A combination of opinion mining and social network techniques for discussion analysis

Anna Stavrianou *, Julien Velcin **, Jean-Hugues Chauchat ***

ERIC Laboratoire - Université Lumière Lyon 2
Université de Lyon
5 avenue Pierre Mendès-France 69676 Bron Cedex, France
* anna.stavrianou@univ-lyon2.fr
** julien.velcin@univ-lyon2.fr
*** jean-hugues.chauchat@univ-lyon2.fr

Abstract. Mining opinion data that reside in online discussions is a way to track opinions of people on specific subjects. Many of the existing techniques model a discussion as a social network of users and they represent it with a user-based graph. In this paper we propose a new framework for discussion analysis. We combine Social Network Analysis and Opinion Mining in order to give structure to a discussion. Such techniques have not been combined until now. We propose the use of an opinion-based graph whose vertices contain message objects and its «reply-to» edges are labeled with opinion polarities. We compare the opinion-based with the user-based graphs and we analyze the different information that can be extracted from them. Our experiments validate the proposed framework and show that the representation of discussions by opinion-based graphs gives information that cannot be provided by a user-based graph.

1 Introduction

The development of Web2.0 has resulted in the generation of a vast amount of blog repositories, review sites, web forums and online discussions. In this type of discussions people express opinions, criticize products and ideas, exchange knowledge and beliefs. Tracking opinions on specific subjects allows the identification of user expectations and needs, feelings of people about certain political decisions or reactions against particular events. As a result, mining and extracting opinion data that reside in online discussions becomes significant.

Opinion Mining is the field that deals with the mining of subjective statements from texts, the identification of opinions, the estimation of opinion orientation and the extraction of arguments that relate to opinions. Mining opinions in online discussions requires an appropriate representation.

An online discussion can be represented as a graph where the vertices are knowledge entities (users, messages etc.) and the edges between them show relationships. Hence, a discussion can be analyzed by techniques of the Social Network Analysis which is the mapping of relationships between people, organizations or other information/knowledge processing entities

(Helander et al. (2007)). Currently most online discussions are modeled by a social network in the form of user-based graphs where the vertices represent the users who participate in the discussion.

In this article we propose a new framework for discussion analysis by combining Social Network and Opinion Mining techniques. To the best of our knowledge, techniques from these two fields have not yet been combined. Our objective is to study the structure of an online debate that takes part in a well-defined domain and analyze the user reactions, preferences, and opinions on a certain subject.

The contributions of our work are summarized in the following:

1. We propose a framework for analyzing online discussions: the structure of the discussion is seen from the point of view of exchanged messages rather than of the users who participated in the discussion. Currently most online discussions are modeled by a social network of users and not messages.
2. A combination of Social Network Analysis and Opinion Mining techniques: application of Opinion Mining knowledge to the link structure between the messages and generation of an opinion-based graph. Social Network techniques and Opinion Mining analysis have not yet been applied together to the analysis of online discussions.

The proposed framework allows the direct identification of the sentiment flow in a discussion as well as the mining of the discussion parts that contain opinions. It enables the acquisition of a content-oriented view of the discussion and the focus on the parts that have caused reactions by the participants. It gives an indication of the variety of opinions received through reply posts by a message and it monitors the opinion behavior of a user and towards a user during the discussion.

The objective of the proposed model is not to replace the social network represented by user-based graphs, but to provide additional, complementary information. It could be used together with the user-based graphs in order to enrich and better handle the knowledge extracted from a discussion.

This article is structured as follows. Section 2 discusses existing research in both Social Network and Opinion Mining fields. Section 3 presents the proposed opinion-based framework and it defines some opinion measures. In Section 4 we validate the proposed framework through some experiments and examples with discussions found in the web. Section 5 concludes by highlighting future perspectives.

2 Existing research

In this article we perform discussion analysis by combining techniques from the Opinion Mining and the Social Network field. These two domains are not complementary and this is why we will present the existing research in both fields separately. The focus is more on the Opinion Mining field.

2.1 Current Opinion Mining techniques

The identification of the opinion polarities and strength inside a text is a fairly recent area of study with plenty of applications and challenges.

Hatzivassiloglou and Mckeown (1997) are among the first to deal with opinion classification. They focus on adjectives and they study phrases where adjectives are connected with conjunction words such as «and» or «but». They construct a log-linear regression model so as to clarify whether two adjectives have the same orientation. The accuracy of this task is declared to be 82%. Their technique is described in the following steps:

- extract from a text the conjoined adjectives that are connected with the words «and»/ «but» etc.,
- run a supervised algorithm that builds a graph where the nodes are the adjectives and the links determine same or opposite orientation,
- run a clustering algorithm to separate the graph into two classes,
- assume that the cluster with the highest frequency is the one that shows positive orientation.

Another important work in the field is that of Turney and Littman (2003) who use a pointwise mutual information (PMI) and a latent semantic analysis (LSA) measure to find out the statistical relation between a specific word and a set of positive or negative words. They construct a seed set which contains words that can be classified as either positive or negative independently of the context e.g. «excellent» is always positive. The LSA-based measure gives better results than the PMI-based one.

Their approach can be described in the following steps:

- part-of-speech tagging in order to identify adjectives and adverbs,
- extraction of 2-word phrases where one word is an adjective or adverb,
- calculation of the association between two words w1 and w2, where w1 is a word in the review and w2 is a word from the seed set. The association is calculated by the statistical measures LSA or PMI, based on the co-occurrence of two words,
- calculation of the sum of the LSA or PMI measurement between a word and the words from the positive seed set. Subtraction of this sum from the sum of the association between the same word and the words in the negative seed set,
- classification of the review as positive or negative according to the average semantic orientation of the review phrases.

Wiebe (2000) deals with the distinction between objective and subjective sentences i.e. between facts and opinions. They deal with 3 subjectivity types: positive, negative and speculation. They follow the process:

- construction of a seed set by manually tagging the subjective adjectives of a corpus and determine a strength score for each of them from 1-3,
- populate the seed set as follows: for each subjective adjective of strength 3, find 20 synonyms or near-synonyms by using a distributional similarity measure Lin (1998) or WordNet,
- add semantic features to adjectives. The features are the semantic orientation and the gradability (whether a word can modify a noun or it can be used in comparative sentences).

Their results show that the probability of a sentence being subjective, given that there is at least one adjective in the sentence is 55.8%. Also, the sentences that contain an adjective that exists in the expanded seed set and the list of automatically identified positive polarity adjectives are subjective by 71%. They claim that ontologies and dictionaries are not sufficient to help distinguishing between facts and opinions because they are not tagged with subjectivity.

Constructing a seed set with the right adjectives is not a straightforward task. Harb et al. (2008) present a work in which they generate automatically a dictionary of adjectives. Initially they collect their data by getting web documents that contain negative and positive opinions. In order to determine the orientation of the opinions they use the seed set of Turney (2002). Then, they expand the initial seed set by extracting more adjectives from the collected documents. The extracted adjectives have to be related to more than one adjective found in the initial set. Finally, each document is classified as positive or negative according to the number of positive or negative adjectives it contains. The experiments show that following this approach the seed set is expanded with relevant adjectives that help in the correct classification of documents according to the opinion polarities.

A significant work in the field of Opinion Mining is that of Hu and Liu (2004). They deal with product reviews written by customers on web sites. Their objective is to produce a structured summary that informs about positive or negative statements that are made for product features. Their process is the following and the model they use is presented in Liu (2007):

- find out product features that are discussed in the reviews (e.g. camera size, camera image etc.). This is done by selecting the frequent words, assuming that people often use the same words to describe features. Label sequential rules and patterns are used,
- identify opinion sentences and their orientation. An opinion sentence is defined as a sentence that contains both a feature and one or more adjectives. They use a seed list of 30 basic adjectives. For each adjective in the reviews, they check whether it exists in the seed list or it is an antonym or synonym of a word in the seed list. Every time the orientation of an adjective is found, the seed list is expanded with this adjective,
- infrequent features are identified by looking for the nearest noun phrases to an opinion word,
- summarization of results. Each sentence is given the orientation of the majority of its part-orientations.

Ding and Liu (2007) improve the previously mentioned (Hu and Liu (2004)) system by assigning an orientation score to each opinion word found in a sentence. The score takes into account the semantic orientation of the opinion word that is located near the feature-word and the distance between the feature and the opinion word. In this way a low score is given to the opinion words that are far from the feature.

The majority of the mentioned approaches focus on adjectives and adverbs. They use a seed list and they attempt to find out the relation between the words that appear in a text and the words of the seed list. The difference lies in the similarity measure used to calculate the association between words. Some use WordNet, others use statistical measures. Some approaches give also importance to the percentage of how positive or negative a word is.

An original way of calculating the polarity and strength of opinions that differs from the techniques mentioned, is proposed in Ghose et al. (2007). They use the feedback comments posted by users in a reputation system such as «eBay» and they calculate the effect of these comments on the prices of the products sold. The orientation and the strength included in the opinion of a user's feedback are inferred by observing the changes in the respective product prices. If, for example, a certain opinion results in the reduction of a product's price, then this opinion is considered to be negative and its strength is measured on the basis of how much the price has been reduced.

2.2 Current Social Network techniques

The Social Network Analysis deals with the analysis of the relationships that exist between entities in a social network. For instance, in a social network of people, the analysis can include who is friend with whom, who can influence which group of people, whom can have access to the information that goes through this network etc.

Lately there has been a growing interest in this field, especially as to how it gets involved with knowledge discovery and data/web mining. For instance, analyzing the behavior of users in online discussions or discover how users form communities and are affected by them are interesting works.

Fisher et al. (2006) analyze newsgroups by applying Social Network techniques and they interpret online communities by assigning roles to the members of the groups. This is done by observing how people relate to each other in a graph-based model of post-reply relations. They notice that short discussion threads point out question-answer exchanges and longer threads indicate proper discussions.

Java et al. (2007) analyze the Twitter's social network and the intentions of the associated users in order to understand the reason why people use such networks. They identify the communities that are formed, they categorize them into communities that create information, communities that receive information and communities that exist only because of friendship. They label the identified communities by the keywords that appear in the various posts.

Scripps et al. (2007) introduce a new measure that defines the number of communities to which a node is attached. Using this measure they assign roles to nodes by considering the community structure in the network of the node. Defining roles in this way, improves the performance of link-based classification and influence maximization tasks.

The rest of this subsection presents existing work in the field of discussion analysis from which our research has been influenced.

One of these works is that of Helander et al. (2007) that analyzes 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 from our work is that our objective is not to find out the degree of innovation of a discussion but to identify opinions. Moreover, in our case, the participants of the discussion come from different backgrounds as opposed to the Innovation Jam - so they have different concepts and beliefs. Also, while in the IBM Innovation Jam the users are known since they are specific IBM employees, in our work users remain anonymous. The anonymity allows people to express more honestly how they feel about a certain issue.

In Maurel et al. (2008) they have analyzed forums in the domain of tourism and they have extracted information regarding user sentiments and tourist destinations. They apply syntactic and semantic processing techniques and they adapt the grammar rules or the opinion words they try to identify according to the domain. They do not, though, represent the discussion as a graph.

Forum analysis has also been dealt with in Zhang et al. (2007). 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 user-based graph. 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.

A work with the objective of separating a set of newsgroup users in those that are for or against a topic is presented in Agrawal et al. (2003). In this work they represent a newsgroup as a user-based graph and they base their analysis on the «reply-to» links between the users. They do not consider the content of each text because they claim that the statistical methods do not work for small messages where users use similar vocabulary. Contrary to this, we consider the content of the text and its semantic attributes in order to identify the sentiment orientation of the text.

3 Opinion-based model

In this section we present the opinion-based model we propose for the representation of the structure of web discussions. We consider that the participants of the discussions identify themselves by a user name and they can participate by either writing a new message or replying to an already posted one. The relations «reply-to» between the exchanged messages point out which message replies to what and they are considered to be known.

The model is based on a graph-based representation. Most graph-based existing works consider users to be the vertices of the graph. In this article, we propose to use message objects as the vertices.

Hence, we represent the online discussions by a directed graph $G = (V, E)$, where V is the set of vertices which denote message objects and E are the edges that show the relations «reply-to». The message objects encapsulate both the content of the message and the author who has written it. In this way information about the author is not lost. The edges are labeled by the opinion polarities included in the reply message. We call this graph *opinion-based graph* and we formally define it as follows:

Definition. We define an **opinion-based graph** to be the graph $G = (V, E)$ with n vertices and m edges. The set of vertices $V = \{v_1, v_2, ..., v_n\}$ contains vertices of the type v_i= (m_i, u_i). v_i is a *message object* which is composed of m_i that represents the message itself and u_i that is the user who has written it. The set of edges is $E = \{e_{1i}, e_{2i}, ..., e_{mi}\}$, and each edge $e_{ij} = (v_i, v_j)$ points out direction from v_i to v_j.
Each edge e_{ij} is weighted by a value that represents the opinion expressed in the message object v_i which is a reply to the message object v_j. The weight is a function $w : E \rightarrow \{-1, 0, 1\}$ and it takes the value -1 when the opinion is negative, the value 0 when there is no opinion and $+1$ when a positive opinion is expressed.

In Figure 1 we can see an example of an opinion-based graph that is composed of three discussion threads. The discussion threads are the connected components of the graph that represents the discussion. The image is taken from our platform that is developed for the purpose of visualizing and analyzing discussions. We have used the JUNG library (http://jung.sourceforge.net) for the implementation of this platform.

3.1 Opinion measures

In this article we concentrate on the opinion measures that are derived from the proposed model. These measures enable us to determine the flow of the opinion inside a discussion. We

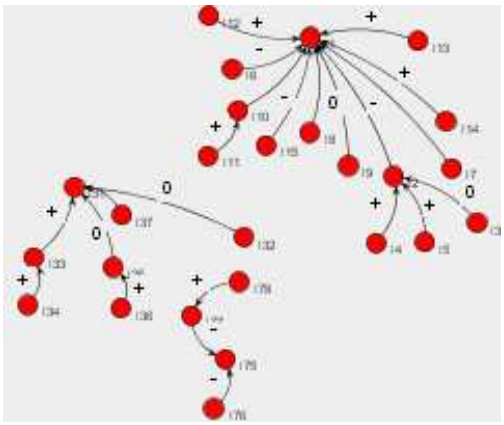

FIG. 1 – *An opinion-based graph composed of three discussion threads.*

have separated the measures into three categories according to whether they characterize the opinion per node, per discussion chain or per discussion as a whole. In order to define the opinion measures that follow, we will first clarify what an *opinion* is in our model.

The $opinion(v_x, v_y)$ is denoted by the weight of an edge $e_x = (v_x, v_y)$ and it expresses the opinion polarity which is present in node v_x which is a reply to the node v_y. It 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 it can be calculated by Opinion Mining techniques, such as the ones mentioned in the previous section.

In the measures that follow we use the concept of **predecessors**. This consists of the set of reply nodes *inReply* towards a vertex v_x, and according to the theory of graphs, it is defined as:

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

It's worth noting that in opinion-based graphs, these nodes do not describe the predecessors in time since the replies to a node v_x take place after (and not before) the generation of the particular node.

Another concept we use in the definition of the opinion measures is the **message objects** that are generated by a certain user. The message objects written by user u are given by :

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

3.1.1 Opinion measures per vertex

A message object $v_x \in V$ may be replied to during the discussion through posts. These posts may contain objective information or they may include the sentiments of the author expressed by positive or negative opinions.

We define the *average opinion* received by a message object v_x with predecessors as:

$$avgMsgOpinion(v_x) = \frac{\sum opinion(inReply(v_x), v_x)}{\mid inReply(v_x) \mid} \quad (1)$$

The average opinion towards a message object is 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 can always distinguish between the different posts according to their opinion polarity. The number of positive posts towards a message object v_x is defined by the number of reply vertices that are connected to the message object by an edge with positive weight. We describe the number of positive, negative and objective replies as:

$$reply(v_x, r) = \mid \{v_y \in inReply(v_x), opinion(v_y, v_x) = r\} \mid, r \in \{-1, 0, 1\} \quad (2)$$

Having described the various nodes according to the opinion polarities included in their reply posts, allows us to define a measure regarding the *opinion information* held by a node. We use the entropy H for this purpose, and we define the amount of opinion information held by a node $v_x \in V$ (that has been replied to), as:

$$H(v_x) = -\sum_{i=-1,0,1} (\frac{reply(v_x, i)}{\mid inReply(v_x) \mid} \log \frac{reply(v_x, i)}{\mid inReply(v_x) \mid}) \quad (3)$$

The opinion information is an indication of the variety of opinions received by a node. If, for instance, a node has received reply posts that are all of the same opinion orientation, then the entropy will be 0. This information can be interpreted as: there is either objective information or unanimous opinion regarding the message expressed by the particular node.

3.1.2 Opinion measures per chain

A *discussion chain* $G_c = (V_c, E_c)$ in the graph G is a path whose starting node is a root and ending node is a leaf when we inverse the direction of the edges. The distinction between a discussion chain and a discussion thread becomes apparent from Figure 2 that shows a graph consisted of 2 threads marked by a rectangle. In this Figure, the «THREAD 1» is consisted of 3 discussion chains: {msgObj1, msgObj2, msgObj3}, {msgObj1, msgObj4}, and {msgObj1, msgObj5, msgObj6, msgObj7}. The «THREAD 2» is consisted of 2 discussion chains: {msgObj8, msgObj9} and {msgObj8, msgObj10, msgObj11}.

A discussion chain connects a series of replies between messages and we consider that it represents a sub-dialogue or even a sub-topic inside the discussion. As a result, defining opinion measures that characterize each 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.

The number of positive edges inside a discussion chain are the edges that belong to E_c having a positive weight. This is described as: $\{(v_x, v_y) \in E_{c_i} \mid opinion(v_x, v_y) = 1\}$. The number of negative and objective edges is described in the same way.

A user may have more than one posts inside a discussion chain by replying, for example, to messages (s)he has already received. In order to capture the *average opinion expressed by a user* u inside a discussion chain G_c, we define the following measure:

$$avgFromUsrChainOpinion(G_c, u) = \frac{\sum opinion(v_x, v_y)}{\mid msgs(u) \cap V_c \mid}, v_x \in msgs(u) \cap V_c \quad (4)$$

This measure denotes on average the opinion reaction *of* the specific user within a sub-dialogue or a sub-topic.

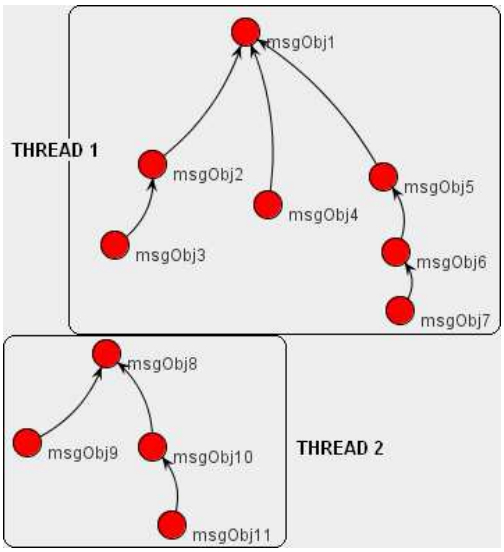

FIG. 2 – *Distinction between discussion threads and discussion chains.*

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

$$avgToUsrChainOpinion(G_c, u) = \frac{\sum opinion(inReply(v_x), v_x)}{\sum \mid inReply(v_x) \mid}, \tag{5}$$

$$\text{where } v_x \in msgs(u) \cap V_c, inReply(v_x) \in V_c$$

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.

Similarly to the *opinion information* measure per vertex $H(v_x)$ we defined previously, we can define the same measure per discussion chain. This enables the identification of the discussion chains that contain the maximum amount of information.

3.1.3 Opinion measures for the discussion

The opinion of a user can also be seen globally for the whole of the discussion. 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. Considering that the discussion is represented by the graph $G = (V, E)$, we define two measures:

the average opinion expressed *by* a user u during the discussion:

$$avgFromUsrOpinion(u) = \frac{\sum opinion(v_x, v_y)}{\mid msgs(u) \mid}, v_x \in msgs(u), (v_x, v_y) \in E \tag{6}$$

and the average opinion expressed *towards* a user u during the discussion:

$$avgToUsrOpinion(u) = \frac{\sum opinion(inReply(v_x), v_x)}{\sum \mid inReply(v_x) \mid}, v_x \in msgs(u) \tag{7}$$

Both of these measures refer to the behavior of a certain user when looking at the discussion as a whole and the results they give may differ from the results given by the respective per chain measures.

3.2 Discussion

The combination of Social Network Analysis methods and Opinion Mining techniques improves the discussion analysis and it allows the extraction of certain information from an online discussion that is not straightforward when we represent it as a social network of users. Examples of this information include:

- the discussion chains that show which users are talking about the same subtopics,
- the posts that have caused many reactions,
- the opinion/sentiment flow of the discussion,
- the sentiment behavior of a user during the discussion,
- the sentiment behavior towards a user during the discussion.

In an opinion-based graph, message objects that appear in the same discussion chain imply similarity in content. In a user-based graph that represents a social network, a «reply-to» relationship does not always mean similarity in topic since two users may have replied to each other many times in many different discussion chains.

The opinion-based graph identifies the messages that have caused many reactions, and allows the concentration of the analysis on the content rather than the participants of the discussion. A message that has received more replies compared to one that has received none is definitely a more interesting message that may worth being analyzed in more detail.

From the opinion identification point of view, the opinion-based graph enables the direct and immediate visualization of the opinion polarities during the discussion. It facilitates the identification of the discussion parts that contain opinions, it enables the distinction between the objective and subjective sides of the discussion and it allows focusing on the parts that show more interest from an opinion exchange point of view. This permits the mining of the information and the focus of the analysis on a subset of the discussion rather than on the discussion as a whole. In addition, the new model, through the proposed measures, permits measuring the average opinion during the discussion, the average opinion per sub-dialogue, as well as, the average opinion received per message.

The proposed model allows us to observe the evolution of the opinion inside a discussion. For instance, if in the beginning of the discussion we have a variety of opinion polarities and in the end the majority of polarities are negative, we can assume that the «atmosphere» of the discussion has turned to really bad. Observing the evolution of an opinion will be, of course, more efficient when it is done in collaboration with Text Mining techniques (Stavrianou et al. (2007)) i.e. the extraction of keyword information from the various posts.

User-based and opinion-based graphs serve different purposes. Both of them give structure to a discussion and they aid the discussion analysis by extracting useful information from the structured representation. Using a combination of these graphs for the analysis of a discussion should be considered.

Main differences between the proposed and the existing representation are summarized in Table 1.

	User-based model	Opinion-based model
Entity	The main entity of the discussion is considered to be the *user* who participates.	The main entity is the *message object* that is posted.
Chains	-	Identify the posts that are connected in discussion chains.
Community	A community of user nodes denotes users that respond to each other (friendship, interest in the same topic).	A community of opinion-based nodes shows content relations and possible similarity in the subtopic references.
Interaction	We can observe how the users interact with each other.	We can observe how the message objects form discussion chains.
Opinion	-	We can see the opinion flow of the discussion and measure the opinion information per post, chain and discussion as a whole.
Popularity	If many edges arrive to a user node, then the specific user is popular because s/he has received messages by many people.	If many edges arrive to an opinion-based node, then the specific post is popular because it has caused many reactions.
Visualization	Users and «reply-to» relations.	Message objects, «reply-to» relations and opinion polarities.

TAB. 1 – *Differences between user-based and opinion-based models.*

4 Model evaluation

In this section, we evaluate the proposed model. The evaluation is done by showing the advantages and the complementary information that can be extracted from an opinion-based graph as compared to the standard user-based graph of the social network model.

For this purpose we applied our model to real discussions found on the site of the Digital Camera Magazine Community (http://community.dcmag.co.uk). In these discussions one can find opinions regarding the use of digital cameras, photography contests and photography in general.

Our approach consists of three steps. For each discussion, we identify the structure and we generate automatically a graph. Then we find out the type of opinions that appear in the messages of the discussion, and finally we label the graph with the opinion polarities. The focus of this article is on the presentation of an opinion-based model that facilitates the discussion analysis and not on the ways to identify opinion data. As a result, we do not describe the approach we followed in detail, since we consider it out of the scope of this article.

In Table 2 we give some information of the discussions we analyzed such as the number of messages exchanged and the number of users that participated. The column «Opinion Edges» shows the number of edges that represent positive or negative opinion so, in other words, the reply posts that include sentiments. The column «Opinion Chains» shows the number of discussion chains which are characterized by a sequence of edges labeled with a positive or negative opinion that are not interrupted by a neutral-opinion edge.

The experiments with real discussions allowed us to observe the characteristics of online discussions and identify the behavior of users. The experiments confirmed the importance of opinion-based graphs since they capture information that cannot be provided by the user-based graphs.

In more detail, from Table 2, we see that in the majority of the discussions, the number of opinion edges is less than the number of messages. This shows that inside the discussions we can find many messages that do not express opinions. These are usually messages that contain a question, a request for an advice or a statement to start a discussion with on a certain topic. This points out the importance of our proposed model since it allows a discussion analyst to concentrate only on the parts of the discussion that contain opinions without losing time in analyzing the whole discussion. The opinion-based graph allows us to see at a glance how the opinion flows inside the discussion, how the positive messages alternate with the negative ones. This useful information is not provided by the user-based model.

A user-based graph allows us to identify who is talking to whom or who could be considered as an expert in a discussion (Zhang et al. (2007)). An opinion-based graph, though, allows us to see how the discussion evolves and extract the chains of the discussion threads. By following the different discussion chains we may find out sub-dialogues or sub-topics, and, as a result, being able to see the reactions of users and the evolution of the opinion per chain. In our experiments we identified the «opinion chains» which are characterized by a series of messages that express opinions.

Having represented the debate from the point of view of message objects instead of users allows us to identify quicker interesting opinion discussion chains. A message that has received few but varied positive and negative opinions can be more interesting than one that has received plenty of messages that are all positive or neutral. In our model, we measure this by the entropy. This measure allows the selection of the right data to focus on, concentrating on

Disc.	Messages	Users	Disc. Chains	Opinion Edges	Opinion Chains
1	24	18	4	16	2
2	19	12	8	9	5
3	19	8	7	4	0
4	18	6	3	8	2
5	16	11	4	5	3
6	16	6	4	7	1
7	15	7	4	7	1
8	13	4	1	3	0
9	12	11	3	8	3
10	12	6	3	5	1
11	11	8	3	3	1
12	11	8	3	4	1
13	11	7	1	3	1
14	11	4	1	5	1
15	10	9	3	3	0
16	10	3	2	3	0
17	9	8	4	3	1
18	9	4	3	4	1
19	8	6	2	4	1
20	7	5	3	5	1

TAB. 2 – *Information about the analyzed discussions.*

discussion chains where some opinions exist and ignoring the neutral statements. If we had a user-based graph where the edges were weighted by opinion information, we would know that at some point there has been an interesting exchange of opinions. What we would not know is the discussion chain in which the exchange of opinions has taken place since two users may exchange many neutral messages before they start expressing opinions.

The opinion-based graph allows us to observe the sequences of opinions. A sequence of two positive edges in a discussion may show agreement between the users. The prerequisite, though, for assuming agreement is that the messages express a positive opinion on the same argument. Keyword and feature extraction with the combination of opinion information will allow the identification of agreement or disagreement between the users. For example, the message « I do not agree with the best photograph chosen for the competition » that receives as a reply the message « Yeah, its subject was usual and boring » is a sequence of two negative opinions that shows agreement between them. On the contrary, the message « It is bad to say this » as a reply to the message « This camera is bad » does not point out agreement even though both messages express a negative opinion and they are connected. Similarly, a sequence of a positive message followed by a negative one does not necessarily show disagreement between the messages, since it could also be a discussion on different aspects of the same topic.

From the experiments we noticed that in short discussions where few users participate, everyone is exchanging messages with everyone else. Users belong to the same community and often this community is a clique. As a result, if we represent the debate from the point

of view of users, finding cliques and communities between users would not make sense, but finding communities between messages has a sense in order to identify discussion chains. Discovering and analyzing communities and roles (relation of a node to its neighbors) in online discussions is a recent area of research (Du et al. (2007), Fisher et al. (2006), Scripps et al. (2007), Zhou et al. (2007)). In a debate represented by an opinion-based graph, knowing the roles of messages and the opinions expressed in them could help the user find out quicker how he can participate in the debate or to whom to talk to in the first place, instead of losing time reading all the messages. Link prediction and community identification algorithms are a future research issue in opinion-based graphs.

In conclusion, the opinion-based graphs provide information that cannot be extracted by the user-based graphs and as such, the proposed model is useful and it has a lot to offer to the discussion analysis.

4.1 Analysis of a short discussion

In this section, we will apply our model to a short artificial discussion in order to see how the measures are used. We consider the participants of this discussion to be 4 (A, B, C, D) and the exchanged messages 13. The message flow is shown in Table 3.

Message Object	Message	Author	Reply-to	Opinion
1	message 1	A	-	no
2	message 2	B	message 1	+
3	message 3	C	message 1	+
4	message 4	D	message 1	+
5	message 5	A	message 2	no
6	message 6	B	message 3	+
7	message 7	A	message 3	+
8	message 8	C	message 7	-
9	message 9	D	message 7	-
10	message 10	D	message 3	-
11	message 11	A	message 10	-
12	message 12	C	message 10	-
13	message 13	B	message 10	+

TAB. 3 – *Message flow of the short discussion.*

In Table 3, the column «Author» shows the author of the specific message, the column «Reply-to» shows which message the reply refers to and the column «Opinion» denotes whether there is a positive (+), negative (-), or neutral (no) opinion in the content of the message.

The representation of this short discussion by an opinion-based graph is depicted in Figure 3. For comparison purposes the user-based graph is shown in Figure 4.

As we can see, the two graphs in Figure 3 and Figure 4 represent different information for the same discussion. The user-based graph shows the interaction between the discussion participants. We do not know who initiated the discussion or the order in which the users spoke to each other. Additionally we cannot identify the parts of the discussion during which the users have participated; did they speak only in the beginning or they were active participants

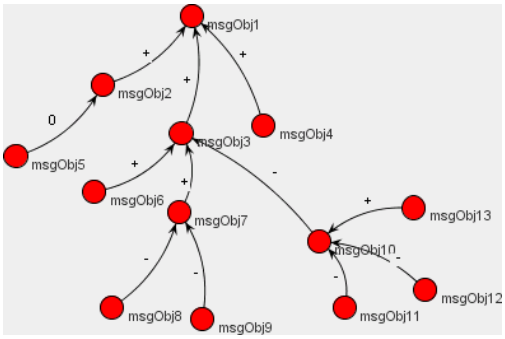

FIG. 3 – *Opinion-based graph of the short discussion.*

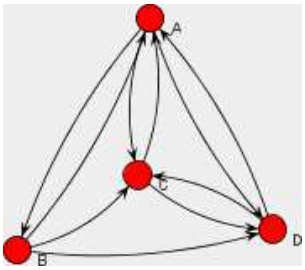

FIG. 4 – *User-based graph of the short discussion.*

throughout the whole discussion? For instance, in the user-based graph of Figure 4, we can see that the users A and B exchanged some messages but it is only in the opinion-based graph that we can identify during which part of the discussion they actually chatted.

In the opinion-based graph the messages are not just a bundle of random messages but they have a structure. They have ancestors and descendants. We can distinguish the discussion chains: {msgObj1, msgObj2, msgObj5}, {msgObj1, msgObj3, msgObj6}, {msgObj1, msgObj3, msgObj7, msgObj8}, {msgObj1, msgObj3, msgObj7, msgObj9}, {msgObj1, msgObj3, msgObj10, msgObj11}, {msgObj1, msgObj3, msgObj10, msgObj12}, {msgObj1, msgObj3, msgObj10, msgObj13}, {msgObj1, msgObj4}. All the chains have as root the initial post. As a result, we know that it is the author A who started the discussion. We can see in which parts of the discussion this author appears and we notice that the first discussion chain is just a short dialogue between the authors A and B. Similarly we can follow the presence of all authors in the discussion.

Additionally we can identify the parts where opinion messages appear. We see, for example, that all discussion chains contain some opinion information. More specifically, the node representing the message object 10 has received replies expressing both negative and positive opinions, and the node of the message object 1 has only received reactions containing positive opinions.

The graph allows us also to notice the most popular messages, which in our example are represented by the message objects 3 and 10. Both of these messages have caused reactions

and they have received replies that contain opinions.

In Table 4 we give the values of some measures per node. From this table, we identify which messages have caused reactions with positive or negative opinion polarities. The reactions of the message object 10, for instance, are on average negative. Moreover, the average opinion values 1 (msgObj1) and -1 (msgObj7) show unanimous positive and negative opinion received respectively.

We can also see that the message objects that have received varied opinion replies have higher entropy than the rest of the message objects. In other words, the message objects 3 and 10 are regarded as the nodes of the graph that hold higher opinion information since they have caused varied opinion reactions.

The combination of the average message opinion value and the entropy reveals more information. For example, by knowing these two values for the popular message objects 3 and 10, we can assume that the msgObj3 has received varied opinion reactions that are mostly positive, while the msgObj10 has had various reactions mostly negative.

Message Object	$avgMsgOpinion(v_x)$	$reply(v_x, 1)$	$reply(v_x, -1)$	$reply(v_x, 0)$	$H(v_x)$
1	1	3	0	0	0
2	0	0	0	1	0
3	$\frac{1}{3}$	2	1	0	0.2772
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	-1	0	2	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	$-\frac{1}{3}$	1	2	0	0.2772
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	0	0

TAB. 4 – *Opinion measures applied to the short discussion.*

In Table 5 we show the results of the opinion measures oriented towards the users. From this table, we notice that the user B has always had a positive reaction during the discussion. Also, the users C and D had a more negative than positive reaction. Furthermore, there was an average positive reaction towards the user A and C and an average negative reaction towards the user D. This is indeed the case in our example.

User	$avgFromUsrOpinion(u)$	$avgToUsrOpinion(u)$
A	0	0.2
B	1	0
C	-0.33	0.33
D	-0.33	-0.33

TAB. 5 – *Opinion measures applied to the users.*

4.2 Analysis of a long web discussion

Applying our model to bigger discussions with hundreds of messages is interesting. We have taken a discussion from the site of a French newspaper (http://www.liberation.fr). The discussion is in French and it consists of 272 messages and 205 users. We have manually identified the opinion polarities and we have automatically created the opinion-based graph that is shown in Figure 5. The message objects appear with an identification number calculated internally by our application. The opinion polarities are omitted for legibility reasons.

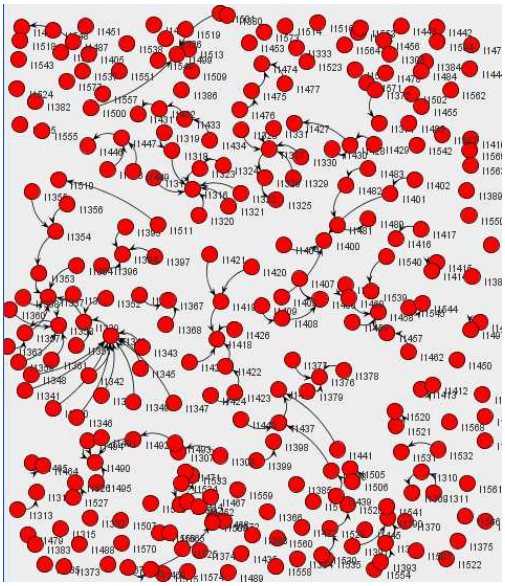

FIG. 5 – *Opinion-based graph of the discussion of the site of « liberation ».*

As we can see in Figure 5, the opinion-based graph is quite complex. There are many nodes that are «lonely» in the sense that they 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. We can also discern some discussion threads that are consisted only of two nodes. The visualization of the opinion-based graph allows us to concentrate on the discussion threads that consist of many nodes or many discussion chains or reactions with varied opinion polarities. Such chains appear in the center and on the left center-side of Figure 5.

Since, we cannot show the analysis for the discussion per node and per chain, we give in Table 6 some global statistics in order to point out the complementary information we can extract from our graph as compared to the graph of the social network. From this table, we see the number of discussion threads and chains that appear in the discussion. This information cannot be extracted by a user-based graph. We can also identify how many of these threads/chains contain opinions. Furthermore, we notice that the negative opinion edges are more than the positive ones (35 negative as opposed to 8 positive edges) which shows that, in this discussion, users have the tendency to speak negatively rather than positively. This agrees with the obser-

vation of Agrawal et al. (2003) who say that, in general, users post more messages when they disagree rather than when they agree.

Messages	272
Users	205
Disc. Threads	40
Disc. Chains	94
Opinion Threads	16
Opinion Chains	9
Opinion Edges	43 (35-, 8+)

TAB. 6 – *Statistics of the discussion of « liberation ».*

In Table 7, we apply some measures on the most popular messages of the specific discussion. We refer to them by their unique code given by our application.

Message	No. of reactions	$avgMsgOpinion$	H
I1340	10 (5-, 1+, 4∅)	-0.5	0.41
I1316	6 (2-, 1+, 3∅)	-0.17	0.44
I1337	5 (3-, 0+, 2∅)	-0.6	0.29
I1358	4 (1-, 0+, 3∅)	-0.25	0.244
I1418	4 (2-, 0+, 2∅)	-0.5	0.3

TAB. 7 – *Information about the most popular messages.*

From Table 7, we see that the message I1340 is the most popular one, having had 10 reactions of which 5 were negative, 1 was positive and 4 contained no opinion. We notice that the message I1316 has the highest entropy of all. Indeed this is the message that has received replies with the highest variety of opinions. We also notice that the average opinion of all messages is negative which indicates the general tendency of the discussion. Again, this information is not given by the user-based model.

The information provided through the opinion-based graph facilitates the mining of the discussion by reducing the dimension space of the data. For instance, graph nodes that do not connect to the rest of the graph are less probable to have an impact on the whole discussion or to contain interesting opinions. As a result, such nodes can be ignored by the user who wants to get quickly an idea of the most interesting messages of the discussion. The reduction of the space can be achieved through a user-based model only from the point of view of users but not from the point of view of posts. This means that from a user-based model we could see the users that influence the discussion but not the exact messages. In other words, the opinion-based model can be seen as a « zooming » process into the user-based one.

Mining the discussion graph and transforming it from a complex to a simpler one by extracting only the nodes that seem to contain important information is significant for such long discussions.

5 Conclusion and future perspectives

In this paper, we have proposed a new framework that represents online discussions by opinion-based graphs. This enables a content-oriented representation of a discussion focusing on its sentiment flow. Such a graph allows a straightforward identification of discussion parts where opinions rather than facts are involved. The proposed opinion measures offer a sentiment-oriented analysis of the online discussion.

The future in opinion-based graphs is prosperous. More measures need to be defined and more large-scale experiments are needed for the formal validation of our model.

One future objective is to combine the user-based and the opinion-based graphs in order to analyze a discussion. For example, we could use the user-based graphs in order to extract the users that are experts Zhang et al. (2007) in the discussion domain. Afterwards, by using this information we could extract from the opinion-graph, the discussion chains where the experts have participated.

An interesting issue is also to monitor how opinion changes over time. This allows observing 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.

Identifying agreement and disagreement is another perspective that cannot be determined by the orientation of a text (Stavrianou and Chauchat (2008)). In the future we are planning to carry out experiments in order to find out if our model facilitates this identification.

References

Agrawal, R., S. Rajagopalan, R. Srikant, and Y. Xu (2003). Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International Conference on World Wide Web*.

Ding, X. and B. Liu (2007). The utility of linguistic rules in opinion mining. In *SIGIR-07*.

Du, N., B. Wu, X. Pei, B. Wang, and L. Xu (2007). Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Analysis*, pp. 16–25. ACM.

Fisher, D., M. Smith, and H. Welser (2006). You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual HICSS*. IEEE Computer Society.

Ghose, A., P. Ipeirotis, and A. Sundararajan (2007). Opinion mining using econometrics: A case study on reputation systems. In *ACL*.

Harb, A., G. Dray, M. Plantié, P. Poncelet, M. Roche, and F. Trousset (2008). Détection d'opinion : apprenons les bons adjectifs ! In *Atelier Fouille des Données d'Opinions (FODOP 08)*, pp. 59–66.

Hatzivassiloglou, V. and K. Mckeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174–181.

Helander, M., R. Lawrence, and Y. Liu (2007). Looking for great ideas: Analyzing the innovation jam. In *KDD '07: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. ACM.

Java, A., X. Song, T. Finin, and B. Tseng (2007). Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Analysis*, pp. 56–66.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING-ACL*.

Liu, B. (2007). *Web Data Mining ? Exploring Hyperlinks, Contents and Usage Data*. Springer.

Maurel, S., P. Curtoni, and L. Dini (2008). L'analyse des sentiments dans les forums. In *Atelier Fouille des Données d'Opinions (FODOP 08)*.

Scripps, J., P.-N. Tan, and A.-H. Esfahanian (2007). Node roles and community structure in networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Analysis*, pp. 26–35. ACM.

Stavrianou, A., P. Andritsos, and N. Nicoloyannis (2007). Overview and semantic issues of text mining. *SIGMOD Record 36*(3), 23–34.

Stavrianou, A. and J.-H. Chauchat (2008). Opinion mining issues and agreement identification in forum texts. In *Atelier Fouille des Données d'Opinions (FODOP 08)*, pp. 51–58.

Turney, P. (2002). Thumbs up or down? semantic orientation applied to unsupervised classification of reviews. In *ACL-2002*, pp. 417–424.

Turney, P. and M. Littman (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS 21*(4), 315–346.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *AAAI-2000*.

Zhang, J., M. Ackerman, and L. Adamic (2007). Expertise networks in online communities: Structure and algorithms. In *Proc. of the 16th International Conference on World Wide Web*, pp. 221–230.

Zhou, D., I. Councill, H. Zha, and C.-L. Giles (2007). Discovering temporal communities from social network documents. In *International Conference on Data Mining (ICDM '07)*, pp. 745–750. IEEE Computer Society.

Résumé

La plupart des recherches existantes représentent les discussions en ligne par un réseau social des participants sous forme de graphes. Dans cet article, nous utilisons une combinaison des techniques de la fouille des données d'opinions et des réseaux sociaux afin d'analyser des débats en ligne. La représentation proposée est orientée par le contenu et la dynamique de la discussion. Elle facilite l'analyse de discussions et l'identification des parties les plus significatives. Elle permet aussi la visualisation des polarités des opinions et l'évolution des opinions par sous-thème.