Identifying Leaders in an Online Cancer Survivor Community

Kang Zhao[1+], Baojun Qiu[+], Cornelia Caragea[+], Dinghao Wu[+], Prasenjit Mitra[+], John Yen[+],
Greta E. Greer[*], and Kenneth Portier[*].
[+]The Pennsylvania State University, University Park, PA 16802, USA.
[*]American Cancer Society, Inc. Atlanta, GA 30303, USA.

Abstract

Online communities are an important source of social support for cancer survivors and their informal caregivers. This research attempts to identify leaders in a popular online forum for cancer survivors and caregivers using classification techniques. We extracted user features from many different perspectives, including user-contribution, network, and semantic features. Based on these features, we further exploited the structure of the network among users and generated new neighborhood-based and cluster-based features. Classification results revealed that these features are discriminative for leader identification. Using these features, we developed a hybrid approach based on an ensemble classifier that performs better than many traditional metrics. This research has important implications for understanding and managing similar online communities.

1. Introduction

Cancer accounts for 1.8 million deaths in China, 0.57 million in the U.S. (both in 2010), and 7.6 million deaths worldwide in 2008. As of 2007, 11.7 million Americans have been diagnosed with cancer and are either living free of cancer or still have evidence of the disease. Social support can help cancer survivors cope better with the disease and improve their life [1]. Moreover, peer support and information can be quite valuable, since cancer survivors' experiences are unique and may not be well understood by their family members, friends or care providers.

As 83% of adult Internet users in the U.S. utilize the Internet for health-related purposes, many cancer survivors also share social support online. An online health community (OHC) typically provides interactive features such as discussion boards and chat rooms that afford users the ability to connect and communicate with one another to share their experiences and support. Research of online communities shows that a small group of leaders may have a significant amount of influence and impact over others [2]. Identifying leaders has utilities for community building and management, marketing, and information retrieval and dissemination. For an OHC specifically, finding leaders may have additional implications in advocating new treatments, guiding the proper use of drugs, and encouraging healthy lifestyles and positive attitude [3, 4].

In this research, we show how leaders can be identified automatically from online web forums. To validate our method, we use de-identified data from the online forum in the American Cancer Society Cancer Survivors Network® (CSN). CSN (http://csn.cancer.org) is an OHC of more than 146,000 registered members created by and for cancer survivors and caregivers. Its online forum consists of 38 discussion boards. Our dataset includes posts from July 2000 to October 2010, comprised of 48,779 threaded discussions with more than 468,000 posts from 27,173 users. This paper reviews related research, describes our classification-based approach and illustrates our results, and discusses implications and future research directions.

2. Related Work

Researchers have proposed many ways to identify or rank key users in various online communities. However, finding the correct metric is an art that is very community-dependent, especially if multiple metrics must be incorporated. The use of general metrics of network centrality hardly suffices. For example, to find influential users in a community that features significant contagion

[1] Corresponding author. Address: 324 IST Building, University Park, PA 16802, USA. Email: kangzhao@psu.edu

phenomena in a network setting, such as information spreading, sharing and viral marketing, one can use contagion maximization [5]. In a Q&A community, an expert finder can exploit the difference in knowledge/expertise between askers and answerers, and the link structure of the Q&A network [6]. The blogosphere depends highly on bloggers' contributions reflected by the number and length of posts, and reader activities (as comments) generated by a blogger's posts, etc. [7]. Research also suggested that analyzing how the content of users' posts relate to each other can help reveal influential users in a dark web forum that aims at ideological propaganda [8].

What is an accurate metric to identify leaders in an online forum of a peer support community? The forum provides multiple types of social support to its users, including information support (e.g. asking and answering questions about a specific treatment or diagnosis), emotional support (e.g. sharing experience, seeking prayers, encouraging others), and companionship (e.g. initiating discussions about weekend plans or playing online Scrabble games) [9]. In such a multi-purpose and interactive community, many features reflect a user's contribution. A combination of features must be considered to form an accurate metric. For example, to become a leader, one may need to actively start and follow threads, possess network centrality, and post content that brings a positive spirit to the community. Consequently, we used classification techniques to select an accurate metric from the "big data" of users' online activities and distributed interaction.

3. Analysis and Results

3.1. Data preparation and initial results.

Three groups of basic features were extracted from the forum for leader classification: contribution, network, and semantic features. Contribution features, as the name implies, measure a user's direct contribution to the forum, such as number of discussion threads (topics) initiated and replies posted, number of days the user has actively posted, length of the user's post, etc. Network features reflect users' centrality (e.g. in/out-degree and betweenness) in a post-reply network, where there is an edge pointing from user A to user B if user A replied a thread started by user B. Semantic features are derived from the content of users' posts. They reflect positive or negative sentiment, emotional strength, diversity of topical coverage (utilizing Latent-Dirichlet Allocation), etc. of a user's posts. Table 1 summarizes these features.

We used classification techniques to distinguish between leaders and regular users. To train a classifier, we first label a set of users as either a leader (a positive record, +1) or non-leader (a negative record, -1). While it is easy to label users whose activity levels are very low as non-leaders, finding a leader requires good knowledge of the user's activity history and other users' reactions to the user's posts over an extended period of time. As this task is very difficult even for someone who knows the forum well, domain experts were consulted. The CSN community manager and two staff members who monitor the forum content on a full-time basis were asked to nominate leaders. 41 members were nominated as leaders. Although a subjective and imperfect labeling, the non-exclusive list provides a good starting point and helped us understand what type of behaviors contributes to leadership and how to improve classifiers.

Identifying community leaders posed a challenge for our classifiers--an unbalanced dataset. As only a small number of users are leaders and most users are not, the dataset contains many more negative than positive records. With such an unbalanced dataset, a regular classifier is biased towards simply classifying all users as non-leaders. To tackle this problem, we processed user features in two steps. First, assuming that it is almost impossible for one to become a leader with little contribution to the forum, we ruled out obvious non-leaders using two empirical rules based on users' activities: (1) users whose total number of posts is below average (17 posts) or (2) users whose total number of active days is below average (7.7 days) are non-leaders. In fact, all 41 nominated leaders made a considerable contribution in terms of total posts and active days. With the two rules, 91.4% of all users in the dataset were eliminated from consideration as leaders, leaving 2,336 users as unclassified, but the much smaller dataset with 2,336 users was still unbalanced. Second, we over-sampled positive records 20 times, a common practice in dealing with unbalanced datasets, to raise the ratio between the numbers of positive and negative records to about 0.36. This step provided us with a more balanced dataset (hereafter referred to as *Dataset-1*) with which to train classifiers.

Table 1. Summary of basic user features

Group	Features
Contribution features	Number of one’s initial posts (i.e., posts that start threads)
	Number of one’s replies to others (i.e., following posts)
	Number of threads that one contributed post(s) to
	Number of other users’ posts published after one’s post in the same thread
	Avg. response delay btwn. one’s post and the next post by others in the same thread (in minutes)
	Total length of one’s post (in bytes)
	Avg. length of one’s post (in bytes)
	Avg. content length of one’s top 30 longest posts (in bytes)
	Number of one’s active days (one published at least 1 post in an active day)
	Time span of one’s activity (from first active day to the last)
	Avg. number of posts per active day
	Avg. number of posts per day during one’s time span of activity
Network features	One’s in-degree and out-degree in the post-reply network
	One’s betweenness centrality in the post-reply network
	One’s PageRank in the post-reply network
Semantic features	Avg. percentage of words w/ positive sentiment in one’s posts
	Avg. percentage of words w/ negative sentiment in one’s posts
	Avg. percentage of Internet slangs/emoticons in one’s posts
	Avg. percentage of words w/ strong emotion in one’s posts
	Ratio between the numbers of words w/ positive and negative words in one’s posts
	Topical diversity (Shannon entropy and log energy of topic distribution in a user’s posts)

We applied 5 classifiers, Naïve Bayesian, Logistic Regression, and Random Forest, one-class SVM, and two-class SVM, to *Dataset-1* using 10-fold cross-validation. We also needed a proper metric to evaluate the performance of these classifiers. Traditional classification metrics, such as F-measure and area under the ROC, place equal importance on precision and recall. However, in our case, correct identification of the 41 nominated leaders (i.e., recall) is more important. Because the 41 nominated users were not the only leaders in the forum, identifying leaders who were not nominated was a helpful undertaking. Although recall is emphasized, we certainly do not want to get a perfect recall of 1 by classifying all users as being leaders. Instead, we use top-*K* recall to evaluate classifiers’ performance. Basically, we have classifiers find *K* users whose probabilities of being a leader are among the top *K* of all users (note that one-class SVM only provides binary decisions. Thus we specify the “outlier” ratio in the classification process, so that the classifier identifies *K* leaders). Then the top-*K* recall metric is the fraction of the 41 nominated leaders that are among the top *K* users. The top-150 recall metrics obtained from the 5 classifiers range from 0.557 to 0.789 (see Table 2 for details).

Table 2. Compare the top-150 recall of 5 individual classifiers on 3 datasets.

	Naïve Bayesian	Log. Regression	Random Forest	One-class SVM	Two-class SVM
Dataset-1	0.789	0.685	0.738	0.557	0.732
Dataset-2	0.796	0.681	0.779	0.561	0.724
Dataset-3	0.781	0.706	0.731	0.781	0.739

3.2. Improving classification with new features and ensemble methods.

We observed that identifying leaders in boards that are not very active is challenging. There are 25 cancer-specific boards, such as “Brain cancer”, and 13 non-cancer-specific ones, such as “Caregivers”. About 90% of all users contributed to only one discussion forum (see Figure 1), so users may be fragmented into multiple sub-communities, each of which could have different leaders. Because sizes of sub-communities vary, leaders in these sub-communities may exhibit different online behaviors. For example, one nominated leader who participated in only one of the most active boards published 5 times more posts and had 3 times higher out-degree than another nominated leader who participated in a much less active board. This is most likely due to the fact that the board with less activity attracts fewer users and posts than highly active ones. While it may take more contribution for one to stand out from the crowd and become a leader in a larger, more active sub-community, it can also be difficult for even a leader to have high in/out-degrees in a sub-community with fewer users.

How can we identify leaders in both large and small sub-communities at the same time? Intuitively, each discussion forum can be considered as a sub-community, from which users’ features are gathered and normalized. Then there are 38 datasets, one for each sub-community (i.e., board) and the potential for leaders within each one. However, the one-to-one mapping between boards and sub-communities is arbitrary and may be disadvantageous to those contributing to multiple boards. The division of user sub-communities should be based on users’ interactions, which are not necessarily separated by the boundaries of boards. When several discussion boards share a similar user base, these users have closer relationship with each other and should be viewed as one sub-community that spans these boards. For instance, a group of more than 3,000 users contributed to both the colorectal and the anal cancer boards, while there were no common users between the stomach and the uterine cancer boards. With this approach, if a leader contributes almost evenly to multiple boards that share a similar user base and gains reputation among this group of users, her contribution will be divided into each board. As a result, her contribution to each individual board may not be high enough for her to be recognized by classifiers as a leader. By contrast, classifiers may favor another user who has made less overall contribution but participates in one board only. Also, while we can get a user’s contribution and semantic features in a board by examining the user’s posts in this board only, the board-based division of sub-communities may lead to inaccurate board-based network features, as it will arbitrarily cut many inter-user ties that are formed through their interactions in other boards. We may fail to find leaders who act as a bridge or broker between two sub-communities that are not well connected.

Instead of arbitrary board-based sub-community division, we emphasized on users’ interactions and proposed two new groups of features to facilitate leader identification in sub-communities, especially smaller or less active ones. The first group, neighborhood-based features, takes an ego-centric approach and focuses on local neighborhood. The idea is that a user whose contribution and centrality are much higher than her network neighbors is more likely to be a leader. We looked at who a user interacts with and measured how the user stands out in her neighborhood. In other words, we measured how a user’s contribution and network centrality differ from those of her neighbors. For each user in Dataset-1, we generated a new neighborhood-based feature from each of the user’s contribution and network features using Equation (1), where N_i is the set of i’s neighbors. As Equation (1) shows, the neighborhood-based feature j for user i ($F'_{i,j}$) is the difference between user i’s value on feature j ($F_{i,j}$) and the average of user i’s neighbors values on feature j. Adding neighborhood-based features to the original Dataset-1, another dataset (*Dataset-2*) is created for classifiers. As Table 2 suggest, adding neighborhood-based features has improved the top-150 recall of 2 of the 5 classifiers.

$$F'_{i,j} = F_{i,j} - \frac{\sum_{k\in N_i} F_{k,j}}{|N_i|}, \tag{1}$$

While neighborhood-based features are simple to calculate and have helped some classifiers find more leaders, disadvantages arise when looking at users who connect to other leaders with high contributions. Consider Figure 2 that depicts two sub-communities in the user network (users' values on feature *f* are listed below their IDs). Users *G* and *I* are leaders and are connected to each other. As a result, *I* has a neighborhood-based feature value of only 7-24/4=1. Meanwhile, user *F*, who connects only to the low-contribution user *B*, gets a neighborhood-based value of 4. As *I* and *F* have the same original feature value 7, the classifier will tend to consider *F*, who has higher neighborhood-based feature value, as a leader instead of *I*, who is actually a "big fish in a small pond" we want to find.

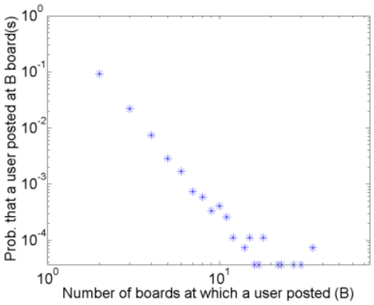

Figure 1. Distribution of the number of discussion boards to which a user contributed.

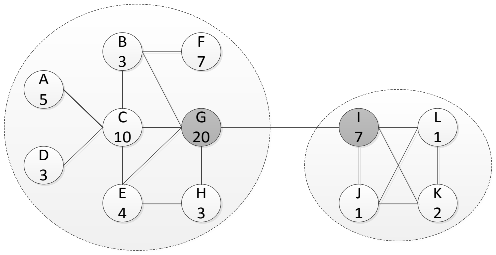

Figure 2. An example of a user network with two sub-communities (leaders are in grey. Users' values on feature *f* are below their IDs).

To complement neighborhood-based features, a second group of cluster-based features were added. The cluster-based features go beyond the local neighborhood and emphasize the sub-community structure in the user network. We applied the modularity-maximization algorithm [10] and clustered the post-reply network among users into 16 sub-communities: the largest has about 5,000 users and the smallest has only 90 users. Similar to neighborhood-based features, the cluster-based contribution or network feature *j* for each user *i* in Dataset-1 was calculated using Equation (2), where C_i is the network cluster to which *i* belongs. Back to the example in Figure 2, cluster-based features of user I do not depend on user *G*'s, but on those of users *J*, *L*, and *K*, who belong to the same cluster as user *I*. Adding new cluster-based features to Dataset-2 produced *Dataset-3*. It turns out the new dataset complements Dataset-2 very well, raising top-150 recalls on 3 classifiers that Dataset-2 fails to improve (see Table 2). In sum, our neighborhood-based and cluster-based features can help all the 5 classifiers improve classification results.

$$F''_{i,j} = F_{i,j} - \frac{\sum_{k\in C_i} F_{k,j}}{|C_i|}, \tag{2}$$

To further improve our classification, we adopted ensemble methods [11]. For each user, a classifier gives a classification result, either as a probability or a binary value to denote whether the user is considered a leader. We then fed each user's five classification results from the five individual classifiers to an ensemble classifier. For individual classifiers, we picked the dataset that enables the classifier to achieve the highest top-150 recall, e.g., Dataset-2 for Na ïve Bayesian, Dataset-3 for Logistic Regression, and so on. Among many ensemble methods, the ensemble classifier based on Random Forest achieves the best performance: an average top-150 recall of 0.854. In addition, when compared to user ranking based on traditional contribution and centrality metrics, our ensemble classifier performs better in terms of top-150 recall.

Table 3. Compare the top-150 recall of various approaches.

Our classifier	Total Posts	Total Active days	Total Degree	Page Rank	Betweenness
0.854	0.732	0.756	0.707	0.731	0.463

4. Conclusion and future work

Our research identifies leaders in an online peer support community for cancer survivors and informal caregivers using classification techniques. At the onset of the project, large-scale labeling was difficult and unavailable. Our classifiers thus face a significant challenge and can only utilize a partial list of subjectively identified leaders from domain experts. We incorporated the structure of the network among users into the extraction of various types of user features, including neighborhood-based and cluster-based features. By taking advantage of user features we derived and combining various individual classifiers, our ensemble classifier achieves reasonable top-150 recall that is higher than those obtained using single-metric-based user ranking methods.

We believe this research could have important implications for both forum users and administrators. Our work allows the forum to encourage users' participation by awarding prestigious status (e.g., virtual badges) to leaders who have been accurately identified. Status as a leader may also be helpful when a user faces conflicting information in the forum or in resolving relationship conflicts. Our research can also help community managers readily identify existing and potential leaders, and reduce labor-intensive monitoring over a long period of time to establish behavior patterns. Early and proactive identification of potential leaders allows community managers to encourage them to assume more leadership when the community loses one of its leaders due to death or other reasons. All online communities face loss of members for a variety of reasons. In addition to these reasons, in an online cancer-related community, users' online activities are often closely related to their health conditions. The loss of few leaders may lead to a less supportive and more pessimistic community. For example, members in CSN are particularly affected by the deteriorating health of community leaders. Absence of leaders from the community due to their medical conditions, depression, or other cancer-related concerns, is acutely felt by community members. However, community managers can encourage other existing or potential leaders to become more active to ameliorate the loss and facilitate the community's recovery.

Our research points us in several directions for further research. A preliminary analysis we conducted, for example, revealed how others' support through online discussions contributes to positive sentimental changes in the posts by the original poster. We believe similar sentiment analysis can refine our understanding of how leaders positively influence other users. We are also interested in using other classification methods, such as boosting, to improve the result. We will continue to work with domain experts to evaluate the outcome of our classification more rigorously and comprehensively. We will build upon these findings to conduct an in-depth investigation of the characteristics and behaviors of leaders in this community.

Selected references

[1] C. Dunkel-Schetter, "Social Support and Cancer: Findings Based on Patient Interviews and Their Implications," *Journal of Social Issues*, vol. 40, pp. 77-98, 1984.

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proceedings of the ICWSM'10*, Washington, DC, 2010, pp. 10-17.

[3] Anonymous, "Calling all patients," *Nature Biotech*, vol. 26, pp. 953-953, 2008.

[4] C. A. Brownstein, J. S. Brownstein, D. S. Williams, P. Wicks, and J. A. Heywood, "The power of social networking in medicine," *Nature Biotech*, vol. 27, pp. 888-890, 2009.

[5] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth international conference on KDD*, Washington, D.C., 2003, pp. 137-146.

[6] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th Intl. Conf. on WWW*, Banff, Alberta, Canada, 2007, pp. 221-230.

[7] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the International Conference on WSDM*, Palo Alto, CA, 2008, pp. 207-218.

[8] C. C. Yang, X. Tang, and B. M. Thuraisingham, "An Analysis of User Influence Ranking Algorithms on Dark Web Forums," in *Workshop on Intelligence and Security Informatics*, Washington, DC., 2010.

[9] A. Bambina, *Online social support: the interplay of social networks and computer-mediated communication*. Youngstown, N.Y.: Cambria Press, 2007.

[10] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577-8582, June 6, 2006 2006.

[11] G. Seni and J. F. Elder, "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 2, pp. 1-126, 2010.