EXPLOITING EMOTIONS IN SOCIAL INTERACTIONS TO DETECT ONLINE SOCIAL COMMUNITIES

Xin Li, Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, xin.li@cityu.edu.hk
Hsinchun Chen, Department of Management Information Systems, University of Arizona, Tucson, Arizona, USA, hchen@eller.arizona.edu
Su Li, Center for the Study of Law and Society, School of Law, University of California - Berkeley, Berkeley, CA, USA, su.li@wichita.edu

Abstract

The rapid development of Web 2.0 allows people to be involved in online interactions more easily than before and facilitates the formation of virtual communities. Online communities exert influence on their members’ online and offline behaviors. Therefore, they are of increasing interest to researchers and business managers. Most virtual community studies consider subjects in the same Web application belong to one community. This boundary-defining method neglects subtle opinion differences among participants with similar interests. It is necessary to unveil the community structure of online participants to overcome this limitation. Previous community detection studies usually account for the structural factor of social networks to build their models. Based on the affect theory of social exchange, this research argues that emotions involved in social interactions should be considered in the community detection process. We propose a framework to extract social interactions and interaction emotions from user-generated contents and a GN-H co-training algorithm to utilize the two types of information in community detection. We show the benefit of including emotion information in community detection using simulated data. We also conduct a case study on a real-world Web forum dataset to exemplify the utility of the framework in identifying communities to support further analysis.

Keywords: Community detection, social network analysis, Web 2.0, user-generated contents
1 INTRODUCTION

The virtual world is more “real” to people nowadays than ever before. Web 2.0 technologies allow people to experience virtual interactions easily through Web forums, Web blogs, social networking Websites, video-sharing Websites, etc. Through these interactions, people gradually form social groups/communities based on their shared values and interests, which are quite different from traditional communities formed based on geographical locations (Wellman, 2005). Virtual communities influence their members’ decisions in online and offline activities, such as in online auctions (Chua et al., 2007) and e-marketing events (Dellarocas, 2003). Identifying and understanding virtual communities is critical to researchers and business managers.

Identifying online communities is not a trivial task, as it is usually difficult to define virtual community boundaries. Previous studies on virtual communities mostly use natural boundaries such as Websites or Web forum boards to define communities. However, within these boundaries, multiple social groups may exist that care about similar topics but hold different opinions. Ignoring the subtle differences of interests and values across these groups may lead to biased conclusions.

Several previous studies have examined the community detection problem in the context of social networks. However, most of these studies deemed social relationships as binary relations (i.e., indicating the existence of relations) and only accounted for the structural factor of social networks in community detection. They usually ignore a substantial amount of information associated with social interactions, such as the (favorable/unfavorable or agreement/disagreement) emotions involved in the interactions. Nevertheless, social scientists have highly valued the role of emotions in the formation of communities. According to the affect theory of social exchange (Lawler, 2001), emotions involved in social exchange impact actors’ collective behavior. We conjecture that the effect of emotion could be more significant for virtual communities, since people tend to express stronger emotions in online interactions than in face-to-face interactions (Sia et al., 2002).

In this research, we try to include the factor of emotions into social network analysis for community detection. We propose a framework that applies opinion-mining techniques to extract social interactions emotions from user-generated contents and design a GN-H co-training algorithm that accounts for both social network structure and emotions in social interactions for community detection. We conduct experiments on simulated data to show the benefit of including emotion information in community detection and a case study on a WalMart-related Web forum dataset to exemplify the utility of the proposed framework in identifying virtual communities.

2 RELATED WORK

2.1 Emotion in Social Networks

The social network model abstracts social actors as nodes and social relationships as links in a graph. The social network theory enables us to account for the relationships among actors together with actors’ attributes as determinants of social activities (Marsden & Friedkin, 1993). In social network analysis, the structural characteristic of social networks was considered a major factor in explaining social phenomena (Wasserman & Faust, 1994).

However, the structure of networks ignores information related to the links such as the emotions involved in the interactions. According to the social exchange theory (Hochschild, 1979), emotions are an integral part of the normative context of exchange. Lawler proposed an affect theory of social exchange (Lawler, 2001), in which he argued that social units (relations, groups, networks) are a source of emotions and emotions involved in the exchange impact actors’ collective behavior.

In spite of these initial theoretical explorations, few empirical attempts have been made to combine social networks with emotions in studying social interactions. This may be due to the seemingly
orthogonal nature between the interactionists’ approach and the social network structural approach in studying social interactions (Gibson, 2005). As a default setting, most social networks in previous studies were based on positive connections. Few of them have considered negative connections (Sampson, 1968; White, 1961). It may due to the difficulty to collect negative social relationships and the fact that people are likely to break the connections if too much negative affects are involved.

The development of computer mediated communication (CMC) technologies and Web 2.0 provides us an opportunity to study emotions in social networks. In online interactions people tend to express stronger emotions than in face-to-face interactions (Sia et al., 2002). The communications (and debates) between actors who hold different opinions may exist for a long time. These channels provide us access to emotion information of social interactions and call for empirical studies that combine social networks with emotions in understanding human online behaviors.

2.2 Community Detection in Social Networks

There are several approaches one can use to extract social groups from online participants (i.e., model the community structure of participants). For example, we can use individual actor’ attributes and apply traditional clustering algorithms, such as hierarchical clustering (Girvan & Newman, 2002) and K-means (Hopcroft et al., 2004), to group similar individuals. However, since communities are formed on the basis of social interactions, it makes more sense to take a social network analysis perspective to detect communities.

Focusing on the graph representation of social networks, a variety of analytic tools in graph theory can be used to inspect the community structure in social networks. For example, fully connected nodes (i.e., cliques) or the chains of connected cliques were considered as indicators of communities (Palla et al., 2005). Social networks can be decomposed to isolated dense components (i.e., communities) by removing possible inter-community links identified using measures such as link betweenness (Girvan & Newman, 2002). Individuals can also be categorized by their positional closeness to each other in a social network (Alves, 2007; Zhou, 2003).

Social networks can be transformed to other representations for building community detection models. According to the spectral graph theory, the eigenvectors of representative matrices of social networks, such as the Laplacian matrix (Donetti & Munoz, 2004) and modularity matrix (Newman, 2006), can reflect the networks’ community structure. One can also build probabilistic models on inter- and intra-community social interactions for given networks and extract communities based on the most possible parameter settings of the model (Airoldi et al., 2008; Newman & Leicht, 2007). In addition, the community detection problem can be formulated as a combinational optimization problem to optimize community assignment quality, in which several search algorithms, such as greedy search (Clauset et al., 2004) and simulated annealing (Guimera & Amaral, 2005), can be applied to find the best community assignment.

Similar to other social network analysis studies, previous community detection studies usually neglected the important emotion component in social interactions. However, social scientists have noticed the importance of emotions in the formation of a community (McMillan & Chavis, 1986). The affect theory of social exchange explains how interaction emotions help interpret group participants’ collective behaviors (Lawler, 2001). Interaction emotions modeled in a signed social network (Wasserman & Faust, 1994), on which a small number of community detection studies have been conducted. For example, Traag et al. (Traag & Bruggeman, 2009) and Doreian et al. (Doreian & Mrvar, 2009) proposed to take a combinational optimization approach to reduce intra community negative links. In general, in order to better utilize interaction emotions to interpret human behaviors in virtual communities, it is worthwhile to examine emotion-aware community detection models.

2.3 Opinion Mining

Online social communities usually involve a large amount of (textual) user-generated contents. To quantify emotions in these contents effectively for community studies, it is necessary to employ
opinion-mining techniques. Opinion mining seeks to determine the level of subjectivity and polarity in textual expressions (Pang & Lee, 2008). Opinion mining can be conducted based on affect lexicons (Mishne, 2006) or patterns learned from tagged text using machine-learning techniques (Dave et al., 2003). Learning-based approaches may better detect the subtle and indirect expressions of sentiment (Pang et al., 2002) than lexical approaches since that are based on the actual data in applications.

In addition to quantifying individual snippets’ subjectivity and polarity, determining the agreement between pairs of documents is another opinion-mining problem (Baym, 1996). If the document pairs focus on identical topic, agreement detection can be done by comparing the sentiment on the topic. However, if the documents are on a variety of issues, more sophisticated models are needed. Galley used lexical features with a maximum entropy model to determine the agreement relations in a sequence of conversations (Galley et al., 2004). Sorower et al. proposed to use syntactic features together with word features to improve classification accuracy (Sorower & Yeasin, 2007). Hahn et al. (Hahn et al., 2006) took an semi-supervised learning approach for agreement detection.

3 METHODOLOGY

Noticing the important role of emotions in the formation of virtual communities and the limited community detection studies that considered interaction emotions, we are interested in studying how interaction emotions would affect community detection. We employ opinion-mining techniques to extract posting-level interaction emotions and aggregate them to a user-level emotion social network (in which links are annotated with favorable/unfavorable emotions). We then design emotion-aware community detection algorithms to identify the community structure of such networks. Our proposed framework (Figure 1) contains four steps: data acquisition, interaction and emotion extraction, community detection, and community analysis.

![Figure 1. A framework for emotion-aware community detection from user-generated contents](image)

3.1 Data Acquisition

At the data collection stage, we collect user-generated contents from Web 2.0 Websites, such as Web forums, Web blogs, and product review Websites, using Web crawlers or through Web APIs provided by the Websites. The meta-data and discussion contents of the extracted data are parsed into relational databases for future analysis.

3.2 Interaction and Emotion Extraction

Since the overall relationships (and emotions) between people are built based on individual instances of communication, we extract posting-level social interactions and emotions first. In some Web 2.0 applications, user discussions focus on specific objects/topics, in which the objects may serve as proxies to reveal posting-level social interactions. For example, in online product review applications, a review can be considered as a response to previous reviews on the same item. In applications with
open discussion, such as in Web forums, detecting the post-reply relationships between users is nontrivial. In our framework, we use a heuristic-based text mining approach to identify social interactions (Fu et al., 2008). This approach follows a procedure of matching quotation information (including quoted user names and contents) in previous postings, finding the direct address of other users in the current posting, and calculating the lexical similarity between the current posting and previous postings to identify possible post-reply relations. The relationships that cannot be identified in these three steps will be considered replying the above post as a rule of thumb.

Table 1. Features for interaction emotion detection

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>Word, N-gram, Word count</td>
<td></td>
<td>First 3 sentences’ subjective word count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>First 3 sentences’ positive word percentage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>First 3 sentences’ subjectivity</td>
</tr>
<tr>
<td>Syntactic</td>
<td>POS tag, Function word, Punctuation char, Content word count</td>
<td></td>
<td>First 3 sentences’ total sentiment score</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>First 3 sentences’ polarity</td>
</tr>
<tr>
<td>Affective</td>
<td>Word sentiment value, First 3 word’s sentiment value, First 3 sentences’ positive word count, First 3 sentences’ negative word count</td>
<td>Affective (cont’)</td>
<td>Posting positive word count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting negative word count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting subjective word count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting positive word percentage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting subjectivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting total sentiment score</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posting polarity</td>
</tr>
</tbody>
</table>

We then take a machine learning approach to reveal the agreement/disagreement or favorable/unfavorable emotions within each identified post-reply pair based on textual contents. In Web forums, the discussions in a thread may change gradually. However, in most cases, the posts still have agreement/disagreement emotions even if the topic is drifted from the original one or new materials are provided. (This is consistent with what we found in the dataset used in this research). In this research, we only care about the relations between two consecutive posts, and we consider this task as a two-class classification problem. We use the contents in posting pairs as the major evidence for classification. Based on previous studies, we compile a list features for this text classification task, including lexical features, syntactic features, and sentiment features (Table 1). To extract the sentiment features, we used SentiWordNet as a lexicon resource (Esuli & Sebastiani, 2006). We apply information gain-based feature selection and use the Support Vector Machine (SVM) algorithm to build classifier on the selected features. We expect the classifier learn from human-annotated instances to estimate the relations between users based on wordings in the post-reply pairs.

3.3 Community Detection

We aggregate the extracted posting-level interactions and emotions into a user-level (actor level) emotion social network. We average the emotion values of all the interactions between a pair of users to annotate the overall emotion between them, where agreement/disagreement and favorable/unfavorable emotions are represented as positive/negative links. In this process, one may want to set up a threshold on the total number of interactions between a pair of users to remove the unimportant interactions or unimportant users. To reduce the complexity of the problem, we do not differentiate emotion strength and only consider emotion polarity in this research.

To extract the embedded social groups from the emotion social networks with both positive and negative links, we propose a GN-H co-training algorithm that combines the GN algorithm (Girvan & Newman, 2002) and hierarchical clustering (Johnson, 1967). As compared with previous efforts on community detection on signed networks (Doreian & Mrvar, 2009; Yang et al., 2007), this approach makes use of both positive and negative links and does not require predefining the number of communities in the network. The algorithm is also easier to interpret for community analysis purposes.

The original GN algorithm proposed by Girvan and Newman can be applied on networks with only
positive links. It gradually removes the high betweenness links from the network to split the network into isolated components. In a real social network, intra-community links tend to hold more communication traffic and have a higher betweenness measure. Removing these links could provide communities with dense intra-community interactions. During the link removal process, the algorithm needs to determine the best community assignment to output. Thus, Newman and Girvan proposed a modularity measure (Newman & Girvan, 2004) to measure the quality of extracted communities. The modularity measure is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(i,j),$$

where $m$ is the total number of links in the network, $A_{ij}$ indicates the existence of the link between node $i$ and $j$ (0 or 1), $k_i$ and $k_j$ are the degrees of node $i$ and $j$, and $\delta(i,j)=1$ when $i$ and $j$ are in the same community.

### 3.3.1 A GN-H Co-training Algorithm

Although the GN algorithm is effective for community detection on positive networks, it cannot be applied on networks with negative links, such as emotion social networks. Thus, we split an emotion social network into a positive part and a negative part, use the GN algorithm to handle the positive part, and hierarchical clustering to handle the negative part. For hierarchical clustering, we define a similarity measure based on the distribution of negative links to other nodes (or communities). Thus, the users with similar “enemies” have a higher probability of being assigned to the same group.

---

**Input emotion social network $G$ with $k$ nodes $n_1, n_2, \ldots, n_k$.**

1) Split $G$ into a positive part $G^+$ with $l^+$ links and a negative part $G^-$ with $l^-$ links.

2) Apply the GN algorithm on $G^+$ to generate $p$ communities $C^+ = \{c^+_1, c^+_2, \ldots, c^+_p\}$.

3) Calculate two nodes’ similarity in $G$ as the inner product of their feature vectors: $S(n_i, n_j) = \langle V_i, V_j \rangle$. Feature vector $V_i$ represents node $n_i$ based on its negative links to the nodes in the $p$ communities of $C^+$: $V_{ic} = V_{ic} = \sum_{w \in c} \text{abs}(w_{ic})/|c^+_s|$, where $\text{abs}(w_{ic})$ is the absolute weight of the link between node $n_i$ and $n_c$, and $|c^+_s|$ is the number of nodes in $c^+_s$.

4) Apply hierarchical clustering based on $S(n_i, n_j)$ to generate $q$ communities $C = \{c^-_1, c^-_2, \ldots, c^-_q\}$.

5) Select $l^-$ links that have the largest $S(n_i, n_j)$ and do not belong to $G^+$. Combine the $l^-$ links and $G^+$ together as the new input of GN algorithm.

6) Generate a new community assignment $C^+$ by applying the GN algorithm on the basis of $C$: i.e., weight the shortest paths between the nodes in the same community of $C$ as a half of other paths when calculating link betweenness in the GN algorithm.

7) Go to step 3 and iterate multiple times. Keep the community assignment $C^+$ with the highest modularity measure $Q^*$ as the entire algorithm’s output.

---

**Figure 2.** Pseudo code of the GN-H algorithm

In the GN-H algorithm, the communities detected from the positive part are incorporated into hierarchical clustering and the communities detected from the negative part are incorporated into the GN algorithm. In hierarchical clustering, the detected communities (from the positive part) are used to calculate user similarities. In the GN algorithm, the links between nodes in a detected community (from the negative part) are weighted smaller when calculating the betweenness measure. Thus, those links have a smaller probability to be removed as an inter-community link. In the iterations between the two modules, information in both parts is utilized for community detection. In order to better deal with networks with a large portion of negative links, which is rare, we generate hidden positive links between nodes with similar (negative) link distributions. These hidden positive links are used together with the original positive links in the algorithm. After multiple iterations of the two modules, the best community assignment is selected as output. (Please see details in Figure 2.)

Similar to the GN algorithm, the GN-H algorithm needs to examine the quality of community
assignments. We propose a modified modularity measure, which is similar to the one in (Traag & Bruggeman, 2009). The measure is defined as:

$$Q^*=\frac{1}{2m^++2m^-}\left[\sum\left(A^+_{ij} - \frac{k^+_i k^+_j}{2m^+}\right)\delta(i,j) - \sum\left(A^-_{ij} - \frac{k^-_i k^-_j}{2m^-}\right)\delta(i,j)\right],$$

where \(m^+\) and \(m^-\) are the total number of links of the positive and negative parts, \(A^+_{ij}\) and \(A^-_{ij}\) indicate the existence of the link between node \(i\) and \(j\) in the positive and negative parts, \(k^+_i\) and \(k^+_j\) are the degrees of node \(i\) and \(j\) in the positive part, \(k^-_i\) and \(k^-_j\) are the degrees of node \(i\) and \(j\) in the negative part, and \(\delta(i,j)=1\) when \(i\) and \(j\) are in the same community. This definition applies the modularity measure on the positive and negative part separately and then combine them together. It prefers community assignments with more intra-community positive links (less inter-community positive links) and less intra-community negative links (more inter-community negative links).

### 3.4 Community Analysis

In order to extend our understanding of the effect of interaction emotions in community detection, we conduct experiments to compare the performances of using different portions of emotion information. First, we can use all available emotion information with social networks for community detection. We can ignore the emotion information and use only the social network built from the online interactions for community detection. We can also create a network for using only partial emotion information by removing links with negative emotions from the network. The network for partial emotion imitates a simple practical approach of dealing with emotion in social networks, i.e., only examine the positive links using existing community detection algorithms to unveil the community structure.

After experiments on simulated data with known community structure to evaluate the effectiveness of the proposed algorithm and examine the effect of using interaction emotions in community detection, we conduct a case study on a Web forum dataset to exemplify the utility of the proposed framework.

### 4 SIMULATION

#### 4.1 Data Generation

We adopted a widely used methodology (Guimera & Amaral, 2005; Radicchi et al., 2004), in which networks with 128 nodes evenly distributed in 4 clusters with an average degree of 16 are generated for evaluation. The links were generated randomly, where we controlled the proportion of positive links (\(\alpha\)), the proportion of intra-community negative links among all negative links (\(\beta\)), and the proportion of intra-community positive links among all positive links (\(\gamma\)). We assumed that there should be more intra-community positive links than negative links. Thus \(\beta\) should be less than 0.25 and \(\gamma\) should be larger than 0.25 (for a network with 4 clusters). We selected the three parameters as: \(\alpha\) from 0.1 to 0.9 (step 0.1), \(\beta\) from 0 to 0.2 (step 0.05), and \(\gamma\) from 0.5 to 0.9 (step 0.1) to imitate more realistic networks. For each parameter setting, we randomly generated 100 networks as testbed.

#### 4.2 Evaluation

To examine the effect of using interaction emotions in community detection, we generated a network with all emotions, a network with partial emotion, and a network with no emotion as explained in Section 3.4. We applied the GN-H algorithm on the first network and the GN algorithm on the other two networks to detect communities from them. The community detection results from these two algorithms were compared with the actual clusters of the dataset for evaluation.

We adopted a widely used normalized mutual information measure (Danon et al., 2005) to evaluate the community detection performances. After constructing an \(n*m\) confusion matrix \(N\), where \(n\) is the number of real communities, \(m\) is the number of generated communities, and \(N_{ij}\) is the number of nodes in the real community \(i\) that has been assigned to community \(j\) by the algorithm, the normalized
mutual information is measured as:

\[ NMI = \left[ -2 \sum_i \sum_j N_{ij} \log \left( \frac{N_{ij} N}{N_i N_j} \right) \right] \left/ \left[ \sum_i N_i \log \left( \frac{N_i}{N} \right) + \sum_j N_j \log \left( \frac{N_j}{N} \right) \right] \right] , \]

where \( N_i \) is the sum of row \( i \), \( N_j \) is the sum of column \( j \), and \( N \) is the sum of all elements. A higher value of the normalized mutual information indicates a better community detection performance.

### 4.3 Results and Discussion

Table 2 reports the comparison of community detection performances on different portions of emotion information in pairwise t-test (i.e., consider each generated network as a data instance, we compare the three programs’ NMI measure on it). In more than 180 of the 225 parameter settings (i.e., 225 pairwise t-tests), including (partial) emotion information significantly improved community detection performances as compared with excluding emotion information at a 99% confidence interval. In addition, using all emotion information could outperform using only partial emotion information in 127 parameter settings at a 90% confidence interval. We visualize the results of the comparison between all emotion and partial emotion in Figure 3, where a darker cell shows the higher performance improvement. It is shown that when \( \alpha \) is larger than 0.2 and \( \gamma \) is larger than 0.5, using all emotion information would have significantly more benefits as compared with using only the positive ones in most of the parameter settings. In other words, the GN-H algorithm can better take advantage of the negative links if a certain amount of positive links already existed in the network. In real life, we expect most networks have a large portion of positive emotions and a small portion of negative emotions, thus the GN-H algorithm could be beneficial in many of the applications.

<table>
<thead>
<tr>
<th># of settings</th>
<th>partial emotion &gt; no emotion</th>
<th>all emotion &gt; no emotion</th>
<th>all emotion &gt; partial emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>p&lt;0.01</td>
<td>186</td>
<td>184</td>
<td>104</td>
</tr>
<tr>
<td>0.01&lt;p&lt;0.05</td>
<td>0</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>0.05&lt;p&lt;0.10</td>
<td>2</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>0.10&lt;p</td>
<td>37</td>
<td>36</td>
<td>98</td>
</tr>
</tbody>
</table>

**Table 2. Comparison of community detection performances (pairwise t-test)**

![Figure 3. Performance improvement of using all emotion over using partial emotion (x: p > 0.1; y: p > 0.05 and < 0.1; z: p > 0.01 and < 0.05)](image-url)

We further inspected the relationship between community detection performances and the values of the three parameters on simulated data. It is found that the community detection performances of all algorithms improve if the network contains more positive links (i.e., \( \alpha \) increases) and the positive links can better show the actually community structure (i.e., \( \gamma \) increases). In addition, if the negative
links can better represent the community structure (i.e., $\beta$ decreases), the performance of the algorithm that uses all emotion information improves, the performance of the algorithm that uses partial emotion information does not change, and the performance of the algorithm that does not use any emotion information has a decreasing trend. Obviously, mistaking negative emotions for positive ones will harm the models’ community detection performances. It is critical for us to have a better assessment of interaction emotions before conducting community detection to analyze virtual communities.

5 CASE STUDY

5.1 Dataset

We collected a Web forum dataset from the “WalMart sucks board” on the WalMart-blows Website (http://www.walmart-blows.com/forum/viewforum.php?f=3) for our case study. The data are from November 2003 to November 2008 and includes 1,354 threads with 19,624 postings contributed by 1,855 users. The forum participants were mainly employees and customers who complained about their experience with WalMart. Although the Website contains mainly negative opinions of WalMart, the social interactions on it still have both positive and negative emotions (i.e., some members may agree with each other about the bad side of WalMart). We chose this Website since it represents the Websites business managers may have an interest in to understand their customers and employees.

As explained in Section 3.2, we applied a heuristic algorithm to extract posting-level interactions. To estimate the effectiveness of this step, we randomly selected 62 threads with 993 post-reply relations from the WalMart dataset and coded them for a small-scale evaluation. Experiments showed that the interaction extraction algorithm achieved 92% accuracy on this testbed.

For the emotion extraction stage, we coded the postings in 62 threads as agreement/disagreement relations. During this process, there are only two post-reply pairs cannot be categorized to this two groups due to sudden topic change. After feature extraction and selection, we build the classifier using the libSVM package (Chang & Lin, 2001) and conduct 10-fold cross validation on the testbed. The emotion identification part achieved 85% accuracy in this experiment.

After interaction and emotion extraction, we aggregated the identified posting-level interactions and emotions to an emotion social network. During this process, we discard the links between user pairs who only have one posting-level interaction to focus on the more reliable social relations. The final network contains 5,450 positive links and 110 negative links between 951 users. The users that are isolated nodes in the network have been removed from the study. We believe they do not significantly affect the results of our analysis due to their small contribution to the forum.

Similar to the experiments on the simulated dataset, we generated networks with all emotion, partial emotion and no emotion. However, since we do not have knowledge on the underlying community structure of this real-world dataset, we focus on exploring the utility of the framework by inspecting the community structure, user activities, opinions, and opinion leaders of the detected communities.

5.2 Results and Discussion

On the Web forum dataset, the community detection algorithm on the networks with all emotion, partial emotion, and no emotion information generated 61, 94, and 92 communities, respectively. Figure 4 shows the community-level topology of the emotion social networks extracted from the Web forum dataset, in which each node indicates a detected community and the node size indicates the number of users in each community. From the visualization and the topological measures, it is clear that the community-level structure detected from the network significantly changes when considering emotion information. Furthermore, the use of emotion information caused intra-community negative links to reduce from 2.32% (no emotion) to 1.06% (partial emotion) and 1.14% (all emotion). The algorithms using emotion information put more negative links between communities.

It should be noted that major communities detected under the three circumstances have large overlaps.
For the three largest communities, about 80% of the users are the same across different algorithms. However, the changes of the other 20% of users’ communities have significantly affected the community-level structure. Further inspections on such users provide us several examples that involved in conflicts/disagreements and better grouped with emotion information.

<table>
<thead>
<tr>
<th></th>
<th>No emotion</th>
<th>Partial emotion</th>
<th>All emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of communities</td>
<td>61</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>Community-level links</td>
<td>93</td>
<td>74</td>
<td>73</td>
</tr>
<tr>
<td>Avg. degree</td>
<td>3.0492</td>
<td>1.6809</td>
<td>1.5870</td>
</tr>
<tr>
<td>Avg. shortest path length</td>
<td>1.6710</td>
<td>2.5423</td>
<td>2.4817</td>
</tr>
<tr>
<td>Avg. clustering coefficient</td>
<td>0.8517</td>
<td>0.7033</td>
<td>0.7434</td>
</tr>
<tr>
<td>Giant component size</td>
<td>22 communities / 93 links</td>
<td>51 communities / 73 links</td>
<td>49 communities / 72 links</td>
</tr>
<tr>
<td>Intra-community negative interaction ratio</td>
<td>2.32%</td>
<td>1.06%</td>
<td>1.14%</td>
</tr>
</tbody>
</table>

Figure 4. Community level structure of the Web forum dataset

We inspected the activities of the users in the three major communities in the Web forum. We found that the joining time patterns of the users in them have some coherency. Specifically, communities A, B, and C contain mainly users who joined the Web forum in the middle of 2006, at the beginning of 2007, and at the end of 2005, respectively. Users who joined the forum at a similar time may have a higher chance to interact with each other. They may also be attracted to the forum by same events. These two reasons increase the probability for them to form communities.

We inspected the major discussion topics of the users in the major communities with the help of a topic extraction tool, Arizona Noun Phraser (Tolle & Chen, 2000). It is found that the three identified communities shared some common topics and displayed a certain extent of topic coherence in their discussion topics. For example, the users in community A were involved in the discussions on “Sam Walton” and “store manager” more frequently than the users in the other two communities. Users in community C show more interest in the topics “low price” and “minimum wage.” Further inspection shows that when using emotion information, the topics in major communities show relatively higher coherency. For example, without the emotion information, community C would not have distinct differences from the other two communities on “low price” and “minimum wage.”

Figure 5. Top 10 high-degree users in the major communities
We inspected the opinion leaders in the major communities according to their degree in the community. In general, these opinion leaders are different from the ones identified using the degree on the entire network. Figure 5 reports the top 10 opinion leaders for the major communities. The top opinion leaders of the major communities generally ranked around 3-15 on the entire network. We believe that the opinion leaders identified in communities have stronger influence in their local communities, whereas the high-degree users on the entire network may have many postings but do not have a consistent interest pattern to build their identities in the communities. In addition, we noticed that the identified opinion leaders have a higher overall ranking (i.e., a smaller ranking index) when we use emotion information, especially in communities B and C. Since the opinion leaders identified without emotion information may rank nearly 100 among the total 951 users, we tend to believe that using emotion information provide us a better set of the opinion leaders.

6 CONCLUSION

In this research, we intend to account for emotions of social interaction to tackle the community detection problem. We propose a framework to extract social interaction and emotions from user-generated contents and design a GN-H co-training algorithm that can combine emotions with social networks for community detection. Experiments using the simulated data show that including emotion information would significantly improve the effectiveness of community detection. Experiments on a real-world dataset show the utility of the proposed framework in virtual community analysis.

The research lends empirical support to the affect theory of social exchange with applications in virtual community studies. Through this research, we expect to promote future studies that combine social network analysis with interaction emotions in examining social phenomena in Web 2.0. In the future, we plan to design more effective algorithms for emotion-aware community detection. We plan to explore the role of interaction emotions in virtual community from a more theoretical perspective.

References


