting functionality is not always provided in online forums, and even when it exists many discussion participants do not use it. Moreover, a message may have many recipients. Consequently, posting it as an answer to another specific one may be insufficient. Gruzd and Haythornthwaite [8] presented an automatic approach to discover and analysize social networks from threaded discussions in online courses. The authors proposed a Name Entity Recognition system to extract name mentions inside the textual content of posts. After a preprocessing step (removing quotations, stop words, etc.), their method used a dictionary of names combined with manually designed linguistic rules. Another textual based method has been proposed by Forestier et al. [9] to extract a network of user interactions. They suggested to infer three types of interactions: structural relations, name citations, and text quotations. While structural relations can be inferred directly from the structure of the forum, name citations and text quotations require analyzing the textual contents. First, name citation relations have been extracted by searching pseudonyms of authors inside the posts. Then, text quotations are extracted by comparing sequences of words inside a message and the messages that have been posted before in the same thread.

On the other hand, existing trust metrics dealing with online forums can be organized in two main categories: structure-based trust metrics and content-based trust metrics. The first category focus on the structure of the website (including the number of postings, the distance between messages, quotes, citations, etc.) [10], while the second one use the textual content of messages to infer trust and reputation. For example Wanas et al. [11] automatically score posts based on their textual content. Their method is inspired from forums that use collaborative intelligence to rate posts. They tried to model how users would perceive a post as good or as bad. However, unlike Wanas, we believe that the textual content of the messages that reply to a user's post may reveal a lot of information regarding the trust or the distrust that the other users have in this post and therefore in its author. Consequently, instead of inferring a user's reputation from his own posts, we suggest to consider the messages replying to his posts. Moreover, we would like to give more importance to a reply made by a trusted user and less to a reply made by an untrusted user. A large effort has been done to include propagation aspects in order to rank webpages [12]. Similarly, we propose a reputation metric that include these propagation aspects.

3 Materials and methods

3.1 Corpus of study

CancerDuSein.org is a French health forum specialized in breast cancer. 1,050 threads have been collected which amounts 16,961 messages posted by 675 users. It represents all the data that have been posted between October 2011 and November 2013. This forum allows users to thank each other using a "like" button, but this functionally is rarely used. Less than 1.4% of messages received at least one "like". On the other hand, *CancerDuSein.org* gives a rank to each user based on the number of posts since his registration. However, we believe that these ranks are not sufficient to infer reputations.

3.2 Theoretical framework

Let $G = (V, E, t, r)$ be a multigraph where: V is the set of users, E is the multiset of 'reply-to' edges between these users, t is a function that returns the transmitter of a reply, and r is a function that returns the recipient of a reply:

$$\begin{aligned} t &: E \to V \qquad & r &: E \to V \\ e &\to t(e) \qquad & e &\to r(e) \end{aligned}$$

Let $v \in V$ be a user. Then $E_v \subseteq E$ is the set of edges that reply to the user v:

$$E_v = \{e \in E : r(e) = v\}$$

Let E_v^+, E_v^- and $E_v^n \subseteq E_v$ be the subsets of trust, distrust and neutral edges that reply to the user v. Note that $E_v = E_v^+ \cup E_v^- \cup E_v^n$ and $E_v^+ \cap E_v^- \cap E_v^n = \emptyset$.

3.3 Extracting the interaction network

We suggest searching nine types of relations using manually designed heuristic rules, checked sequentially in the following order:

Explicit quoting: *CancerDuSein.org* allows users to explicitly quote another user's post. However, only 349 posts on the Website are explicit quoting. They have been detected automatically using the HTML tag <*quote*>.

Second posts: Messages posted at the second place in each thread have been considered as replying to the first one.

Names and pseudonyms: If a message contains the pseudonym or the name of a user who previously posted a message in the same thread, then this user is considered as the recipient of the message. The following preprocessing steps were been applied to detect names and pseudonyms: 1) Remove all non-alphabetic characters except spaces; 2) Replace all accented characters by the corresponding non-accented ones; 3) Lowercasing.

Grouped posts: If a message contains a group marker ("hello everyone", "Hi girls", "Thank you all", etc.) then all the users who previously posted in the same thread are considered as recipients for this post.

Second person pronouns: In French, singular second person pronouns and plural second person pronouns are different. If a singular second person pronoun is used then the recipient is considered to be the author of the previous post.

Activator posts: If the activator[3] posts a new message in the same thread, we consider that his new message is adressed to all the users who posted after him.

Questions: If the message contains a question, then the message is addressed to all the users who previously posted in the same thread.

Answers: If there is a question posted before in the thread, the recipient is the user who posted this question.

Default: If none of the above rules before are satisfied, we consider that the recipient of the message is the activator.

3.4 Predicting trust and distrust

Once the interaction network is constructed, we need to classify each post with one of the following three classes: (1) Positive: the post expresses trust to its recipient; (2) Negative: the post expresses distrust to its recipient; (3) Neutral: otherwise.

Building lists of trust and distrust expressions: We manually created two lists of expressions that should indicate if a message expresses trust (or distrust) to its recipient. These lists have been obtained by manual annotations of a set of threads using the brat tool[4]. The annotators were asked to choose trust, distrust or neutral for each thread post and to indicate the expressions that justify their choice. These expressions have been manually validated, and then corrected, lowercased and lemmatized.

[3] The user who opened the thread by posting the first message.

[4] www.brat.nlplab.org

Handling negation: If a trust expression is under the scope of a negation term, it is considered as a distrust expression and vice versa.

Computing the frequencies and classifying the posts: All posts have been automatically lowercased, lemmatized, and corrected using the Aspell[5] spell checker. Then, each post is assigned to the majority category carried by its words.

3.5 Proposed metrics

For each user v, we define a reputation value R(v) as follows:

$$R_{n+1}(v) = \begin{cases} \dfrac{\sum_{e\in E_v^+} R_n\big(t(e)\big)}{\sum_{e\in E_v^+} R_n\big(t(e)\big) + \sum_{e\in E_v^-} R_n\big(t(e)\big)}, & if\ E_v^n \neq E_v \\ 0.5 \qquad , & Otherwise \end{cases}$$

This equation is recursive and can be computed by starting with reputations equal to 1 and iterating until it converges. The proposed reputation equation depends on both the number of trust and distrust replies a user receives and the reputations of the users who posted these replies.

We also define two complementary metrics: the neutral rate of the user $NR(v)$, and the reliability of the computed reputation value $Rel\big(R(v)\big)$.

$$NR(v) = \begin{cases} \frac{|E_v^n|}{|E_v|}, & if\ |E_v| \neq \emptyset \\ 0\ , & Otherwise \end{cases} \quad , \qquad Rel\big(R(v)\big) = \begin{cases} \frac{|E_v|}{maxR}, & if\ |E_v| < maxR \\ 1\ , & Otherwise \end{cases}$$

Where $maxR$ is a constant that represents the maximum replies that a user should receive in order to have a reliability of one in his reputation.

4 Results

4.1 Evaluating the network extraction step

Two datasets were used to test our rule based heuristic. The rules have been designed according to a development set (10 threads) and tested on other 10 unseen threads.

Prior-assessment: 15 non-expert annotators, unaware of the designed rules, annotated our two datasets. Each one annotated between 1 and 5 threads so that each thread had 3 different annotators. The goal was to find the recipient(s) of each post without knowing the results of our heuristic.

Post-assessment: Three expert annotators (the authors) annotated the links found by the heuristic in the two datasets. The goal was to validate or not the links found

[5] www.aspell.net

automatically with the possibility of adding a link which was not found by the heuristic.

Evaluation: Using these annotations, the quality of the developed heuristic was evaluated. The links obtained automatically were compared with those obtained from the annotations by considering only those that have been validated by two or more annotators (a majority vote). We compare the results of the prior-assesment and the post-assesment with two baselines. The first one considers the activator of the thread as the recipeint of all the messages posted in this thread (activator). The second baseline considers the author of the previous message as the recipient (previous).

Table 1. Precision (P), recall (R) and F1-score (F1) of baselines and our heuristic obtained on both dataset using prior and post assessments

		P	**R**	**F1**
Development set	Baseline1 (activator)	0.39	0.24	0.30
	Baseline2 (previous)	0.76	0.45	0.57
	Prior-assessment	0.70	0.68	0.69
	Post-assessment	0.80	0.84	0.82
Test set	Baseline1 (activator)	0.55	0.35	0.45
	Baseline2 (previous)	0.63	0.43	0.51
	Prior-assessment	0.81	0.83	0.82
	Post-assessment	0.83	1	0.91

Discussion: Our heuristic obtained higher F1-scores than both baselines. The results obtained using a post-assessment are better than those obtained using prior-assessment. This observation can be explained by the nature of the prior-assessment itself which gives much more freedom in choosing the links. Surprisingly, the results obtained on the test set have been better than those obtained on the development set.

4.2 Evaluating the trust prediction step

Two new datasets have been used to evaluate the automatic trust inference. Unlike the first step where both datasets had prior-assessment and post-assessment, here prior-assessment has been done only for the first dataset and post-assessment has been done only for the second one.

Prior-assessment: Three annotators annotated the trust expressed in 97 messages without knowing the results of the automatic system. The agreement between them was less than the recipient assessment but still acceptable.

Post-assessment: The same three annotators annotated the trust expressed in 102 other messages. The results of the automatic system have been displayed, and annotators can chose the same value of trust or another one.

Evaluation: The results obtained by comparing the classification made by the system with the annotations (majority vote) are presented Table 2.

Table 2. Precision, recall and F1-score of the trust inference system using prior-assessment and post-assessment

Datasets	Class	P	R	F
Prior-assessment	Trust	0.67	0.93	0.78
	Distrust	0.50	0.25	0.33
	Neutral	0.96	0.83	0.89
	Global	**0.86**	**0.84**	**0.84**
Post-assessment	Trust	0.77	0.87	0.82
	Distrust	0.23	0.60	0.33
	Neutral	0.92	0.75	0.83
	Global	**0.84**	**0.78**	**0.80**

Discussion: The results obtained for the trust class are good but the recall is higher than the precision using both assessments. Therefore, our list of trust expressions seems to be sufficient to find the majority of trust posts. Moreover, the results obtained on the neutral class are also good, but the precision is higher than the recall. Finally, the results obtained on the distrust class have been the worst but it is difficult to make conclusions regarding the small number of distrust posts.

4.3 Evaluating the proposed metric

In our experiments, the constant *maxR* has been fixed to the average number of replies received by each user. The reputations of 157 users had reliabilities greater than 0.5.

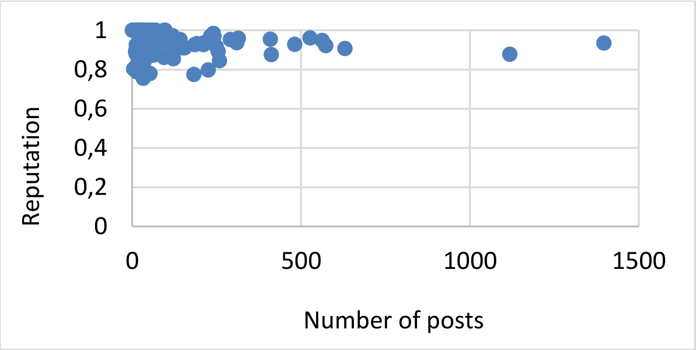

Fig. 1. User reputations that had more than 0.5 of reliability according to the number of posts

Discussion: Figure 1 shows that all considered reputations are greater than 0.7. This observation can be explained by the fact that *CancerDuSein*.org is a forum where little distrust is expressed, since the users aim at first to exchange emotional support. Moreover, the user reputations seem to be independent from the number of posted messages, which reinforces our opinion that the number of postings do not represent a good estimation of user reputations or ranks.

5 Conclusion

Many perspectives can be considered in order to improve the work and to better explore the idea. First, the user's reputation can be computed for each thread topic in addition to the global reputation in the whole forum. In fact, the user's expertise may change according to the discussed topic. Then, we are now scrolling other French forums in order to apply our method on a larger number of forums. Finally, we are planning to compare ourselves to PageRank or HITS based models built on the user interaction network.

6 Bibliography

[1] C. Lampe and P. Resnick, "Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2004, pp. 543–550.

[2] P. Sztompka, *Trust: A Sociological Theory*. Cambridge University Press, 1999.

[3] J. Golbeck, "Trust and Nuanced Profile Similarity in Online Social Networks," *ACM Trans. Web*, vol. 3, no. 4, pp. 12:1–12:33, Sep. 2009.

[4] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge Sharing and Yahoo Answers: Everyone Knows Something," in *Proceedings of the 17th International Conference on World Wide Web*, New York, NY, USA, 2008, pp. 665–674.

[5] A. Stavrianou, J. Velcin, and J.-H. Chauchat, "Definition and Measures of an Opinion Model for Mining Forums," in *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, 2009, pp. 188–193.

[6] H. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the Signatures of Social Roles in Online Discussion Groups," *The Journal of Social Structure*, vol. 8, no. 2, pp. 1–32, 2007.

[7] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise Networks in Online Communities: Structure and Algorithms," in *Proceedings of the 16th International Conference on World Wide Web*, New York, NY, USA, 2007, pp. 221–230.

[8] A. Gruzd, "Automated Discovery and Analysis of Social Networks from Threaded Discussions. Paper presented at," University of Illinois, Urbana-Champaign, 2009.

[9] M. Forestier, J. Velcin, and D. Zighed, "Extracting Social Networks Enriched by Using Text," in *Foundations of Intelligent Systems*, Poland, 2011, pp. 140–145.

[10] F. Skopik, H.-L. Truong, and S. Dustdar, "Trust and Reputation Mining in Professional Virtual Communities," in *Proceedings of the 9th International Conference on Web Engineering*, Berlin, Heidelberg, 2009, pp. 76–90.

[11] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, "Automatic Scoring of Online Discussion Posts," in *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, New York, NY, USA, 2008, pp. 19–26.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab., 66, 1999.