Proceedings of 2014 INFORMS Workshop on Data Mining and Analytics (DMA 2014)
D. Sundaramoorthi, H. Yang, eds.

PREDICTING USER ENGAGEMENT IN ONLINE HEALTH COMMUNITIES BASED ON SOCIAL SUPPORT ACTIVITIES

Xi Wang
Interdisciplinary Graduate Program in Informatics (IGPI)
The University of Iowa
Iowa City, IA
xi-wang-1@uiowa.edu

Kang Zhao
Department of Management Science
The University of Iowa
Iowa City, IA

Nick Street
Department of Management Science
The University of Iowa
Iowa City, IA

Abstract

Online health communities (OHCs) have become a major source of social support for people with health problems. It has been shown that getting engaged in an OHC and interacting online with peers are beneficial to community members. This research studied members' engagement in OHCs from the perspective of social support. Using a case study of an OHC among breast cancer survivors, we first illustrated that members' levels of engagement in an OHC are related to their activities in various types of social support. Then we predicted members' engagement based on their activities in early stages of their online participation, including seeking and receiving (directly or indirectly) different types of social supports, as well as their activity dynamics. The outcome of this study will have implications for the management and design of OHCs to keep users long-term involved.

Keywords: Social Support, Online Health Communities, Engagement and Churn, Predictive Model, Text Mining.

Introduction

During the past decade, more and more people surfed on the Internet for health-related purposes. Different from traditional health-related websites, which only allow members to retrieve information, online health communities (OHCs) provide members a more convenient

environment to share information, interact with peers suffering similar diseases, and as a result better meet their immediate needs for social support. Nowadays, 5% of all Internet users participated in an OHC [1]. Obtaining psychosocial support is one of the key benefits of engagement in OHCs [2,3]. An OHC can help patients adjust to the stress of living with and fighting against their diseases, and also serves as an outlet for members' emotional needs and improves their offline life [4]. A sustainable OHC with an engaged user base will not only help users, operators of the OHC can also better advocate new treatment and healthy life styles in the community. Thus it is important to keep members engaged in an OHC.

Three types of social support have been identified in OHCs according to previous literature: informational support, emotional support, and companionship (a.k.a., network support) [5,6]. Informational support is the transmission of information, suggestions or guidance to the community members [7]. The content of such a post in an OHC is usually related to advice, referrals, education and personal experience with the disease or health problem. Emotional support contains the expression of understanding, encouragement, empathy affection, affirming, validation, sympathy, caring and concern, etc. Such support can help one reduce the levels of stress or anxiety. Companionship or network support consists of chatting, humor, teasing, as well as discussions of offline activities and daily life that are not necessarily related to one's health problems. Examples include sharing jokes, birthday wishes, holiday plans, or online games. Companionship helps to strengthen group members' social network and sense of communities.

Previous studies of OHCs have examined social support among OHC members in different ways. For instance, [8] proposed that members who received more emotional support tended to stay longer, while receiving more informational support does not keep a member engaged. However, none has built predictive model of members' engagement systematically by examining members' seeking, receiving, and provision of various types of social support. Are a member's involvement in and exposure to different types of social support related to her/his long-term engagement in an OHC? Can we predict the "churn" of members based on their early stage behaviors? In this research, we tried to address these two problems.

Detecting Social Support from Texts

We used Breastcancer.org as a case study in this research and designed a web crawler to collect data from its online forum. Our dataset consists of all the public posts and member profile information from October 2002 to August 2013. There are more than 2.8 million posts, including 107,549 initial posts, from 49,552 members. As we mentioned earlier, informational support, emotional support, and companionship are the three major types of social supports in OHCs. To determine whether a post was seeking informational support (SIS), providing informational support (PIS), seeking emotional support (SES), providing emotional support (PES), or simply about companionship (COM), we need to examine the content of the post. We did not differentiate the seeking and provision of companionship, because the nature of companionship is about participation and sharing. By getting involved in activities or discussions about companionship, one is seeking and providing support at the same time. It is also possible that a post could belong to more than one of the categories above. For instance, a post can provide information and emotional support at the same time.

As it is almost impossible to label all 2.8 million posts manually, we used classification algorithms to decide what kind(s) of social support each post contains. We randomly selected 1,333 (54 initial posts and 1,279 comments) out of our dataset. Five human annotators were trained to read each post and decide whether the post is related to one or more categories of social supports (SIS, PIS, SES, PES, COM). Meanwhile, the final label(s) of a post was determined by majority vote among three human annotators with the best quality[1]. The 2nd column in Table 1 shows the number of posts in each category of social support in the annotated dataset. Because there were five categories of social supports and a post may be related to more than one category, we built a classifier for each category. To capture different writing styles or linguistic preferences of members, we extracted various types of features (including basic features, lexical features, sentiment features, and topic features) from each post. For each of the five classifiers, we applied different classification algorithms and picked the best performing algorithm (using 10-fold cross-validation). Overall, our classifiers achieved decent performance with accuracy rates ranging from 0.8 to 0.91 for the five classification tasks. After applying the five classifiers on the remaining of 2.8 million posts, each post received 5 labels, each of which indicated whether the post belong to one of the five social support categories. More details about the classification can be found in [9]. The total numbers of posts in each category are listed in the 3rd column in Table 1.

Table 1: Total numbers of posts in each category of social supports.

Social Support Category	Annotated sample	Whole dataset
Companionship (COM)	435	932,538
Seeking Informational Support (SIS)	96	284,027
Seeking Emotional Support (SES)	22	227,188
Providing Informational Support (PIS)	411	1,034,682
Providing Emotional Support (PES)	249	497,096

Explanatory and Predictive Modeling

We first conducted survival analysis with Cox Proportional-Hazards Model to see if members' involvement in and exposure to different types of social support are related to their engagement in an OHC. The model assessed the importance of different independent variables on the "survival time" it takes for the event of "leaving the OHC" to occur. A member's survival time was measured by the difference between her/his last and first posts in the OHC. We summarized 13 independent variables based on members' activities within the first month of their online activities (shown in Table 2). These variables not only reflected members' own behaviors in seeking and providing social support (e.g., posting to ask or answer questions), but also represented how much social support of various types they received directly and indirectly. In the result, independent variables with hazard ratio lower than 1 contribute positively to the "survival" (i.e., engagement) of members, whereas those with hazard ratio higher than 1 are considered "hazardous" to keep members in this OHC. We removed *TotalPost* and *NumThread* to build the full model, as they are highly correlated with several other independent variables (with correlation coefficients greater than 0.8). According to the results (shown in Table 2), four

[1] To control the quality of human annotations, we also added 10 posts that have been annotated by domain experts. We only accepted results from annotators whose performance on the 10 quality-control posts was among top 3.

independent variables (*SES, SIS, COM, RIS_D*) were statistically significant. Specifically, seeking and receiving informational support directly from the others did not help members' long-term engagement. However, getting involved in more companionship and actively seeking emotional support contributed to their "survival" in this OHC.

While the survival analysis has explained factors related to members' engagement, could we predict whether a member will "churn" from the OHC? Thus we built predictive models based on members' early-stage behaviors. We included all 13 variables in Table 2 as predictive features. At the same time, to measure how members' behaviors changed over time, we calculated slopes for cumulative curves of these variables. For instance, a member who became active quickly will have a high slope for *TotalPost*. Similarly, to detect temporal regularity (daily posting behaviors) in members' behaviors, we computed 13 Entropy values and 13 Temporal Variations [10] of these independent variables. Interestingly, some did not immediately publish a post after registering as a member. We hypothesized that this type of delay may signal a member's intention to join the OHC and thus included the value of such delay as a feature as well.

Table 2: Independent variables in survival analysis

Indep. Variables	Hazard Ratio	Descriptions
TotalPost	-	The total number of posts a member has published
InitPost	0.990	The total number of threads a member initiates
NumThread	-	The number of threads a member contributed to (excluding those initiated by the member)
PES	1.015	The number of a member's posts that provided emotional support
PIS	0.977	The number of a member's posts that provided informational support
SES	0.958***	The number of a member's posts that sought emotional support
SIS	1.055***	The number of a member's posts that sought informational support
COM	0.907***	The number of a member's posts that were related to companionship
RIS_D	1.048*	Direct informational support received--the number of informational support posts a member received after initiating a support-seeking thread.
RES_D	0.993	Direct emotional support received--the number of emotional support posts a member received after initiating a support-seeking thread.
RIS_I	1.040	Indirect informational support received--the number of informational support posts a member was exposed to in threads that she/he did not initiate but contributed to.
RES_I	0.970	Indirect emotional support received--the number of emotional support posts a member was exposed to in threads that she/he did not initiate but contributed to.
RCOM	0.968	Companionship received--the number of companionship posts a member was exposed to in threads that she/he did not initiate but contributed to.
Note: for RIS_I, RES_I, and *RCOM*, we assumed that a member read others' replies that were posted within 7 days before the member's replies.		

*:$p<0.05$, ***: $p<0.001$

First, we tried to predict if a member would eventually churn from this OHC after 4 weeks since her/his first post, according to her/his behaviors in the first 4 weeks. Among 47,581 members in our dataset, we labeled some as "churn" (the positive class), if their online "life span", measured by the time difference between their last and first post, was less than 4 weeks. Others were considered as "staying" (the negative class). We implemented 7 different classification

algorithms to build this predictive model (Naïve Bayes, Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoost and SVM). As the goal was to find those who churned after 4 weeks, we focused the recall for the positive class (using 10 fold cross-validation). We also removed some features to see how they contributed to the classification. In fact, we found that while slopes of cumulative curve of independent variables were helpful to our classification, removing Entropy value and Temporal Variations led to improvement in classification outcome. Among all classifiers, Logistic Regression model has the highest accuracy value of 76.5%. More importantly, the recall for the positive class is 0.911, which is highly desirable.

While the first predictive model can provide a "big picture" on who eventually churn or stay, our second model took a more fine-grained approach and focused on finding those who would churn in their fourth week as an OHC member. We removed from the dataset members whose time spans of activities were less than 21 days. Members, whose time span of activities was between 22 and 28 days, were labeled as the "positive" class (899 members in total). Members who stayed in this OHC for more than 4 weeks were in the "negative" class (18931 members in total). Different classification algorithms were used to predict whether a member would churn in their 4^{th} week. The dataset we dealt with was highly unbalanced--the size of the "negative" class was 20 times more than that of the "positive" class. Thus we under-sampled the majority class in training sets to improve the performance of our predictive model. As Table 3 illustrates, decreasing the ratio of instances between majority class and minority class (class distribution spread) improved the performance of classifiers (evaluated by 10 fold cross-validation). We chose Area under ROC (AUC) to show the general performance of classifiers, and recorded the recall and precision of the "positive" class to evaluate the utility of the predictive model in churn detection. For the under-sampled dataset, even though the Random Forest model did well in distinguishing the two classes (higher AUC), Naïve Bayes achieved the best recall among all algorithms.

Table 3: Performance comparison of predictive models

Algorithm	Original Data			Class distribution (1:1)			Class distribution (1:1) and with feature selection		
	AUC	Recall	Precision	AUC	Recall	Precision	AUC	Recall	Precision
Naïve Bayes	0.654	0.531	0.071	0.657	0.803	0.597	0.664	0.84	0.604
Logistic Regression	0.68	0	0	0.684	0.691	0.626	0.644	0.682	0.582
KNN	0.532	0.068	0.067	0.554	0.56	0.555	0.651	0.657	0.658
Decision Tree	0.653	0.092	0.272	0.645	0.654	0.667	0.74	0.715	0.682
Random Forest	0.679	0.027	0.462	0.742	0.636	0.702	0.748	0.615	0.695
SVM	0.5	0	0	0.633	0.736	0.611	0.644	0.682	0.582

To understand what feature(s) in the dataset positively contribute to the prediction, we conducted feature selection based on the under-sampled dataset. Except Logistic Regression, the performance of all the other classification algorithms improved. This was especially true for the Naïve Bayesian classifier, which can reveal 84% of members who left at their 4^{th} week. Four features were chosen by CFS subset evaluator feature selection algorithm for the Naïve Bayesian model: Slope of cumulative curve, Entropy value, Temporal Variation of *TotalPost* and Slope of

cumulative curve of *NumThread.* Basically, features related to the amount of a members' raw contributions was an important indicator to predict churn. By contrast, features related to social support were not significant contributors to this prediction.

Discussion and Future Work

In this research, we studied members' engagement in OHCs. We first used survival analysis to show that members' involvement in different types of social support is related to their engagement. While many would expect that members join an OHC mainly for informational support, sharing stories from personal daily life and online activities that are not directly related to health turn out to be the key for keeping this community together. Then we built prediction models on OHC members' churn behaviors by leveraging their early-stage activities. Our model was able to predict members who would churn after 4 weeks with high accuracy. Also, we also explored predicting whether a member would churn during a specific time period (4^{th} week in this study). Our research can help OHC managers identify members who may churn so that they can take proactive measures to keep members in the community.

Admittedly, this research is still preliminary in many ways. For future research, we plan to further improve the performance of our predictive models. We also would like to explore member churn prediction for different time periods (e.g., churn in the second week, or the seventh month), so that we can reveal what factors led to churn at different time.

References

1. Chou, W. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P. & Hesse, B. W. Social Media Use in the United States: Implications for Health Communication. J Med Internet Res 11, (2009)
2. Kim, E. et al. The Process and Effect of Supportive Message Expression and Reception in Online Breast Cancer Support Groups. Psycho-Oncology 21, 531–540 (2012)
3. Rodgers, S. & Chen, Q. Internet Community Group Participation: Psychosocial Benefits for Women with Breast Cancer. Journal of Computer-Mediated Communication 10, 00 (2005)
4. Maloney-Krichmar, D. & Preece, J. A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community. ACM Trans. Comput.-Hum. Interact. 12, 201–232 (2005)
5. Bambina, A.: Online Social Support: The Interplay of Social Networks and Computer-Mediated Communication, Youngstown, N.Y.: Cambria Press. (2007)
6. Keating, D. M. Spirituality and Support: A Descriptive Analysis of Online Social Support for Depression. J Relig Health 52, 1014–1028 (2013)
7. Krause, N. Social Support, Stress, and Well-Being Among Older Adults. J Gerontol 41, 512–519 (1986)
8. Wang, Y. -C., Kraut, R. & Levine, J. M. To Stay or Leave? : The Relationship of Emotional and Informational Support To Commitment in Online Health Support Groups. in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 833–842 (2012)

9. Wang, X., Zhao, K. & Street, N. Social Support and User Engagement in Online Health Communities.in Proceedings of International Conference for Smart Health, 97-110 (2014)
10. Zhao, K., and Kumar, A. "Who Blogs What: Understanding Publishing Behaviors of Bloggers," World Wide Web (16: 5-6), 621-644, (2013)