Definition and Measures of an Opinion Model for Mining Forums

Anna Stavrianou, Julien Velcin, Jean-Hugues Chauchat
ERIC Laboratoire - Université Lumière Lyon 2
Université de Lyon
Lyon, France
Email: anna.stavrianou@univ-lyon2.fr, julien.velcin@univ-lyon2.fr, jean-hugues.chauchat@univ-lyon2.fr

***Abstract*—Online discussion systems in the form of forums have recently been analyzed by using graphs and social network techniques. Each forum is regarded as a social network and it is modeled by a graph whose vertices represent forum participants. In this paper, we focus on the structure and the opinion content of the forum posts and we are looking at the social network that is developed from a semantics point of view. We formally define an opinion-oriented model whose purpose is to provide complementary information to the knowledge extracted by the social network model. We define and present measures that can give important information regarding the opinion flow as well as the general attitude of users and towards users throughout the whole forum. Applying our model to a real forum found on the Web shows the additional information that can be extracted.**

***Keywords*-Opinion Mining; social networks; forums; forum analysis;**

I. Introduction

The abundance and popularity of online discussion systems that usually come in the form of forums, blogs or newsgroups, has pointed out the need for analyzing and mining such systems. Monitoring how the users behave and interact with each other, their ideas and opinions on certain subjects, their preferences and general beliefs is significant.

Most existing works view an online discussion as a network in which users meet and contact each other, form communities and acquire certain roles. Forums are usually modeled by a graph whose vertices represent users that are connected with each other according to who speaks to whom. Such graphs are analyzed by social network techniques ([1]).

The application of the social network model to a forum provides information about how the users interact with each other. The opinion information contained in the forum is lost. By taking into account the structure of the posts and their opinion content, we can become more familiar with the users and get to know better their attitude during the discussion. We can observe whether there is an important opinion presence in the forum and if so, we can measure its amount. In this way, it can be easily identified when users agree with each other, in which parts of the forum they are contradicted or whether they keep talking negatively.

In this paper we present a theoretical work that has been carried out with the purpose of looking at the social network developed in a forum from the point of view of the opinions expressed by the participants. The contribution of our work is a new opinion-oriented model which is complementary to the social network model. It is supposed to be applied to an online discussion together with the social network model in order to enrich the information extracted from a discussion.

This paper is structured as follows. Section 2 discusses related work. Section 3 presents the proposed opinion-oriented framework and the components it is consisted of. In Section 4 we define metrics that allow us to measure the opinion presence and flow inside a forum. In Section 5 we show how the proposed model can be applied to a real web forum and what information we can extract from it. Section 6 concludes by highlighting future perspectives.

II. Background

Most research regarding forum analysis focuses on analyzing the interaction between users or discovering how users form communities and are affected by them. Until now we have seen no works that examine automatically how the opinion content appears, influences and flows within a network of messages. Nevertheless, our work has been influenced by existing research in the social network domain.

One of these works is presented in [2] where they analyze the Innovation Jam 2006 among IBM employees and external contributors. The representation of the discussion is seen from the point of view of posts rather than users. The difference is that they do not consider the opinion flow inside the discussion. Also, while in the IBM Innovation Jam the users are known, in our work users remain anonymous, so they express more honestly their opinion.

Forum analysis has also been dealt with in [3]. They analyze the Java Forum by using Social Network Analysis methods for the purpose of automatically identifying user expertise. They represent the social network of the forum with a graph whose vertices represent users. Their objective is different from ours since we concentrate on the content rather than the participants of a discussion and we do not seek to find experts.

In [4] they attempt to separate a set of newsgroup users in those that are for or against a topic. They represent a newsgroup as a graph with user nodes and they base their analysis on the "reply-to" links between the users. Again, they focus on the users and not on the posts. Although they consider the presence of agreement and disagreement, they do not actually take the opinion into account.

Roles are assigned to user nodes of a graph in [5] and [6]. We have been inspired by these works in the sense that each node is different and its position in the network carries information about how it affects the rest of the network. Both works, though, differ from ours both in the representation and the objective aspect.

III. OPINION-ORIENTED FRAMEWORK

In this section we present a framework which achieves a forum representation complementary to the existing techniques. The new representation allows us to exploit the structural characteristics of a forum and analyze it from a semantics-oriented point of view. The proposed framework represents a forum by an "opinion-oriented graph", whose definition follows.

Opinion-oriented graph.: An opinion-oriented graph is a graph $G = (V, E)$ with a set of V vertices and a set of E edges. Each vertex v_i represents a "message object" v_i= (m_i, u_i), where m_i is the message and u_i the user that has written it. Each edge $e_{ij} = (v_i, v_j)$ points out direction from v_i to v_j, and it is weighted by a value that represents the opinion expressed in the message object v_i as a reply to what has been said in the message object v_j. The weight is a function $w : E \to \mathbb{Z}$ and it takes negative values when the opinion is negative, the value 0 when there is no opinion and a positive value when a positive opinion is expressed.

The weight shows not only the orientation but also the strength of an opinion e.g. $w(e_{ij}) = +4$ shows a stronger positive opinion than $w(e_{ij}) = +1$.

The notion of time is encapsulated in the proposed model, so the future and the past of a vertex can be easily traced. The successor of a node v_x, which is unique in the case of opinion-oriented models, is a message object that has taken place immediately before the message object represented by the node v_x. Similarly, the predecessors of the node v_x, $\{v_y \in V \mid (v_y, v_x) \in E\}$ contain message objects that have been posted after the post represented by v_x.

An opinion-oriented graph is consisted of components whose identification allows us to define measures in order to extract useful information from such graphs. In this paper, we present two basic components; the discussion threads and the discussion chains. The distinction between a discussion thread and a discussion chain becomes apparent from Figure 1 that shows a graph consisted of 2 discussion threads. For the visualization of the graph we have used the JUNG library (http://jung.sourceforge.net).

In Figure 1, the first thread is consisted of 3 discussion chains: $\{msgObj1, msgObj3, msgObj6\}$, $\{msgObj1, msgObj4\}$, $\{msgObj1, msgObj2, msgObj5\}$. The second thread is consisted of 2 discussion chains: $\{msgObj10, msgObj11, msgObj13\}$, $\{msgObj10, msgObj12\}$.

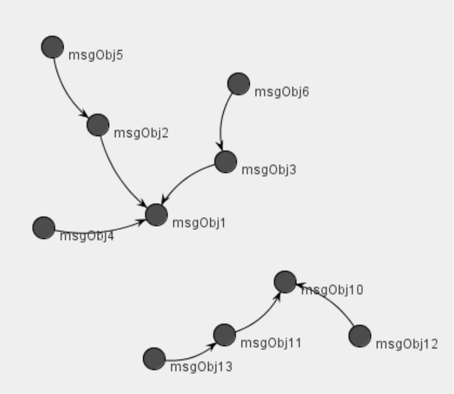

Figure 1. Discussion threads and chains

The set of the discussion threads in an opinion-oriented graph G is the union of all the maximal connected components of G. The discussion threads can be "queried" either by a message m or a user u. For example, the threads where the user u has participated can be found by tracing the vertices of each thread of the graph until a message object $v_i = (m_i, u)$ is found.

The discussion chains consist of the paths in the graph whose starting node is a root and ending node is a leaf when we inverse the direction of the edges.

In order to define a discussion chain, we consider $roots(G)$ to be the set of vertices of the graph G which represent message objects that do not reply to another message. Moreover, $inReply(v_x)$ is the indegree of the node v_x which is discussed more extensively in the next section. A formal definition of a discussion chain follows:

Discussion chain.: We define a discussion chain in the graph G as the subgraph

$$G_c = (V_c, E_c)$$

where
$V_c = \{v_i, v_{i-1}, v_{i-2}..., v_{i-x}\}$, $v_i \in roots(G)$, $inReply(v_{i-x}) = \emptyset$, $v_i \neq v_{i-x}$,
$v_{i-k} \in inReply(v_{i-(k-1)}), \forall k, k \in [1, x]$ and
$E_c = (V_c)^2 \cap E$.

Similarly to the discussion threads, the discussion chains can also be queried by a specific message or user. The discussion chains where a message m appears are all the chains G_c of the graph G for which $\{\exists v_x \in V_c \mid v_x = (m, u_x)\}$. Similarly, the chains where the user u has participated are the chains G_c of the graph G for which $\{\exists v_x \in V_c \mid v_x = (m_x, u)\}$.

The chains are important in an opinion-oriented graph. The longest discussion chain can point out the longest exchange of messages in a forum and it can be measured by the maximum number of edges that start from a leaf node and end up to a root node.

If we have more than one chain in the graph, then there is at least one node v_x that has received more than one reply. Additionally, if there exists a node $v_x \in V$ for which its reply has received another reply, then we assume that we have a generation of possible subtopics that start from v_x. Otherwise we consider to have only reactions and not subtopics starting from v_x. These arguments become clearer in Figures 2 and 3.

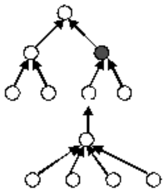

Figure 2. Message object that has possibly caused subtopics

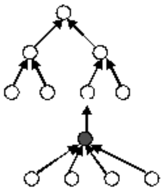

Figure 3. Message object that has only caused reactions

For example, in Figure 2 we can assume that the root of the graph has caused the generation of two different sub-discussions, one of which is initiated by the node in black. This black node, in turn, is dividing the discussion into two parts. The black node in Figure 3 has caused four reactions that have not moved the discussion forward, so we cannot assume that there are four different subtopics that have been invoked.

IV. Opinion-oriented measures

After having defined the opinion-oriented model, we now present some measures that enable us to determine the flow of the opinion inside a discussion and the opinion status of the participants.

We start by clarifying what an *opinion* is in our model.

Opinion.: We define *opinion* as $opinion(v_x, v_y) = orientation(v_x) * strength(v_x)$, where $orientation(v_x)$ takes values in $\{-1, 0, 1\}$ if the opinion expressed in the message object v_x is negative, objective (i.e. no opinion) or positive respectively, and $strength(v_x) \in \mathbb{Z}$ shows how strong this opinion is.

The $orientation$ and the $strength$ are calculated by Opinion Mining techniques ([7], [8], [9], [10], [11]). The description of such techniques are out of the scope of this paper and this is why we will consider that the *opinion* expressed by a vertex is known.

A. Basic measures

In order to define the opinion-oriented measures we use the concept of the direct predecessors which is the set of reply nodes *inReply* towards a vertex v_x. According to the theory of graphs, it is defined as:

$$inReply(v_x) = \{v_y \in V \mid (v_y, v_x) \in E\}. \tag{1}$$

The number of the predecessors shows how many reactions have been caused by the post represented by the vertex v_x. This is a measure of the popularity of a message object and it can be an indicator for a classification of the posts from the most to the least popular. Popular posts point out the "heart" of a forum and their identification facilitates the mining of a forum by directing the analysis towards the most popular messages.

Each forum participant may post many messages inside a forum. These messages are encapsulated in the message objects represented by the vertices of the graph. The message objects written by user u are given by:

$$msgs(u) = \{v_1, ..., v_x\}, v_x \in V, v_x = (m_x, u). \tag{2}$$

This concept is used later in order to define other measures.

Considering a discussion chain G_{c_i} of the graph G, we can calculate the number of edges that point out positive opinion p as:

$$replyCh(G_{c_i}, p) = \{(v_x, v_y) \in E_{c_i} \mid opinion(v_x, v_y) > 0\}. \tag{3}$$

Similarly we define the negative opinion n and the objective edges o.

B. Main opinion measures

A message object $v_x \in V$ is replied to during a discussion through posts. These posts may contain objective information or they may include the sentiments of the author expressed by positive or negative opinions. Measuring the average opinion received by a message object v_x can give us an indication of the reactions of the participants towards the specific post. If, for example, the average opinion is 0, this means that either the reply posts contained objective information, or there is a balance between positive and negative opinions.

We define the average opinion received by a message object v_x that has caused reactions as:

$$avgMsgOpinion(v_x) = \frac{\sum opinion(inReply(v_x), v_x)}{\mid inReply(v_x) \mid}. \tag{4}$$

The $\mid inReply(v_x) \mid$ points out the popularity of the node v_x and it actually shows how many replies the post represented by the vertex v_x has received.

A discussion chain G_c in the opinion-oriented graph G connects a series of replies between messages. We consider that it represents a sub-dialogue or even a sub-topic inside a forum. Defining opinion measures for a discussion chain could give an idea of the sentiment flow inside the particular sub-dialogue/topic. Moreover, by observing the opinion during the time, we could observe the evolution of the opinion in this chain.

A user may own more than one post inside a discussion chain. By capturing the average opinion expressed by a user u inside a discussion chain G_c, we identify the average opinion reaction of the specific user within a specific sub-dialogue or sub-topic. We define this concept by the following measure:

$$avgFromUsrChain(G_c, u) = \frac{\sum opinion(v_x, v_y)}{\mid msgs(u) \cap V_c \mid} \quad (5)$$

where $v_x \in msgs(u) \cap V_c$.

In the same way, we can define the average opinion expressed towards a user inside a chain as:

$$avgToUsrChain(G_c, u) = \frac{\sum opinion(inReply(v_x), v_x)}{\sum \mid inReply(v_x) \mid}, \quad (6)$$

where $v_x \in msgs(u) \cap V_c, inReply(v_x) \in V_c$.

This measure describes on average the opinion expressed in the reactions towards the posts of the specific user, within a sub-dialogue or a sub-topic. The pre-requisite is that at least one of the posts of the user has been replied to.

Another measure that characterizes a discussion chain is the opinion information measured by the entropy $H(Gc_i)$. This measure facilitates the identification of the discussion chains that contain the maximum amount of opinion information. The measure is defined as:

$$H(Gc_i) = -\sum_{k=n,o,p} \left(\frac{replyCh(G_{c_i}, k)}{\mid E_{c_i} \mid} \log \frac{replyCh(G_{c_i}, k)}{\mid E_{c_i} \mid}\right) \quad (7)$$

where n, o and p point out the negative, the objective and the positive edges respectively.

The opinion information is an indication of the variety of opinions inside a discussion chain. Similarly, we can use the entropy $H(v_x)$ to define the amount of opinion information held by a node $v_x \in V$.

The opinion of a user can also be seen globally for the whole forum. In this way, we can observe users that keep a negative or positive position throughout the discussion or we can identify tendencies such as whether people tend to write more when they are unhappy or when they are satisfied with a certain situation. We define two measures:

- the average opinion expressed by a user u during the forum:

$$avgFromUsr(u) = \frac{\sum opinion(v_x, v_y)}{\mid msgs(u) \mid} \quad (8)$$

where $v_x \in msgs(u), (v_x, v_y) \in E$ and

- the average opinion expressed towards a user u (having received at least one answer) during the forum:

$$avgToUsr(u) = \frac{\sum opinion(inReply(v_x), v_x)}{\sum \mid inReply(v_x) \mid} \quad (9)$$

where $v_x \in msgs(u)$.

Both of these measures give results that may differ from the results given by the respective per chain measures.

V. Application

In this section we show how the proposed model can be applied to a real forum and what information the measures we have defined provide us with.

We have taken a forum from the site of a French newspaper (http://www.liberation.fr) that consists of 121 messages and 97 users. We have automatically created the opinion-oriented graph that is shown in Figure 4. The message objects appear with an identification number calculated internally by our application developed for the purpose of visualizing and analyzing forums. The opinion polarities are omitted for legibility reasons.

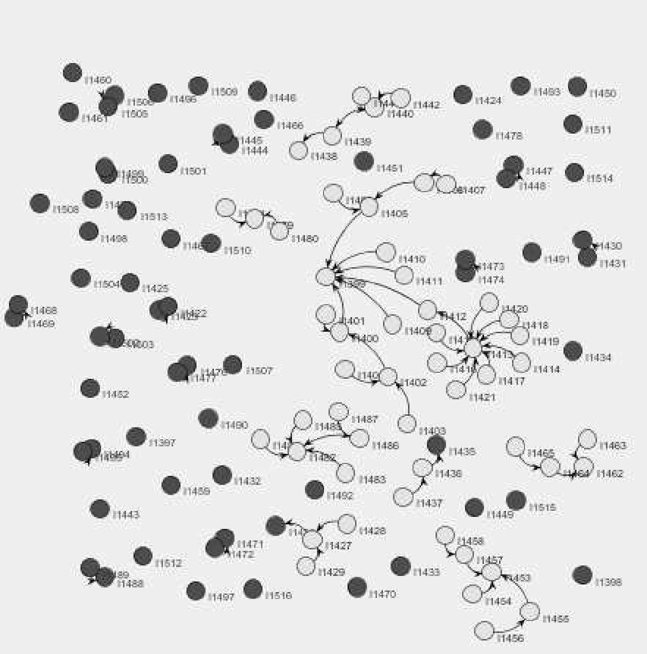

Figure 4. Opinion-oriented graph of a forum

As we can see in Figure 4, the opinion-oriented graph consists of some nodes that do not connect to the rest of the graph. These nodes represent message objects that do not reply to any other message and they have not received any reply either. Visualizing the forum structure by an opinion-oriented graph allows us to concentrate on the discussion threads that consist of many nodes or many discussion chains. Such threads appear in the center of Figure 4 and their nodes are colored in a light gray. We can also identify quicker and focus on the threads that contain posts that have received reactions with varied opinion polarities.

Let us assume that we want the following information:

- the most popular message m_p and which user u_p has written it
- what is the average opinion towards the message m_p
- how varied are the opinions expressed in the replies towards the m_p
- if the author of the m_p has written other messages and what on average is his/her general opinion status
- the average attitude of the rest of the authors towards the user u_p during the forum
- the discussion thread and the discussion chain that contain the message m_p
- whether the particular discussion thread contains another popular message or not
- what is the message that has led to the most popular message m_p
- whether there are subtopics in the discussion thread that contain the m_p

Apparently this information cannot be given by the social network model of the forum, but it can be extracted from the opinion-oriented model.

First of all, the most popular messages are easily identified. In Table I, we present them in descending order by the number of reactions they have received. We refer to them by the unique code given by our application and we provide also information regarding the average opinion of their replies and their variety given by the entropy.

Table I
MOST POPULAR MESSAGES OF THE FORUM

v_x	$\vert inReply(v_x)\vert$	$avgMsgOpinion(v_x)$	$H(v_x)$
I1413	8	-0.375	0.39
I1399	6	-0.5	0.3
I1482	4	0	0
I1453	3	0	0.48

From Table I, we can see that the message I453 has the highest entropy of all. Indeed this is the message that has received replies with the highest variety of opinions. The message I1482 has 0 entropy which shows lack of opinion variety in the replies it has received. In combination with the value 0 of the $avgMsgOpinion$ we understand that all the replies are objective so they contain no opinion (otherwise the entropy would not be 0).

The author u_i of each message m_i is known since the message and its author are both encapsulated in the concept "message object". By definition, each message object v_i is a relation (m_i, u_i). The fact that we know the author of the most popular messages, allows us to look at the position and role of the particular authors in the social network model and derive some conclusions about them, such as how much they can influence the rest of the discussion, the communities in which they exist etc.

In Table II we show the results of the opinion measures oriented towards the authors of the most popular messages. Instead of giving the actual pseudo of each user, we name them "A", "B", "C", "D". From this table, we notice that the user A has a negative attitude during the discussion, while the user D has a balanced attitude (sometimes positive, sometimes negative). The measure is not defined for the users B and C because they have written messages that do not reply to any other message. Furthermore, there was an average negative reaction towards the users A, B and D.

Table II
MEASURES APPLIED TO THE USERS

Msg Object	User	$avgFromUsr(u)$	$avgToUsr(u)$
I1413	A	-1	-0.44
I1399	B	-	-0.5
I1482	C	-	0
I1453	D	0	-0.33

Regarding the discussion threads and chains, they are easily identified by the opinion-oriented model. We notice that both popular messages I1399 and I1413 belong to the same thread. This encourages us to give priority in analyzing the particular thread among all the rest, since its content should be more interesting having caused a discussion around it. Additionally, from our model we can distinguish the message that has led to the most popular message, just by following the respective edge that relates the specific message with its successor (predecessor in time). Finally the particular thread is divided into two subtopics.

In conclusion, the application of the opinion-oriented model to a forum results in the extraction of knowledge that cannot be provided by the social network model. This shows the worth of the proposed model in the domain of discussion analysis.

VI. CONCLUSION AND FUTURE WORK

This paper presents a theoretical work that consists in defining formally an opinion-oriented model. The novelty of our proposal is that we integrate into the model the opinion content of the exchanged forum posts, information that is lost when we represent a forum by a social network model. We define measures that give information regarding the opinion flow and the general attitude of users and towards users throughout the whole forum. The application of the proposed model to real forums shows the additional information that can be extracted and the interest in combining the social network and the opinion-oriented models.

We believe that the future in opinion-oriented graphs is prosperous. Future work will pass from the theoretical to an experimental state by performing large-scale experiments with real forums.

One future objective is to add the time dimension in our model. This will permit monitoring how opinion changes over time. In this way, we could observe whether a product improves as the time passes, whether people become more satisfied with certain services, or even whether people are finally convinced after a long discussion in a forum.

Furthermore, an interesting future issue is to combine the social network and the opinion-oriented models for an improved discussion analysis. For example, we could extract the experts [3] of the discussion domain through the social network representation, and we could, then, use this information in order to extract from the opinion-oriented model their attitude or the discussion chains in which they have participated.

The information extracted by the opinion-oriented model can be used in many ways. We have experimented by using it in order to rank forum messages from the most to the least interesting. This is a combination of many criteria such as how many reactions a message causes, whether it receives reactions that contain opinions, whether these opinions have the same strength or not. Initial results are promising but more extensive experiments are needed.

REFERENCES

[1] P. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, 2005.

[2] M. Helander, R. Lawrence, and Y. Liu, “Looking for great ideas: Analyzing the innovation jam,” in *KDD*, 2007.

[3] J. Zhang, M. Ackerman, and L. Adamic, “Expertise networks in online communities: Structure and algorithms,” in *Proc. of the 16th international conference on WWW*, 2007, pp. 221–230.

[4] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, “Mining newsgroups using networks arising from social behavior,” in *Proc. of the 12th international conference on WWW*, 2003.

[5] D. Fisher, M. Smith, and H. Welser, “You are who you talk to: Detecting roles in usenet newsgroups,” in *Proc. of the 39th Annual HICSS*. IEEE Computer Society, 2006.

[6] J. Scripps, P.-N. Tan, and A.-H. Esfahanian, “Node roles and community structure in networks,” in *WebKDD/SNA-KDD ’07*. ACM, 2007, pp. 26–35.

[7] X. Ding and B. Liu, “The utility of linguistic rules in opinion mining,” in *SIGIR-07*, 2007.

[8] A. Ghose, P. Ipeirotis, and A. Sundararajan, “Opinion mining using econometrics: A case study on reputation systems,” in *ACL*, 2007.

[9] V. Hatzivassiloglou and K. Mckeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 8th conference on European chapter of ACL*, 1997, pp. 174–181.

[10] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *KDD*. ACM, 2004, pp. 168–177.

[11] P. Turney and M. Littman, “Measuring praise and criticism: inference of semantic orientation from association,” *ACM TOIS*, vol. 21, no. 4, pp. 315–346, 2003.