Social Computing and Weighting to Identify Member Roles in Online Communities

Robert D. Nolker
University of Maryland, Baltimore County
Rnolker1@umbc.edu

Lina Zhou
University of Maryland, Baltimore County
Zhoul@umbc.edu

Abstract

As more and more people join online communities, the ability to better understand members' roles becomes critical to preserving and improving the health of those communities. We propose a novel approach to identifying key members and their roles by discovering implicit knowledge from online communities. Viewing an online community as a social network connected by poster-poster relationships, the approach takes advantage of the strengths of social network analysis and weighting schemes from information retrieval in identifying key members. Experimental studies were carried out to empirically evaluate the proposed approach with real-world data collected from a Usenet bulletin board over a one year period. The results showed that the proposed approach can not only identify prominent members whose behaviors are community supportive but also filter chatters whose behaviors are superficial to the online community. The findings have broad implications for online communities by allowing moderators to better support their members and by enabling members to better understand the conversation space.

1. Introduction

As more and more people participate in online communities, there is an emerging need to discover deep-level relationships from their interactions. An improved understanding of human-human and human-information relationships can lead to more effective use of the space [1]. Therefore, we aim to support and improve the health of online communities.

Identifying key members and their roles in an online community is an important way to support the needs of an online community [1, 2]. Its significance lies in that members whose relationships exhibit those traits that promote the health of the community should be encouraged and even rewarded. Much work has focused on analyzing the space of online conversations using information visualization [1-4], which is primarily based on straightforward participation frequency. Other factors that are potentially useful to identifying key members are poorly understood and under studied. It is highly desired to improve our understanding of key members in an online community by looking deep into members' behavior and relationships.

Viewing an online community as a social network connected by member-member relationships, the approach takes advantage of the strengths of social network analysis and weighting schemes from information retrieval in identifying key members. By overcoming the limitations of extant approaches that are based on raw frequencies, the proposed approach is expected to provide new insight into member's roles in an online community and reveal members that would otherwise be misidentified.

This paper makes multifold contributions to the body of literature on online communities: 1) it advances our understanding of the complex relationships that exist in an online community; 2) it refines the classifications and definitions of key members in online communities based on deep-level relationships; and 3) it proposes a social computing approach to identifying key members. Compared with a traditional ethnographic approach to online community data, which requires one to read and analyze hundreds, if not thousands, of postings manually, the social computing approach presented here is much more efficient and effective by automatically identifying those members that warrant additional attention.

The remainder of the paper is organized as follows: Section 2 introduces member relationships and key members in online communities; Section 3 proposes an approach to identifying the roles of members in online communities. Section 4 reports and discusses the evaluation results on real world Usenet data; and Section 5 concludes the findings of the paper and presents future research directions.

2. Background

The first step in understanding roles within online communities is identifying key members and assessing

Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)
0-7695-2415-X/05 $20.00 © 2005 **IEEE**

their potential impact on the communities. In particular, key members and roles are discussed from a social network perspective. Existing methods to identify key members and asses their roles are also introduced.

2.1 Member relationships

A social network is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange [5]. The social networks of today are no different from the pre-PC age except now we include computer networks that connect people. These networks of people connected in cyberspace become online communities.

To identify key members in an online community, it is important to understand the types of relationships that people engage in. Despite direct relationships created by member-to-member interactions, the strength of relationships can vary widely from very weak to very strong. Moreover, as the number of members in an online community increases, the complexity of the relationships between any pair of members increases exponentially, for member A may interact with member B indirectly via many other members. It is even more challenging to identify the role of a member to the community as a whole. Community members post information for the entire community to access. The access of information that others have produced or contributed also produces relationships. While these relationships are indirect [6], they are useful to understanding the relationships within a community. To best serve the needs of a community, it is highly desired to fully understand all of the relationships.

2.2 Key roles

Understanding organizational structure of online communities requires identifying those individuals that have either a positive or negative effect on the group. Different communities have different needs and the roles that support these needs are different. A software bulletin board needs experts whereas a health support group may need a combination of information providers and empathetic listeners. An open subject discussion group needs both listeners and conversationalists. These roles within a community can be defined by relationships between different members, member behaviors, or a combination of the two. Relationship-based roles (e.g., gatekeepers [7]) follow the traditional network node structure of hubs, brokers, and bridges [8]. Behavior-based roles are defined on the basis of an individual's behavioral patterns. Examples are pollinators, debaters, spammers, and conversationalists [9]. The identification of roles is more of a point of reference than a factual statement.

Key members are defined as those who contribute to the success and health of the community. For this research our focus was on an open discussion bulletin board. Based on the review of relevant literature and our analysis of the online community data, we identified two roles that are important to this type of community as follows:

Leaders: individuals that are in a position to spread knowledge; and provide cohesiveness and consistency

Motivators: individuals that keep conversation going.

These members are crucial to the success of a community because they provide activities that help maintain critical mass [10]. Both of the key roles are defined based on the combination of member behaviors, conversation, member relationships, and social networks.

2.3 Chatters

Based on our experience and observations of online communities, high volume members may not be supportive of the community. This raises the issue of separating the contributing key members from those whose conversation patterns are not community supportive. This involves identifying activity patterns that are characteristics of chatter and then using those patterns to filter chatters. In a Usenet group, Peter and Sam are engaged in protracted discussions within single threads, but they rarely get involved in other discussions. Based on the traditional definition of key roles, Peter and Sam are likely to be key members due to the sheer volume of their postings. However, they do not really contribute to the community conversation as a whole. Therefore, these members may well be characterized as chatters. This chatter is not based on the content of their responses but rather on the pattern of the response relationships. We believe that members that fill key roles should have low chatter rates.

2.4 Identifying the key members

It is crucial to identify key members within a community so that group administrators know whom to support and whom to watch. Much of the current work on identifying members who fill the key roles within the Usenet community has focused on frequency of participation and volume of contribution [1, 2, 11, 12]. A few studies have incorporated threads and members' relationships to threads in their analyses [1]. However, the roles of members in an online community have not been investigated by taking a holistic view of member relationships in the entire community. Moreover, there is a lack of attention to non-supportive chatters in online communities.

Drawing on the relevant literature in online communities, social computing, and Web intelligence, we develop a new approach to identifying both key members and chatters in online communities, which will be introduced in the next section in detail.

3. An Approach to Identifying the role of Members

We propose an approach to analyzing the roles of memberships in online communities by using social network analysis and member weighting. "Social network analysis conceives of social structures as the patterned organization of network members and their relationships [13]". It helps simplify the complexity of social networks by applying network analysis. As a result, users can better understand the network and the embedded relationships. Our hypothesis is that by applying social network analysis to online communities, we would be able to discover deeper knowledge about posting patterns that will in turn lead to better identification of members that fill in two key roles. Usenet groups, a popular form of online communities, were selected as the community platform for analysis.

3.1 Key concepts and definitions

We begin by viewing postings in Usenet groups as parts of conversations. By defining conversation we are able to add a new dimension to the relationships that is beyond the scope of volume/frequency of postings, as used in previous studies. The conceptualization of conversation is central to our methodology. Classifying conversation types can lead to better identification of specific roles. Thus, we first give key concepts and their definitions:

One-way conversation – if member A responds to member B, A has a one-way relationship with B. The response is limited within the thread. The one way response counts were closely related to the volume but differred in that posts that did not receive response and thread heads (initiating posts) were not counted.

Two-way conversation – if there are reciprocal one-way relationships from member A to B and from B to A, A has two-way relationships with B and vice versa. These two way relationships were the heart of our proposed method. By tying two members together in a stronger relationship than just volume, we were able to look for community conversation patterns. The two-way relationships were further divided into direct and indirect categories.

Direct two way conversations - if Sam responded to Peter, who in turn responded to Sam within the same thread, they have a direct two-way conversation.

Indirect two way conversations - if Sam responded to Peter and, over the timeframe for analysis, Peter responded to Sam in a different thread, they have an indirect two-way conversation

Discussion ratio – It is the ratio of indirect two-way conversations to direct two way conversations.

The above concepts introduced in this study play an important role in distinguishing roles and community conversations. For example, a member with a low discussion ratio is involved in fewer direct two-way conversations relative to their indirect conversations indicating less interplay (discussion) with the members they interact with. On the contrary, a high discussion ratio indicates that a large percent of their two way conversations are more intense back and forth discussions.

3.2 Membership metrics

According to the conceptualization of conversation, replies to posts are analyzed for their position in the pattern of conversation. Member-member matrices are created based on one-way and two-way conversation relationships respectively. Instead of using binary values to represent elements of the matrices, we computed the frequency of conversations, which is expected to be a more accurate measure of the conversation relationships.

To understand conversation and its interplay with relationships, the relationship-based measures from the social network paradigm (i.e., degree, betweeneess, and closeness) were combined with behavior-based measures from the information retrieval realm (i.e., TF*IDF) to determine the key members.

Degree – the number of conversations that an member is engaged in / the number of members that an member has conversed with.

Betweenness – the number of pairs of other members who can converse with each other indirectly through an member with shortest relay.

Closeness – average conversation distance between an member and all the others in the community.

TF*IDF (Term Frequency Inverse Document Frequency) has been traditionally applied to identify key terms (T) from text documents (D), as shown in Formula (1). It consists of two parts: TF and IDF. The basic idea is that the importance of T to D is proportional to the frequency of T in D and inversely proportionally to the total number of D that T appears in.

$$TF*IDF = \log(TF+1)*\log(|D|/DF) \qquad (1)$$

We innovatively apply TF*IDF to determine the conversation importance of a poster to a thread and the conversation importance of one poster to another poster by assigning new meanings to variables T and D. For the former (post-to-thread importance), D represents the member responding posts and T a thread. For the latter (poster-to-poster importance), D denotes the member responding and T the poster being responded to. Applying this weighting to the different conversation types, we build a TF*IDF spectrum ranging from those who have a few posts to lots of members to those who have lots of posts to a few members. The position of members within this continuum provides an in-depth analysis of the pattern of their relationships. It is noted that the TF*IDF schema was not used to analyze the content of postings but rather to weight the conversation relationships between members. Moreover, we extend the scope of the weighting schema by introducing the metrics of overall importance of T, which is measured with the average TF*IDF weights of T across all D.

3.3 Attributes of key roles

Based on the definition of key roles and conversation relationships, we can identify attributes indicative of key roles and develop criteria for measuring the roles with membership metrics, as shown in Table 1a-c.

By applying TF*IDF to the conversations we aim to filter those members that are just chatters and identify those that are involved in more community supportive conversation patterns. High overall TF*IDF scores indicate members whose conversation pattern falls into the category of chatter, for they have protracted conversations with very few other members. It should be noted that chatters are not inherently bad, and may play a role in the community, but our focus is to identify those key members that play a community supportive role. The key community members will have lower average TFIDF scores as they would be involved in conversations with many community members.

Table 1a. Attributes and Measures of Leaders

Attributes	Measures
Leaders	
This poster contributes and receives many responses.	High Degree in all conversation sets.
Many other conversations pass through this member	High Betweenness in both one-way and two way conversations.
The poster responds to many other posters but does not respond a lot to any one poster.	Low TFIDF based on one-way and two way conversations.
Has a mix of responses, both direct and indirect two way.	Moderate discussion ratio

Betweenness plays an important role in predicting the relationship potential that a member has with other members. This potential can be derived from the past relationship patterns in community conversation. We believe that in-closeness and out-closeness will further delineate the leaders from other members. A high in-closeness would indicate a member who provides consistency and high out-closeness would indicate a member who spreads knowledge [12]. Motivators do not necessarily have a high betweenness ranking, because they may be central to cliques rather than the community as a whole.

Table 1b. Attributes and Measures of Motivators

Attributes	Measures
The average distance to all other members, puts this individual in the middle	High closeness
High posting count spread evenly over lots of threads	Low thread IDF and a low one- way conversation IDF
Has a mix of responses, both direct and indirect two way.	Moderate discussion ratio

Table 1c. Attributes and Measures of Chatters

Attributes	Measures
Talk a lot but only to a few people.	High TF*IDF in two way conversations
Majority of their two way conversations are direct	High discussion ratio

IDF is important to identifying motivators. The low one -way IDF indicates the member is involved with many other members. However, it alone does not tell the whole story. The second factor, low thread IDF, rounds out the information. By also having a low thread IDF, the pattern of communication becomes a member involved with many other members over many threads. The motivator is active throughout the community.

4. Evaluation and Results

The proposed approach was evaluated against a baseline created using network analysis tools on the volume/ frequency data for the group. The network analysis was accomplished through the use of the software tool, Ucinet [14]. By using the volume / frequency network data we were able to compare our results to those currently available to community administrators.

4.1 Selection of online community

The focus on conversation drives our selection of the online community group. In Usenet groups where the role of information provider/expert is salient, it is relatively easy to identify key roles. Thus, we selected open subject discussion type of Usenet bulletin board,, in which.the roles of members was implicit or hidden. The group selected was alt.support.loneliness (ASL), a community, whose size places it 28th out of 385 groups in the alt.support category. This group was selected from a list of open discussion groups in the alt.support community for its relative position. Not the largest group, but one having reached some form of critical mass [10]. The group is not moderated. Our goal was to capture deeper meaning than the traditional volume frequency approach alone yielded.

4.2 Data collection and preparation

Our data was harvested from the Giganews Usenet news server which contained 25,737 postings from 1-1-2004 to 12-31-2004 for ASL.

The following fields of Usenet bulletin boards were collected: header ID, From, Subject, Message-ID, Date and References. The references field was parsed to gather only the reference to the post the message was posted to. This data set was cleaned by removing records with missing fields, yielding 25,308 records. Additional 2702 thread heads were removed because the posters were not responding to another member and therefore there was no conversation to measure. The remaining 22,606 posts were used to analyze relationships and communication patterns. The statistics of UseNet group is shown in table 2.

The posts were transformed into five different matrices, including a raw volume/frequency member-member matrix, a one way conversation member-member matrix, a two way indirect conversation member-member matrix, a two way direct conversation member-member matrix and a conversation weighted member –thread matrix. Each of the matrices is then processed with both social network and weighting schema to discover implicit patterns.

Table 2. Statistics of the UseNet group

Members	1,319
Members in one-way relationships	1,140
One-way conversations	6,831
Members in two-way relationships	638
Two-way conversations	3,533

4.3 Results and discussion

The combination of centrality measures from social networks and the TF*IDF weighting metrics yielded interesting results. The discovered patterns of relationships support the identification of three roles, leaders, motivators and chatters, in the top posters.

Among the 1,319 members in the group, the top 42 high-volume responders were selected for manual validation. These responders contributed 72% of the responses and encompass all the identified significant members while representing the members most likely to be key members from the perspective of community conversation. The contribution by the 42nd member ranked by volume is less than 8 responses per month on average, the group mean is 20 with a median of 2. Table 3 lists the top responders, by volume, with their role, key social network measurements and TF*IDF measures.

4.3.1 The leaders

One leader clearly emerged for this group. Member 258 stood out in network measures such as in-degree closeness, betweenness, and degree as well as in the TF*IDF scores. This member was ranked 8th in volume of posts. Other members in the top ten posters list ranked significantly lower on the variables. The network measures clearly identified this individual as an important member, but it is the communication weighting schema that distinguishes their roles. Therefore, it was the combination of member 258's rank in the network and their communication behavior pattern that separate them as a leader. The leader role is not exclusive and may be occupied by several members. In the ASL community, the other high volume members display traits more in line with other roles, motivators or chatters, in terms of their network position and communication patterns.

4.3.2 The motivators

These members have many of the same characteristics of the leaders. Five members were identified as being motivators. All five ranked high in in-degree closeness, betweenness and degree on the network side and had low TF*IDF scores. In addition, they also had the lowest IDF scores in the one way response and thread data sets. The low one way IDF score indicated that they were involved in responding to many different members. The low thread IDF indicates that they participated by responding in many threads. These members are involved at a community level, responding to many other members in many conversations (threads). It is this response involvement that leads to a critical mass and is an important part of a healthy community. Four additional members had conversation weighting scores that indicated they may be motivators, but their network position and IDF scores did not support their being in that role. Our methodology enabled these four members to be identified such that they can be encouraged to contribute and play a greater role in the community.

4.3.3 The chatters

Five members of the top 42 responders by volume were identified as chatters. Although all the five had weak network positions, they were set apart from other members by their high TFIDF scores and their high percentage of direct two way responses in relation to indirect two way responses. The high TFIDF indicates that they talk to only few people but talk a lot with those few, and the high direct two-way response ratio indicates a protracted

Table 3 Top posters in ASL 2004 ranked by volume

	Volume		Two Way Indirect			Discussion	One Way	Thread
Responder	Ranking	Role	TFIDF	Between	Degree	Ratio	IDF	IDF
344	1	Motivator	4.393	3.097	87.683	47.17%	1.614	0.94
693	2	Motivator	3.829	3.607	71.457	28.68%	1.668	1.078
1082	3	Motivator	4.572	2.029	68.563	49.03%	1.988	1.439
782	4	Motivator	4.804	1.424	44.260	34.51%	2.076	1.964
1135	5		6.205	1.003	32.357	31.82%	2.528	1.851
511	6	Chatter	9.136	0.209	22.011	25.76%	2.880	1.812
889	7	Motivator	4.782	2.055	42.033	26.56%	2.126	1.823
258	8	Leader	2.873	5.972	92.546	45.23%	1.802	2.335
533	9		7.446	0.272	20.269	20.46%	2.911	2.007
339	10		5.863	0.456	24.728	32.92%	2.735	2.525
	…							
742	26	Chatter	8.063	0.400	16.908	86.16%	3.743	5.216
340	27		7.006	0.222	11.808	29.58%	3.483	3.370
834	28	******	4.877	0.434	17.791	32.11%	2.819	3.502
436	29	Chatter	9.244	0.486	7.808	44.96%	3.861	3.235
	…							
1157	37	******	4.964	0.679	13.882	51.25%	3.050	3.851
	…							
968	41	******	4.833	0.541	15.234	27.38%	3.232	3.657
1256	42	******	4.403	0.585	15.573	44.74%	3.147	4.204
	…							
1193	120	Chatter	21.966	0	1.666	50.00%	5.940	6.314
842	220	Chatter	17.044	0	1.293	83.30%	6.346	7.007

****** - Member has strong community communication pattern but is a low contributor. This member should be encouraged to participate more.

two way conversation pattern. Despite the potential contribution of these individuals to the community, this study suggests that their participation behavior is not community supportive.. Two of the chatters were found to have the highest TFIDF scores, 0.29 and 0.14 respectively.. The members identified by our method require further scrutiny by a forum/group administrator. This method makes no statement about the content of conversations but focuses on the conversation patterns that distinguish some members from the rest of the community.

A comparison of the selected key members with those in the volume baseline showed that six of the ten top posters filled roles. Four members that may have been over looked in volume / frequency analysis were identified as having strong community conversation patterns. These four members are candidates to receive additional support and encouragement from the community. Finally using the methods presented in this paper, the pattern of communication put several members at the chatter end of the communication spectrum, talking to only a few people but talking with those people a lot. These chatters while not contributing to community conversation are not inherently bad for the community.

5. Conclusion and future work

The novel approach of combining TF*IDF weighting with social computing enabled us to identify key members at a deeper level in Usenet groups. Our approach was found to be very effective in identifying leaders and filtering high-volume members whose conversation patterns were not community supportive. Being able to single out members who are just high volume allows administrators and managers to focus on those group members of whom additional support will provide the greatest return for the community.

A critical step in processing the real-world data from the Web is data preparation. Analyzing Usenet group postings is no exception. We had to deal with missing information and inconsistent information. As a result, we developed some heuristics in cleaning the data and some data were dropped due to the lack of information. Our experience in preparing the data can be valuable to other researchers who are interested in the same line of research.

Several issues become apparent during this study that would be avenues for continued study. First, a support-oriented discussion bulletin board was selected as the test-bed for analysis in this study. We believe that the proposed approach to identifying key members, which integrates social computing and weighting, is generalizable to other types of open subject discussion bulletin boards such as sports or politics. This needs to be validated in a future study. Second, since a poster has the flexibility to bypass other posters and respond where they feel fit in a Usenet group, the power of betweenness requires re-interpretation and re-evaluation. Third,. The time interval chosen, one year, was arbitrary. Usenet bulletin boards can be a transitory space, so further research needs to be conducted to determine the time intervals that produce the most accurate results. Finally, the ecological validity of the proposed method can be significantly enhanced by involving moderators and administrators in the evaluation of the findings.

7. References

[1] M.A. Smith and A.T. Fiore, "Visualization components for persistent conversations", in *Conference on Human Factors in Computing Systems*, 2001, Seattle, Washington, United States, ACM Press New York, NY, USA, 136 - 143

[2] J. Donath, K. Karahalios, and F.B. Viegas, "Visualizing Conversation", *Journal of Computer-Mediated Communication*, Publisher, June 1999.

[3] J.C. Paolillo, "Visualizing Usenet: A Factor-Analytic Approach", in *33rd Hawaii International Conference on System Sciences*, 2000, Hawaii, IEEE, 1-10.

[4] F.B. Viegas, M. Wattenberg, and D. Kushal, "Studying Cooperation and Conflict between Authors with history flow Visualizations", in *CHI 2004*, 2004, Vienna, Astria, ACM Press, 1-8.

[5] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying Online Social Networks", *Journal of Computer-Mediated Communication*, Publisher, June 1997.

[6] F.-r. Lin and C.-h. Chen, "Developing and Evaluating the Social Network Analysis Systems for Virtual Teams in Cyber Communities", in *37th Hawaii International Conference on System Sciences*, 2004, Hawaii,

[7] J. Xu and H. Chen, *Untangling Criminal Networks: A Case Study*, H.e.a. Chen, Editor. 2003, Springer-Verlag: Berlin Heidelberg. p. 232-248.

[8] P.J. Denning, "Network laws", *Communications of the ACM*, Publisher, November 2004**,** pp. 15-20.

[9] F.B. Viegas and M.A. Smith, "Newsgroup Crowds and AuthorLines:visualizing the Activity of Individuals in Conversational Cyberspace", in *HICSS 2004*, 2004,

[10] J. Preece, *Online Communities Designing Usability, Supporting Sociability*. 2000, New York: John Wiley & Sons.

[11] R. Xiong, M.A. Smith, and S.M. Drucker, "*Visualizations of Clooaborative Information for End-Users*. 1998, Microsoft.

[12] R.A. Hanneman, "*Introduction to Social Network Methods*. 2001. p. 1-150.

[13] B. Wellman, "For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community", in *1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*, 1996, Denver, Colorado, United States, ACM Press New York, NY, USA, 1-11.

[14] S.P. Borgatti, M.G. Everett, and L.C. Freeman, *Usinet 6.0 Version 1.00*. 1999: Natick: Analytic Technologies.