Social Media Analytics of Smoking Cessation Intervention: User Behavior Analysis, Classification, and Prediction

A Thesis
Submitted to the Faculty
of
Drexel University
by
Mi Zhang
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
March 2015
Acknowledgements

I’m grateful to my supervisor, Dr. Christopher C. Yang, whose generous guidance and support made it possible for me to complete my research and this thesis. Dr. Yang provided a flexible environment for me to explore different areas of research, and always gave professional suggestions on my studies and writings. It was very helpful and efficient to discuss and work with Dr. Yang. He listened to my ideas carefully, and offered a lot of valuable advices with great patience.

I also would like to express my great thanks to all the committee members of my prospectus, proposal and final dissertation. During these years, Dr. Weimao Ke, Dr. Jiexun Li, Dr. Kayo Fujimoto, Dr. Erjia Yan and Dr. Paulina Sockolow gave many suggestions on my research that contribute and improve to this thesis.

In addition, I would like to express my gratitude to Dr. Xia Lin for his kindly support, understanding and advices. With his help, I had chances to enroll in many excellent projects, and developed useful skills that greatly benefit my research and future career.

Moreover, many thanks to my collaborators and friends: Xuning Tang, Katherine Chuang, Xuemei Gong, Haodong Yang, Ling Jiang, Jia Huang, Zunyan Xiong, Haozhen Zhao, Haemin Kim, Rachel Magee, Alan Black, Zhan Zhang, and other PhD students at Drexel University.

Finally, I would like to express my endless thanks and gratitude to my parents. Without their love and support, I would never have finished my study and this thesis.
Table of Contents

List of Tables .................................................................................................................................. vi

List of Figures ................................................................................................................................ vii

Abstract ......................................................................................................................................... viii

1.  INTRODUCTION ................................................................................................................... 1

    1.1  Background ...................................................................................................................... 1

    1.2  Related Works .................................................................................................................. 3

    1.2.1  Health Research on Social Media ............................................................................ 3

    1.2.2  Social Support for Health Intervention ................................................................. 5

    1.2.3  Social Network Analysis ....................................................................................... 9

    1.2.4  Data Mining Techniques for Health Social Media ................................................ 11

    1.3  Research Framework ..................................................................................................... 15

2.  A Comparative Study of Smoking Cessation Communities .................................................. 18

    2.1  Problem Definition ......................................................................................................... 18

    2.2  Measurements ................................................................................................................ 19

    2.2.1  Quit Statuses and Quit Stages (RQ 1.1) ................................................................. 19

    2.2.2  Measurement of User Behavior (RQ1.2) ............................................................... 21

    2.3  Experiment and Result ................................................................................................... 23

    2.3.1  Comparison of User Quit Stages (RQ1.1) ............................................................. 23

    2.3.2  Comparison of User Behaviors (RQ1.2) ............................................................... 23

    2.3.3  User Behavior of Different Quit Stages (RQ1.3) ................................................... 27
2.4 Discussions ................................................................................................................... 28

2.5 Conclusion ..................................................................................................................... 31

3. Social Support Exchange Patterns and User Behaviors ......................................................... 32

3.1 Problem Definitions ....................................................................................................... 32

3.2 Discussion Themes and Types of Social Support Exchange (RQ2.1) ........................... 33

3.2.1 Coding Scheme ...................................................................................................... 34

3.2.2 Result of Content Analysis .................................................................................... 38

3.2.3 Discussions ............................................................................................................ 42

3.3 Patterns and User Behaviors of Social Support Exchange (RQ2.2) ................. 44

3.3.1 Measurement of Exchange Patterns ...................................................................... 44

3.3.2 Analyzing Result of Exchange Patterns ................................................................. 48

3.4 User Interactions of Social Support Exchange (RQ 2.3) .......................................... 49

3.4.1 Network Exposure Model ...................................................................................... 50

3.4.2 Blockmodel Based on Quit Stages ......................................................................... 52

3.5 Conclusion ..................................................................................................................... 54

4. Thread Classification and User Recommendation ................................................................. 56

4.1 Classification of Forum Messages ............................................................................... 57

4.1.1 Problem Description .............................................................................................. 57

4.1.2 Approaches ............................................................................................................ 61

4.1.3 Results and Discussions ......................................................................................... 65

4.1.4 Conclusion ............................................................................................................. 72
List of Tables

Table 2.1 User Distributions Through Quit Stages ................................................................. 23
Table 2.2 Statistics of User Average Response Time of QuitNet Forum and QuitNet Facebook . 24
Table 2.3 Degree Centrality and Average Response Time in Each Group of Quit Status of QuitNet Facebook ................................................................. 28
Table 2.4 Result summary of the comparative study of QuitNet Forum and QuitNet Facebook . 31
Table 3.1 Number of Messages in Each Category ............................................................... 39
Table 3.2 Patterns and User Behaviors of Social Support Exchange .................................. 48
Table 3.3 Density Matrix and Block Matrix of Informational Support Network ................. 53
Table 3.4 Density Matrix and Block Matrix of Nurturant Support Network ..................... 53
Table 4.1 Health Feature Sets ......................................................................................... 60
List of Figures

Figure 1.1 Research Flow .............................................................................................................. 16

Figure 2.1 User Distribution Across Time Groups of QuitNet Forum and QuitNet Facebook .... 25

Figure 2.2 Social Network of QuitNet Forum ............................................................................. 26

Figure 2.3 Social Network of QuitNet Facebook .......................................................................... 27

Figure 3.1 Messages in Main Categories ....................................................................................... 40

Figure 3.2 Types of Informational Support Offered in Posts and Comments ............................... 41

Figure 3.3 Types of Nurturant Support Offered in Posts and Comment ....................................... 41

Figure 3.4 Description Types of Requesting Support in Posts and Comments ............................. 42

Figure 3.5 Types of Receiving Support in Comments .................................................................. 42

Figure 3.6 Exchange Patterns and User Behaviors of Social Support Exchange ....................... 47

Figure 3.7 Network Exposure of Users ....................................................................................... 51

Figure 3.8 Reduced Graph of Informational Support Network ..................................................... 53

Figure 3.9 Reduced Graph of Nurturant Support Network .......................................................... 53

Figure 4.1 Discussion Scenarios on QuitNet Forum ..................................................................... 57

Figure 4.2 Main Process to Develop Classifiers ......................................................................... 62

Figure 4.3 Results of Post Classification of Intentions ................................................................ 67

Figure 4.4 Results of Post Classification of Social Support Types .............................................. 69

Figure 4.5 Results of Comment Classification of Intentions ....................................................... 70

Figure 4.6 Results of Comment Classification of Social Support Types ..................................... 72

Figure 4.7 Improving CF with Classification Results ................................................................. 81

Figure 4.8 Improving Hybrid Method with Classification Results ........................................... 83

Figure 4.9 Improving basic CF with Users’ Quit Stages ........................................................... 86

Figure 4.10 Improving Hybrid Method with User Quit Stages and Classification Results ........ 88
Abstract

Social Media Analytics of Smoking Cessation Intervention: User Behavior Analysis, Classification, and Prediction
Mi Zhang
Supervisor: Christopher C. Yang, Ph.D.

Tobacco use causes a large number of diseases and deaths in the United States. Traditional intervention programs are based on face-to-face consulting, and social support is offered to help smoking quitters control stress and achieve better intervention outcomes. However, the scalability of these traditional intervention programs is limited by time and location. With the development of Web 2.0, many intervention programs of smoking cessation are developed online to reach a wider population. QuitNet is a popular website for smoking cessation that provides different services to help users quit smoking. It builds communities on different social media for people to discuss issues of smoking cessation and provide social support for each other. In this dissertation, we develop a comprehensive study to understand user behavior and their discussion interactions in online communities of smoking cessation. We compare user features and behaviors on different social media channels, analyze user interactions from the perspective of social support exchange, and apply data mining techniques to analyze discussion content and recommend threads for users.

Health communities are developed on different types of social media. For example, QuitNet has Web forums on its own Web site while it also has its appearance on Facebook. The user participation may vary on different social media platforms. Users may also behave differently depending on the functions and design of the social media platforms. So, as the first step in this dissertation, we carry out a preliminary study to compare smoking cessation communities on different social media channels. We analyze user characteristics and behaviors in QuitNet Forum and QuitNet Facebook with statistical analysis and social network analysis. It is found that most
users of QuitNet Forum are early smoking quitters, and they participate in discussions more actively than users of QuitNet Facebook. However, users of QuitNet Facebook have a wider spectrum of quitting statuses and interaction behaviors.

Second, we are interested in user behaviors and how they exchange social support in online communities. Social support is “an exchange of resources between two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient”. As QuitNet Forum attracts much more active users than QuitNet Facebook, it provides a better platform for our research purpose. So, we focus on QuitNet Forum, developing a classification scheme through qualitative analysis to categorize discussion topics and types of social support on the forum. Patterns of user behaviors are defined and identified. Social networks are built to analyze user interactions of social support exchange. It is found that users at different quit stages have different behaviors to exchange social support, and different types of social support flow between users at different quit stages.

Discussion topics, user behaviors and patterns of social support exchanges are thoroughly analyzed. However, due to a huge amount of information on QuitNet Forum, it is difficult for users to find proper topics or peers to discuss or interact with. It would be helpful if we could apply machine learning techniques to understand user generated information in online health communities, and recommend discussion topics to users to participate in. We develop classifiers to categorize posts and comments on QuitNet Forum in terms of user intentions and social support types. User behaviors and patterns are used to help developing various feature sets. Then, we develop recommendation techniques to recommend threads for users to participate in. Based on traditional Collaborative Filtering and content-based approaches, we integrate classification results and user quit stages to develop recommendation systems. The experiments show that
integrating classification results or user health statuses can achieve the best recommendation results with different percentages of unknown data.

In this dissertation, we implement all-sided studies for online smoking cessation communities, including comprehensive analytics and applications. The proposed frameworks and approaches could be applied to other health communities. In the future, we will apply more analytics and techniques to a larger data set, and develop user-end applications to serve and improve online health intervention programs and communities.
1. INTRODUCTION

1.1 Background

Smoking and tobacco use are acquired behaviors and the most preventable cause of death in the U.S. (Agaku, King, & Dube, 2012). In 2011, an estimated 19.0% (43.8 million) of U.S. adults smoked cigarettes (Agaku et al., 2012). Lung cancer, ischemic heart disease and chronic obstructive pulmonary disease are highly related to smoking. At least 30% of cancer deaths are associated with smoking, and 440,000 U.S. citizens die from tobacco use every year (NIH, 2012). To reduce the burden of tobacco use, many intervention programs have been developed for smoking cessation. Traditional intervention programs are based on face-to-face consulting. Although they are effective to help people quit smoking (Russell, Wilson, Taylor, & Baker, 1979), the scalability of these programs is limited.

The Internet provides a widely accessible communication channel, which could reach a large number of people by easily overcoming geographical or time limitations. With this advantage, more and more intervention programs of smoking cessation are developed on the Internet. QuitNet1 is a popular online intervention program of smoking cessation founded in 1995. It provides different services to help users quit smoking, including interactive diagnostic tools, quit planning tools, online expert counseling, online communities and one-to-one messaging (An et al., 2008). There are 11 forums on QuitNet website, which are traditional communities for discussions among smokers, smoking quitters, medical professionals and researchers. Registered users of QuitNet website can start new threads (write posts) on the forums. They can also make comments and participate in discussions of other users’ threads. In recent years, QuitNet has also created a public page on Facebook2 to attract broader population. The admin of QuitNet

1 http://www.quitnet.com/qnhomepage.aspx
2 https://www.facebook.com/QuitNet
Facebook posts messages on the public page every day, inviting discussions. Facebook users could browse QuitNet Facebook and participate in the discussions. Users who “like” the public page could receive messages in their Facebook “news feeds” and get updated information in a timely manner. QuitNet Forum and QuitNet Facebook represent two different types of online communities for healthcare topics. Although both of them are supervised by the same organization – QuitNet, we do not know whether they attract users with similar characteristics, or whether users behave similarly in the two different communities.

In health intervention programs, social support plays an important role in helping people to achieve better intervention outcomes. Social support is “an exchange of resources between two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient” (Shumaker & Brownell, 1984). Online communities of smoking cessation provide ideal platforms for smoking quitters to interact and exchange social support with each other. Different types of social support are offered in health intervention programs (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Eichhorn, 2008), which can benefit support receivers in different aspects. Identifying types of exchanged social support could help us understand user features and interactions in online health intervention programs.

Wactlar et al. (Wactlar, Pavel, & Barkis, 2011) discussed different ways that computer science research helps resolving healthcare problems. One of the important ways is empowering patients to play a substantial role in their own health and treatment in cyber environment. Health communities on different social media, like QuitNet forum, have the potential to provide such function, because patients can be fully involved in the discussions and interactions regardless of time and locations. However, due to the large amount and unstructured content of data on the forum or other health communities, it is difficult for users to identify relevant topics or peers to discuss or interact with. It would be helpful if we could recommend interesting topics or predict potential users for online health communities.
Nowadays, many communities are built on different social media sites for smoking cessation and other health issues. A large amount of user generated content can be found on healthcare social media, which provides knowledge and resource to investigate user features and actions. However, few studies investigated user interactions and their discussions on different social media sites. To our knowledge, there is no comprehensive framework to investigate user interactions on health social media. In this dissertation, we make a systematic, deep and complete study on user features, behaviors and interactions of different health social media. We also apply data mining techniques to analyze the discussion content from the perspective of social support exchange, and develop recommender systems to recommend proper topics for users.

1.2 Related Works

1.2.1 Health Research on Social Media

With the plethora of social media websites available, many of them are adapted to healthcare peer-to-peer usage, which introduces the term “Health 2.0” (Hughes, Joshi, & Wareham, 2008). Patient-oriented care is taking place of traditional disease-oriented care. Patients are playing an ever-greater role based on a principle of autonomy (Sacristán, 2013). Health 2.0 shows a variety of features like social networking, participation, apomediation, collaboration and openness (Eysenbach, 2008) (Van De Belt, Engelen, Berben, & Schoonhoven, 2012). Communication becomes an important factor to promote Heath 2.0 in current Internet environment. Generally, health support communities can be divided into two categories, some of which are health focused and other that are general audience. Health specific online communities include MedHelp and PatientsLikeMe, where user group forums are developed on these sites for specific health conditions or problems. Other online health communities are built on popular social networking sites, including Facebook, Twitter, LinkedIn, Blogs, YouTube, Second Life, etc (Backman et al.). These social network sites are developed for broader social uses, not necessarily limited to health interventions. Currently, 67% of Internet users use social networking sites like Facebook and
Twitter (Duggan & Brenner, 2013). Health communities based on these websites have the potential to reach a wider target population regardless of socioeconomic and health-related characteristics (Chou, Hunt, Beckjord, Moser, & Hesse, 2009). Fox et al. reported that “the social life of health information is robust”, with the fact that 52% online health inquired involved interaction with others (Fox & Jones, 2009).

QuitNet Forum and QuitNet Facebook represent two different types of online communities for smoking cessation intervention. A lot of studies have demonstrated that online intervention programs, especially online forums for social support, are effective in helping smokers quit, and well-designed online services increase the ratio of users who quit in a target group (An et al., 2008; Cobb, Graham, Papandonatos, & Abrams, 2005; Graham, Cobb, L., Still, & Young, 2007; Shahab & McEwen, 2009). While online intervention programs are targeted towards smokers of either gender, previously documented research show that a majority of voluntary online intervention participants as female, older, and non-smokers (Cobb & Graham, 2006; Cobb, Graham, & Abrams, 2010). Web-based, tailored, and interactive smoking cessation interventions are more effective (Shahab & McEwen, 2009). There are four key areas of online smoking cessation resources: cessation, prevention, social support and professional development and training (Norman, McIntosh, Selby, & Eysenbach, 2008). Compared to earlier programs, present online smoking cessation interventions focus more on providing advice to quit, practical counseling, and enhancing motivation to quit smoking through personal relevance and risks (Bock, Graham, Whiteley, & Stoddard, 2008).

In previous research, data of user behavior are collected through surveys, experiments or Internet contents. Survey (Cobb & Graham, 2006; Cobb et al., 2005; Graham et al., 2007) and randomized controlled trial are the most common methods to acquire data (Japuntich et al., 2006; Pike, Rabius, McAlister, & Geiger, 2007; Shahab & McEwen, 2009). Surveys can acquire user information about their characteristics, whereas controlled trials can explore and compare user behaviors
under different situations. However, both of the two methods are time-consuming and only refer to a small proportion of users. Another method for data collection is to extract information directly from website contents of online communities, including posts and comments published by users (Burri, Baujard, & Etter, 2006; Cobb et al., 2010; Selby, Mierlo, Voci, Parent, & Cumminghm, 2010). This method could get raw data of all user behavior during a certain period of time. However, this kind of information is not thoroughly analyzed or fully utilized in most current research. Cobb and Abrams (2010) extracted a social network based on data of QuitNet, including information of exchanged messages, buddy list and posts on forums. They studied characteristics of the whole social network as well as five subgroups of it. It was the first study which applied formal social network analysis to online smoking cessation programs.

Bishop (2007) proposed an ecological cognition framework for understanding why members of online communities either participate or do not participate. In this framework, an actor’s desire to participate in discussions is limited by his/her goals, plans, values, beliefs and interests. So, to encourage users participate in discussions in online communities, it is important to understand users’ goals, values and interests, and lead them to proper information and threads. From another level of this ecological cognition framework, an actor will participate in communities based on how they perceive their environment (Bishop, 2007). For patients or health consumers, it is important to build an environment for them to reach the peers who have similar conditions or experiences with them, and who could provide useful information and proper social support for them. Thus, to encourage users to participate in and retain in a health community, we can recommend information to meet users’ interests, and lead users to participate in threads where they could connect with proper peers.

1.2.2 Social Support for Health Intervention

(1) Social Support Introductions
In health intervention programs, social support plays an important role in helping people to achieve better intervention outcomes. Social support is “an exchange of resources between two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient” (Shumaker & Brownell, 1984). In traditional health intervention programs, social support is usually offered to patients through face-to-face communication. It can directly benefit physical and psychological outcomes of health interventions (Wright & Bell, 2003). Social support can also shield patients from the negative effects of stress, and help patients fight against and control stress (Lichtenstein, Glasgowb, & Abrams, 1986; Wright & Bell, 2003). In traditional intervention programs of smoking cessation, it was found that social support was positively correlated with better outcomes, including successful cessation and maintenance (Lichtenstein et al., 1986). Enlisting a support person can increase the success quitting ratio (Pirie, Rooney, Pechacek, Lando, & Schmid, 1997). Traditionally, smoking quitters can only get limited social support through face-to-face communication in pairs or small groups at specific times and locations. Although spouses or significant others have the potential to provide social support to smoking quitters, it is difficult to change their behavior to offer effective support, because smoking cessation may not be a topic of their concern in daily life (Lichtenstein et al., 1986).

With the advent of the “Health 2.0” age (Hughes et al., 2008), large numbers of online communities have been developed for discussions of health issues, and social support is widely exchanged online, which has certain advantages. Comparing to traditional social support, geographic and transportation barriers are absent on the Internet. Also, the number of participants can be unlimited in online virtual communities (White & Dorman, 2001). Online comforting communications can easily establish a safe environment in which participants can discuss personal issues with anonymity (Caplan & Turner, 2007). Burleson and Goldsmith proposed three conditions for effective comforting conversations. First, participants must be willing to enter into the conversation; second, talks must focus on relieving feelings of upsetting matter; and third, the
distressing matter must be discussed in a way that facilitates reappraisals. Caplan and Turner demonstrated in their study that the unique features of online social interactions could facilitate the three conditions for effective comforting communication (Caplan & Turner, 2007). According to a longitude study of randomized clinical trial, Murray et al. (Murray, Johnston, Dolce, Lee, & O'Hara, 1995) found that smoking quitters supported by ex-smokers were more likely to quit than those supported by smokers. To achieve better outcome, supporting people should be also included in the same intervention programs. In online intervention programs of smoking cessation, a lot of ex-smokers participate in the communications to offer social support for smoking quitters. So, online community is an ideal channel for social-support exchange.

(2) Types of Social Support

There are different types of social support identified from online intervention programs of different health problems, like stress control (Cutrona & Suhr, 1992), weight loss communities (Hwang et al., 2011), HIV/AIDS (Mo & Coulson., 2008), and alcoholism community (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011). Usually, the types of social support are identified by content analysis of discussions. The aim of content analysis is to “attain a condensed and broad description of the phenomenon”, and the outcome of the analysis is “concepts or categories describing the phenomenon” (Elo & Kyngäs, 2008). There are two types of content analysis, which are inductive content analysis and deductive content analysis. Inductive analysis uses open coding to write notes and headings of content texts, and list hierarchical categories. Deductive content analysis is used applying already-existed categories or codes in a new context. It involves testing categories, concepts, models or hypotheses.

Generally, most types of social support can be divided into two categories: action-facilitating support and nurturant support (Cutrona & Suhr, 1992). Action-facilitating support is intended to assist patients to solve or eliminate the problems. Nurturant support encompasses efforts to
comfort or console, without direct efforts to solve the problems. Finn (1999) observed 718 posts of an online bulletin of disability, and found that they could be divided into two realms: task-oriented messages and socioemotional messages, which correspond with the two categories of social support.

Based on content analysis of posts of alcoholism community on MedHelp, Chuang and Yang (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011) developed a comprehensive code scheme for types of social support. They developed three main types of social support, including informational support, nurturant support and instrumental support. Informational support is similar to the category of “action-facilitating support” mentioned above. It proposes problem solutions for patients. Informational support includes advice, fact, personal experience, opinion and referral. Nurturant support includes esteem, network and emotional support. Instrumental support provides materials, financial aid or services to assist patients.

(3) User Exchange Behavior of Social Support and Information

In online health communities, people adopt different strategies to request social support. North identified five strategies, including self-deprecating comments, shared experiences, requests for information, statements of personal success and statements of extreme behavior (Eichhorn, 2008).

Social support could be regarded as a specific format of information. To study information seeking behavior, Savolainen (1995) defined two types of information, which are practical information and orienting information. When seeking for practical information, people look for answers to discrete and specific information needs. When seeking for orienting information, people do not have specific questions. They put themselves around the “information neighborhood”, in which there is information related to their ongoing interests and concerns (Burnett, 2000). Practical and orienting information are two main types of information describing
information exchange patterns. Savolainen (1995) pointed out that seeking practically effective
information is “active”, and seeking orienting information is “passive”.

1.2.3 Social Network Analysis
In our work, social network analysis is used to model user behaviors on different social media
channels. Social network analysis has been widely applied in communication studies. A social
network is defined as a set of $g$ actors and a collection of $r$ relations which specify how actors are
relationally tied together (Wasserman, Robins, & Steinley, 2007). Social network could simulate
and describe the macro structure of actors’ connections or interactions. A series of measurements
and technologies are developed for social network analysis. Two properties of relations are
important in social network analysis: whether the relation is directional or nondirectional, and
whether it is dichotomous or valued (Wasserman & Faust, 1994).

(1) Network Exposure Model
Network exposure models can be used to measure the extent to which an actor is exposed to
neighbors with a specific behavioral attribute in a social network (Valente, 1995, 2005).
Generally, the network exposure $E$ of an actor $i$ in a social network is defined as:

$$E_i = \frac{\sum w_{ij} y_j}{\sum w_{ij}}$$  (1.1)

where $w$ is the social network weight matrix, $y$ is a vector of actors’ attributes, and $j$ is a neighbor
of actor $i$. $E_i$ considers the attributes of all neighbors of $i$. Usually, the weight matrix $w$ can be
built on different factors, including relation, position and centrality (Valente, 2005).

Network exposure has been applied in many studies of public health. Usually, it is calculated for
a behavior attribute like smoking, drinking, syringe sharing, etc. (Fujimoto, Unger, & Valente,
2012; Gyarmathy & Neaigus, 2006). Each element in the attribute vector $y$ has a binary value of 1
or 0, indicating whether the corresponding actor had this behavior. For example, Gyarmathy and
Neaigus (2006) analyzed a social network of injecting drug users. For every actor in the network, the personal network exposures for sharing cookers, sharing filters, receptive syringe sharing, distributive syringe sharing, and backloading were respectively calculated using closeness matrix \( w \). For each of the equipment sharing behaviors, the correlation between the actor’s own behavior and his/her personal network exposure was investigated.

(2) Blockmodel

For a social network, a blockmodel partitions actors into \( N \) discrete subsets called positions. There are ties within or between positions to represent their relations (Wasserman & Faust, 1994). For a social network, the blockmodel is represented by a \( N \times N \) matrix \( B = \{b_{kl}\} \), with entries \( b_{kl} \) equaling 1 or 0, indicating the presence or absence of a tie from position \( B_k \) to \( B_l \). Each entry \( b_{kl} \) represents a block in the blockmodel. It is an oneblock when \( b_{kl}=1 \), and is a zeroblock when \( b_{kl}=0 \).

To decide the value of \( b_{kl} \), we calculate the density \( \Delta_{kl} \) in this block (Wasserman & Faust, 1994) using the following formula:

\[
\Delta_{kl} = \begin{cases} 
\frac{\sum_{i \in B_k} \sum_{j \in B_l} w_{ij}}{g_{B_k} \times g_{B_l}}, & \text{for } k \neq l \\
\frac{\sum_{i \in B_k} \sum_{j \in B_k} w_{ij}}{g_{B_k} \times (g_{B_k} - 1)}, & \text{for } k = l 
\end{cases}
\]

where \( w_{ij} \) is the value of tie from actor \( i \) to actor \( j \) in the social network, \( g_{B_k} \) is the number of actors in the position \( B_k \), and \( g_{B_l} \) is the number of actors in the position \( B_l \). \( \Delta_{kl} \) denotes the average value of ties from actors in \( B_k \) to actors in \( B_l \).

There are different criteria to decide oneblock and zeroblock. One common criterion is a density criterion (Wasserman & Faust, 1994), which is defined as:

\[
b_{kl} = \begin{cases} 
0, & \text{if } \Delta < a \\
1, & \text{if } \Delta \geq a 
\end{cases}
\]

(1.3)
A reduced graph could be used to present ties within and between positions. In the reduced graph, positions are represented as nodes and ties between positions are represented as arcs. There is an arc between positions of $B_k$ and $B_l$ if $b_{kl}=1$, and there is no arc between $B_k$ and $B_l$ if $b_{kl}=0$.

1.2.4 Data Mining Techniques for Health Social Media

Users generate a large amount of text on different social media. It provides knowledge and resources to investigate user features and actions. However, the user generated content is usually huge and unstructured. We need to apply data mining techniques to analyze and understand the content. Different data mining techniques are widely used and applied to analyze social media data, including association mining, classification, clustering, graph mining and so on. In our work, techniques of classification and recommendation are applied and developed.

(1) Classification of User Generated Content

Classification is a supervised learning technique that is used in many different intelligent systems. Kotsiantis summarizes four different types of classification algorithms, including logic based algorithms, perceptron-based techniques, statistical learning algorithms, and support vector machines (Kotsiantis, 2007). Decision tree is the representative of logic based algorithms, which leans a set of rules from the training data. Perceptron-based techniques usually run algorithms repeatedly through the training set until finding a prediction model that fits all the training data. Neural network is a widely-used perceptron algorithm. Naïve Bayes and Bayesian Networks are statistical learning algorithms. They develop probability models to predict for each class. Vector support machines develop hyperplanes to separate data points of different classes, which maximizes the margins of hyperplanes. It is proved accurate for many different classification tasks. Different classification algorithms vary in time and space efficiencies. Usually, Naïve Bayes algorithms are much faster than neural networks and vector support machines. They also require little space during the training and classification processes.
In online forums, questioning and answering are the general actions and user interactions that could be extracted from discussion content. For question/answer detections, different classification methods are used to analyze user generated content. Kim et al. (J. Kim, Chern, Feng, Shaw, & Hovy, 2006) classified threads in a student discussion board. They considered speech act patterns and proposed methods to detect conversation focus. However, their classification is implemented manually on a small dataset. Antonelli and Sapino (Antonelli & Sapino, 2005) developed a rule-based classifier to identify the relations of different postings. In some studies (Hong & Davison, 2009) (S. N. Kim, Wang, & Baldwin, 2010), different classification techniques are applied and compared for question/answer detection, including algorithms of maximum entropy, SVM and CRF. A variety of features were also selected and compared. Classification method is also used to evaluate the quality of post content (J. Huang, Zhou, & Yang, 2007; Weimer & Gurevych, 2007). To analyze the completeness, solvedness, spam and problem types of threads in a Linux user forum, Baldwin et al. (Baldwin, Martinez, & Penman, 2007) extracted text features from different positions of threads, and used different classification algorithms for thread classification.

In health area, classification and other data mining techniques are applied to analyze structured biological data, such as attributes of cells, genes, proteins, etc. (Saeys, Inza, & Larranaga, 2007). Academic articles in Medline/PubMed are classified based on text features (Cohen & Hersh, 2005). For social media analysis of healthcare, text mining and social network analysis are used to propagate infectious diseases with hospital records, predict pandemic increase with Twitter data, model hospital structure network, or analyze health social network for some websites (Wegrzyn-Wolska, Bougueroua, & Dziczkowski, 2011). Some studies applied text mining techniques to analyze the post content in online health forums and groups (S. Kim, 2009; S. Kim, Pinkerton, & Ganesh, 2012; Oh & Park, 2013). However, these studies did not analyze the complete text of user generated content. They used medical vocabularies (e.g. Medical Subject
Headings) to extract concepts and terms for analysis. As a result, some important user interactions, like social support exchange, cannot be indicated from the vocabulary-based text mining.

(2) Recommendation Techniques

Collaborative filtering and content-based approach are basic methods used in recommender systems (Adomavicius, 2005). Collaborative filtering is based on the assumption that users with similar preferences are likely to rate items similarly, while content-based approaches assume that items with similar features will be rated similarly (Schafer, Frankowski, Herlocker, & Sen, 2007). There are hybrid systems combining these two technologies for recommendation (Burke, 2007; Grujić, 2008; Melville, Mooney, & Nagarajan, 2002). However, although recommender systems have been widely used by different websites for product recommendation, they are seldom deployed in online forums or communities in practice.

There are two categories of collaborative filtering algorithms: neighborhood-based algorithms and model-based algorithms (Breese, Heckerman, & Kadie, 1998). Neighborhood-based algorithms have the advantages of simplicity, justifiability, efficiency and stability (Desrosiers & Karypis, 2011). They extract user-item relations and construct user similarity matrix or item similarity matrix to make predictions (Z. Huang, 2007; Z. Huang, Li, & Chen, 2005). In a forum, the number of threads continues to increase and the popularity of a thread is temporal in a rather short period. As a result, item similarity matrix is not the best similarity measurement in predicting user participation of forums. In this proposal, we construct user similarity matrix for collaborative filtering. For two different users, the similarity between them is calculated according to the same items recommended or rated by them. The user similarity can be calculated by different methods, including correlation-based similarity, cosine-based similarity, Jaccard’s similarity, etc (Z. Huang, 2007; Z. Huang et al., 2005). Based on the user similarity matrix,
different methods are used to predict users of an item, of which the most prevalent is k-nearest neighbors (Desrosiers & Karypis, 2011; Schafer et al., 2007).

User similarity matrix is usually sparse. Many users in the matrix do not have direct neighbors. Thus, neighborhood-based algorithms often suffer from the cold start problem for prediction (Desrosiers & Karypis, 2011; Schafer et al., 2007). To address this problem, some studies select neighbors based on other sources like individuals’ social network (Zheng, Wilkinson, & Provost, 2008). Improved methods are also proposed to calculate user similarity and build user matrix (Ahn, 2007; Fouss, Pirotte, Renders, & Saerens, 2006; Z. Huang, Chen, & Zeng, 2004).

Content-based approach is also used in many recommender systems (Melville et al., 2002; Zhu, Greiner, & Häubl, 2003). Each user and item is represented as a content-based profile, and “similar” items are recommended to a user by calculating their profile similarities. The content information could be grabbed from different sources, including local information like item descriptions, web pages and web logs (Grujić, 2008; Liu, Zhou, Wang, Zhang, & Guo, 2010; Melville et al., 2002; Zhu et al., 2003), and global information like related content from Wikipedia (Katz, Ofek, Shapira, Rokach, & Shani, 2011). In traditional content-based recommendation, items and users are presented as term vectors, which are usually sparse. The efficiency of matching users with items is quite low by calculating with the sparse term vectors.

Some studies applied recommendation technology in online communities to predict discussion topics and participants. To predict users in a community, Fung et al. (Fung, Li, & Cheung, 2007) adopted collaborative filtering approach, and analyzed user-thread relations with Zipf’s law and tf-idf. To recommend twitter users to follow, Hannon built user similarity matrix for collaborative filtering recommendation, and tried to boost the prediction by detecting user interests and matching their interests with tweets’ content (Hannon, Bennett, & Smyth, 2010). Yano et al. introduced topic models to predict participants of blog posts (Yano, Cohen, & Smith, 2009). They
combined topic models of LinkLDA and CommentLDA to generate blog posts. In our previous work, we proposed another topic model, UTD (User Interest and Topic Detection Model) (Tang, Yang, & Zhang, 2012; Tang, Zhang, & Yang, 2012), to detect user interests and thread topics in online communities, which was used as a content-based approach for user prediction.

1.3 Research Framework

In this dissertation, we develop a comprehensive study to understand user behavior and their discussion interactions in online communities of smoking cessation. We compare user features and behaviors on different social media channels, analyze user interactions from the perspective of social support exchange, and apply data mining techniques to analyze discussion content and recommend topics for users. Our work includes three parts as described below.

(1) A comparative study of online smoking cessation communities. As described above, QuitNet Forum and QuitNet Facebook represent two different types of online communities for smoking cessation intervention. User participation and behavior may be different on different social media platforms depending on the functions and interface design. In this work, we conduct a preliminary study to analyze user characteristics and behaviors in QuitNet Forum and QuitNet Facebook, which can help us to understand the features of different social media (Chapter 2).

(2) Study of social support exchange patterns and user behaviors. Comparison between different social media brings a general description about user characteristics and behaviors. However, to fully understand user behavior and how they exchange social support in online communities, we need to implement a more thorough research. Different types of social support are offered in health intervention programs (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Eichhorn, 2008), which can benefit support receivers in different aspects. In this dissertation, we develop a coding scheme to categorize messages in QuitNet forum from the perspective of social support exchange.
Moreover, motivated by the theory of user information behavior, we analyze exchange patterns and user behaviors of social support. An in-depth study is designed and carried out with content analysis, statistical analysis and social network analysis (Chapter 3). This research can help us understand what are discussed in smoking cessation communities. Also, it provides a broad overall view for user interactions of social support exchange.

(3) Thread classification and user recommendation for QuitNet forum. Discussion topics, user behaviors and social support exchange patterns should be scaled to larger data for different applications. To understand discussion content on QuitNet forum from a large scale, we develop classifiers to categorize posts and comments on QuitNet Forum from the perspectives of user intentions and social support types. Also, a recommendation system is built to recommend threads for users, which can help users to find proper topics and peers (Chapter 4). Based on Collaborative Filtering and content-based approaches, we integrate classification results and user quit stages to develop recommendation systems. This study utilizes knowledge that we learned from user interactions, and applies data mining techniques to process a large number of messages. This research could be applied to improve services of online health communities.

<table>
<thead>
<tr>
<th>Preliminary Study</th>
<th>User Behavior and Interaction Pattern</th>
<th>Classification and Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goals</strong>: compare user health features and behaviors in communities of different social media platforms</td>
<td><strong>Goals</strong>: analyze discussion content, investigate user behaviors and exchange patterns of social support in a selected community</td>
<td><strong>Goals</strong>: develop machine learning applications, including classification and recommendation</td>
</tr>
<tr>
<td><strong>Approaches</strong>: statistical analysis and social network analysis</td>
<td><strong>Approaches</strong>: content analysis and social network analysis</td>
<td><strong>Approaches</strong>: machine learning techniques</td>
</tr>
</tbody>
</table>

Figure 1.1 Research Flow
The three parts of study are in a flow of preliminary study, in-depth study and applied study. The
descriptions for each part is shown in Figure 1.1Figure 2.1. The rest of this dissertation will
describe each part of the study in a chapter. In each chapter, the research questions are proposed,
the data set and methodologies are described, and the experiment results are discussed and
concluded. In the last chapter, we will summarize the whole study.
2. A Comparative Study of Smoking Cessation Communities

QuitNet Forum and QuitNet Facebook are smoking cessation intervention communities based on different social media channels. They have their own advantages and shortcomings. As a professional online smoking cessation program, QuitNet Forum is better organized and managed. Some experts are invited to take part in QuitNet Forum regularly so that they could provide professional suggestions for users. In addition, users of QuitNet Forum are much more motivated so they perform more active than users of QuitNet Facebook. Their discussion topics are usually more targeted. QuitNet Facebook provides a public page on Facebook, which is a convenient way for people to communicate on the topic of smoking cessation. Once a Facebook user “like” the QuitNet Facebook page, s/he will receive all its discussion threads. Any Facebook user can participate in Quitnet discussion easily while they are interacting with other Facebook friends as their regular online social activities. Facebook has been a popular social media platform for information dissemination and consumer opinion collection. In contrary, QuitNet Forum requires users to register a specific account with a password and go to a specific website to participate in discussion. Given different online health intervention programs, in this chapter we investigate and compare user characteristics and behaviors on different channels.

2.1 Problem Definition

QuitNet Forum and QuitNet Facebook, the representatives of two different types of online health communities, are supervised by the same organization – QuitNet. However, we do not know whether they attract users with similar characteristics, or whether users behave similarly in the two different communities. In light of this, the following research questions are proposed:

RQ1.1: What are the differences between user quit statuses of QuitNet Forum and QuitNet Facebook?
RQ1.2: What are the differences of user behaviors between QuitNet Forum and QuitNet Facebook?

RQ1.3: Are user behaviors associated with quit statuses on QuitNet Forum and QuitNet Facebook?

To solve these problems, we collected discussion thread data from QuitNet Forum and QuitNet Facebook, each with different accessibility features. All data collected are publicly and freely accessible. On the QuitNet Forum website, any registered user is allowed to initiate new threads (posts) and make comments. However, only the creator of the public page could start new threads on QuitNet Facebook. Only Facebook users who subscribed to the page by “liking” it can comment on posts. As a result, there are substantially more threads on QuitNet Forum than those on QuitNet Facebook during the same time period. In order to keep the sample sizes relatively similar from each website, we collected one-month data from a QuitNet Forum (May 1, 2011 – May 31, 2011), and three-month data of QuitNet Facebook (April 1, 2011 - June 30, 2011). This extraction process resulted in 3,017 posts and 24,713 comments made by 1,169 users from QuitNet Forum, and 111 posts and 2,574 comments made by 664 users from QuitNet Facebook.

User quit statuses were acquired from both QuitNet Forum and QuitNet Facebook. On QuitNet Forum, users voluntarily post their quit dates on their profile pages and their quit statuses could be calculated. On QuitNet Facebook, a weekly post of “shout your quit status” is created on every Friday. Users reply to this post and disclose their quit statuses. By analyzing the comments, the days of abstinence of these users can be calculated. From our collected data set, we extract 534 out of 1,169 users (45.6%) of QuitNet Forum and 394 out of 664 users (59.3%) of QuitNet Facebook whose days of abstinence can be identified.

2.2 Measurements

2.2.1 Quit Statuses and Quit Stages (RQ 1.1)
In most studies of online intervention programs of smoking cessation, there are only two types of quit status defined for users: smoking and abstinence. A user is described as either smoking or abstinent in these studies for analysis (Cobb & Graham, 2006; Cobb et al., 2010; Cobb et al., 2005). However, quitting smoking is a continuous process that may last for many years. It is difficult to develop a single measurement for all cases (Velicer & Prochaska, 2004). In this dissertation, we use a continuous value to represent quit status. Concretely, the quit status of a user is the number of days that she has been abstinent from the self-reported day she stops smoking to the day she posts the last message on QuitStop in our data sample. We calculate the quit statuses of support givers and receivers identified from the data sample.

Velicer et al. (1992) defined five stages of quitting smoking, which are precontemplation, contemplation, preparation, action and maintenance. Once a smoking quitter stops smoking, she enters the action stage, which is composed by two periods: an early action period and a late action period. In the early action period, the user has been abstinent for 0 to 3 months; and in the late action period, the user has been abstinent for 3 to 6 months. After being abstinent for 6 months, the smoking quitter moves into the maintenance stage, which is suggested to be 6 to 60 months after quitting smoking (Velicer, Prochaska, Rossi, & Snow, 1992).

According to quit statuses, users are categorized into five groups of quit stages in this study. Based on stages since taking actions to quit smoking, the first group is composed by users at the early action stage with the quit status of 0 to 3 months. The second group is composed by users in the late action stage with quit status of 3 to 6 months. Users in the third and fourth groups are at the early and late maintenance stages with the quit statuses of 6 months to 2 years and 2 years to 5 years, respectively. The fifth group is composed of users who have been abstinent for more than 5 years.
To compare the quit stages of users on QuitNet Forum and QuitNet Facebook, we implement statistical analysis to compare the user distributions of quit stages on these two different communities.

2.2.2 Measurement of User Behavior (RQ1.2)
To solve RQ1.2, user behavior is studied from two aspects: user response time and social network analysis. User response time reflects the user’s immediacy to respond to a post. Social networks are built from user interactions in these two communities, and a series of analysis are implemented to analyze user behavior.

(1) Analysis of Response Time
After a thread is initiated and a post is published, it takes some time for people to respond to the post and participate in the discussion. The response time to a post could reflect a user’s interests in the discussion topic as well as his/her activeness in the community. On a smoking cessation forum, Selby et al. (Selby et al., 2010) found that posts with the theme of struggling and seeking for support or advice can receive replies from users in a shorter time. In our study, Facebook is a much more popular social networking site than QuitNet, and general users spend more time on Facebook every day. However, as Facebook is not a professional website for smoking cessation, QuitNet Facebook may not attract as many active users as QuitNet Forum who have strong determination and urgent desire to quit smoking. Thus, users of the two communities may have different patterns to participate in discussions by responding to posts.

In this study, for each user of QuitNet Forum and QuitNet Facebook, we extract all posts that he/she comments on and calculate the amount of time (in seconds) that he/she takes to comment on each post since the post is published. The average value of response time to all posts that a user responds to is defined as his/her average response time, which reflects the response immediacy of that user. User average response time could indicate how fast a user responds to a
post on average if participating in the thread. In our study, we calculate the average response time of all users on QuitNet Forum and QuitNet Facebook, and a T test is used to compare the means of average response time between users of the two communities. In addition, for each of the community, users are divided into seven groups according to their average response time. In each group, users have average response times of 10 minutes, 10 to 30 minutes, 30 minutes to 1 hour, 1 hour to 2 hours, 2 hours to 5 hours, 5 hours to 1 day, and more than 1 day, respectively.

(2) Social Network Analysis

Social network analysis is an important methodology to study user behavior, which can identify connections between user pairs and investigate user interactions from a macro perspective. Cobb et al. built a social network on different types of user interactions on QuitNet website to extract subgroups and analyzed features of core users and periphery users in the social network (Cobb et al., 2010).

In this study, two undirected graphs are constructed as social networks of QuitNet Forum and QuitNet Facebook. For each of the social network, actors (nodes) represent users of the community, and a tie (link) connects two actors who have participated in the same thread of discussions. Each tie carries a weight indicating the number of threads that the corresponding users participate in together. For the social networks of QuitNet Forum and QuitNet Facebook, the network size and average weighted degree centrality are analyzed and compared. The network size of a social network is defined as the number of actors involved in linkages of the network (Chang, 2009). The weighted degree centrality of an actor is defined as the sum of weights of all ties linked to it. This indicator reflects the frequency and activeness of a user’s participation in discussions.
2.3 Experiment and Result

2.3.1 Comparison of User Quit Stages (RQ1.1)

534 and 394 users with valid quit status are extracted from QuitNet Forum and QuitNet Facebook respectively. As mentioned earlier, these users are divided into five quit stages as shown in Table 2.1. The Pearson’s chi-square test is used to compare the distributions of user frequency with different quit statuses between QuitNet Forum and QuitNet Facebook. The result shows a significant difference on user frequency distributions between these two communities (p<0.001). Users of QuitNet Forum and QuitNet Facebook distribute in different groups of quit stages. According to Table 2.1, it is observed that most users of QuitNet Forum are in the stage with 0-90 days of abstinence, which indicates that they have been abstinent for less than three months and are at the early action stage of smoking cessation. But for QuitNet Facebook, the numbers of users in different stages are not substantially different from each other except the last stage.

<table>
<thead>
<tr>
<th>Days of Abstinence</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-90</td>
<td>232</td>
<td>69</td>
<td>100</td>
<td>72</td>
</tr>
<tr>
<td>91-180</td>
<td>69</td>
<td>100</td>
<td>72</td>
<td>61</td>
</tr>
<tr>
<td>181-720</td>
<td>72</td>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>721-1800</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;1800</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3.2 Comparison of User Behaviors (RQ1.2)

(1) Analysis of Time

The average response time is calculated for each user of QuitNet Forum and QuitNet Facebook. Among 1,169 users of QuitNet Forum, 1,040 of them have made comments in threads. Others only initiate new threads and do not respond to any posts. For the average response time of 1,040 users of QuitNet Forum and 664 users of QuitNet Facebook, two outliers are detected for each community, which are removed for analysis. Table 2.2 lists the statistics of response time of the
two communities. For QuitNet Forum, the mean of average response time is 10768.24 seconds (nearly 2 hours), and that for QuitNet Facebook is 19765.38 seconds (more than 5 hours). A T test is conducted, which indicates significant difference (p<0.001) on average response time between users of these two communities. In general, users of QuitNet Forum respond faster than Facebook users. On QuitNet Facebook, the range of user average response time is larger, and the standard deviation is higher.

Table 2.2 Statistics of User Average Response Time of QuitNet Forum and QuitNet Facebook

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forum</td>
<td>1038</td>
<td>10768.24</td>
<td>21813.598</td>
<td>[54, 232995]</td>
</tr>
<tr>
<td>Facebook</td>
<td>662</td>
<td>19765.38</td>
<td>45521.298</td>
<td>[29, 426377]</td>
</tr>
</tbody>
</table>

Users of QuitNet Forum and QuitNet Facebook are divided into seven groups according to their average response time. Figure 2.1 presents the user distribution in each group for QuitNet Forum and QuitNet Facebook. There are 1024 out of 1038 Forum users (98.7%) and 625 out of 662 Facebook users (94.4%) with the average response time of less than 24 hours. For QuitNet Facebook, the numbers of Facebook users in the first six groups (different groups within 24 hours) are close. However, the Forum users in the group with average response time of 1 hour to 2 hours have the highest percentage.
(2) Social Network Analysis

Two social networks are developed for QuitNet Forum and QuitNet Facebook respectively as shown in Figure 2.2 and Figure 2.3. The social network of QuitNet Forum has 1,169 actors and 40,540 ties, and that of QuitNet Facebook has 664 actors and 19,904 ties. In the social networks of QuitNet Forum and QuitNet Facebook respectively, one and two actors do not involve in any linkages with other actors. So the network sizes of Forum and Facebook are 1168 and 662.

The weighted degree centralities of all actors are calculated for the two social networks to understand the distribution of user activity in the network. The average weighted degree centralities of QuitNet Forum and Quit Facebook are 213.223 and 78.374 respectively. The former is nearly as three times as that of the latter. However, the network size of QuitNet Forum (1168) is about twice as that of QuitNet Facebook (662), and the Forum data is collected in a shorter time period. On average, every two Forum users contact with each other 0.322 \( (213.223/1168) \) times in one month, but every two Facebook users contact with each other 0.118 \( (78.374/662) \) times in three months. So generally, users of QuitNet Forum behave more actively.
than users of QuitNet Facebook. They participate in discussions more frequently on average, and closely interacted with others.

Figure 2.2 Social Network of QuitNet Forum
2.3.3 User Behavior of Different Quit Stages (RQ1.3)

For QuitNet Forum and QuitNet Facebook, we compare user average response time between the five quit stages using ANOVA with LSD post-hoc tests. For both of the communities, the P values of ANOVA tests were above 0.1 (P = .47 for QuitNet Forum and P = .25 for QuitNet Facebook). For QuitNet Forum, LSD post-hoc tests showed no significant difference on user average response time between any groups at the level of P = .1. However for QuitNet Facebook, there was significant difference between Group 1 (0-90 days) and Group 3 (181-720 days) with P = .096, and between Group 3 (181-720 days) and Group 5 (>1800 days) with P = .094. The mean of average response time of users in different quit stages is shown in Table 2.3. Users of Group 3 were likely to respond faster than users of most other groups. The mean of their average response time was about 4 hours.
For QuitNet Forum and QuitNet Facebook, we also compare user weighted degree centralities between the five groups of different quit stages. Kruskal-Wallis test and Independent Sample Median test were carried out for the two social media channels respectively. Results show that there is no significant difference on degrees across the five groups of QuitNet Forum (P = .14 and P = .07), but there is significant difference for that of QuitNet Facebook (P = .03 and P = .02).

The average degree of Facebook users in each group is shown in Table 2.3. To further explore the differences, ANOVA with LSD post-hoc test is implemented for QuitNet Facebook. Although P = .19 for ANOVA test, the LSD post-hoc tests showed significant differences between Group 1 (0-90 days) and Group 3 (181-720 days) with P = .035, and between Group 3 (181-720 days) and Group 5 (>1800 days) with P = .1. It indicates that the actors of Group 3 (181-720 days) are likely to have higher degree centrality than those of other groups, which is similar to the difference on user response time across five quit stages.

So, on QuitNet Forum, users at different quit stages have similar behaviors. But on QuitNet Facebook, users at early maintenance stage behave more actively.

<table>
<thead>
<tr>
<th>Table 2.3 Degree Centrality and Average Response Time in Each Group of Quit Status of QuitNet Facebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quit Status (days)</td>
</tr>
<tr>
<td>Mean of Degree Centrality</td>
</tr>
<tr>
<td>Mean of Average Response Time</td>
</tr>
</tbody>
</table>

2.4 Discussions

Summarizing all the results, there are differences of the two samples for average response time, social network structures, and distribution of quit stages. Generally, users of QuitNet Forum perform more actively than Facebook users, with the evidences that Forum users respond faster and more frequently. Different users of QuitNet Forum behave similarly to each other, because most users respond to post in 1 to 2 hours, and there is no difference of user behavior between different quit stages. Users of QuitNet Facebook behave differently from each other, with the
evidences that Facebook users have variant response time, and users at early maintenance stage perform more actively than others. The features of the two social media channels may explain user differences between the two communities.

Active participation of QuitNet Forum requires each user to have an account registered. Only recent smoking quitters who have great enthusiasm and determination are likely to log on QuitNet Forum frequently to participate in discussions. So, most forum users are active and behave similarly to each other. QuitNet Forum is a community of practice (CoP), which is a group of people who “share a concern or a passion for something they do and learn how to do it better as they interact regularly” (Wenger, 2005). In QuitNet Forum, people share the same purpose of smoking cessation and practice together to achieve this purpose. With high average weighted degree centrality, the social network of QuitNet Forum has strong ties between different actors, which indicate strong and deep relations between users in this community (Granovetter, 1973).

QuitNet Facebook is a community page on a popular and open platform for social communications, requiring Facebook accounts. Because their main account is their personal account with QuitNet Facebook as an ancillary mailing list, they can passively receive messages from QuitNet in their “news feeds”. This allows for discussions of smoking cessation while interacting with other friends on Facebook. As a result, people from a wider range are attracted to QuitNet Facebook. Their main purposes of logging on Facebook may not be smoking cessation. They participate in the discussions casually and frequently, as evidenced by the longer average response time and more scattered network shape. QuitNet Facebook is a community of interest (CoI) that is regarded as a “community of communities” (Fischer, 2001). It brings together people with different purposes and the same interest. The ties in the social network of QuitNet Facebook are weak, with the evidence that it has low weighted degree centrality on average. This finding is consistent with a general finding that Facebook emerges a publicly open structure with
loose behavioral norms of participants (Papacharissi, 2009). The weak ties play an important role of “bridge”, which connects to different networks and expands user relations (Granovetter, 1973).

The response patterns of users on QuitNet Forum are similar. A relatively higher percentage of messages have an average response time of 1 hour to 2 hours. The reason for this quicker response may be that QuitNet Forum has a unified and static interface for all users. The latest posts and comments always appear at the top of the Forum page. So the interface and content are consistent for all users when they log on QuitNet Forum at the same time. Old posts are replaced by new posts and removed from the first page of the forum. Users need to make additional efforts to go to the second or later pages to read and respond to old posts. As a result, all users tend to reply more posts at the top of the first page, which might be published in a few hours. So their average response time is very similar.

However, the response time of QuitNet Facebook users varies substantially from each other. Users’ average response times are distributed widely along a spectrum. This may be because the users of QuitNet Facebook are notified of new posts automatically in their main page, which contains a “news feeds”. It is not necessary for them to navigate to the public page of QuitNet Facebook for the ability to interact with others of the community. From their homepage of Facebook through a computer device, they can read messages from their subscribed “news feeds”. During the same periods, different Facebook users receive different numbers of messages in their “news feeds”. Some users may receive a lot of messages from variant sources on Facebook, so posts of QuitNet Facebook could not stay for a long time at the top of their “news feeds”. But some users may receive fewer messages, so posts of QuitNet Facebook could stay much longer at the top of their “news feeds”. So, different users of QuitNet Facebook make different efforts to read old posts of QuitNet Facebook. As a result, the response immediacy of QuitNet Facebook users varies greatly from each other.
2.5 Conclusion

The results summarized in Table 2.4 show that the characteristics and participation behaviors of users of QuitNet Forum and QuitNet Facebook are different. A large portion of users participating in the QuitNet Forum is in the earlier stage of smoking abstinence. However, the numbers of Facebook users are equally distributed across different levels of quit statuses. Users of QuitNet Forum participate more consistently, whereas Facebook users behave sporadically. Generally, users of QuitNet Forum perform more actively than Facebook users, with the evidences that Forum users respond faster and more frequently.

QuitNet Forum and QuitNet Facebook represent two general types of communities for healthcare topics as mentioned above. Our study initially compares specific examples of the two types of communities with user behavior investigation. However, it is only a preliminary study with basic data analysis and social network analysis. The interaction process between users in the communities is not fully observed or analyzed. To get a more thorough idea about how active users behave and interact with each other, we carry out an in-depth study for QuitNet Forum in the next chapter.

| Table 2.4 Result summery of the comparative study of QuitNet Forum and QuitNet Facebook |
| --- | --- | --- |
| **User Behavior** | **QuitNet Forum** | **QuitNet Facebook** |
| **Response Immediacy** | Mean of Average Response Time | About 3 hours | About 5 hours |
| | User Distribution in Groups of Average Response Time | Centralized: most users respond in 1 to 2 hours | Scattered: users are equally distributed in different groups |
| **Social Network Analysis** | Average Weighted Degree | 213.223 | 78.374 |
| | Network Centralization | 2.96% | 10.17% |
| **Quit Status** | User Distribution in Groups of Quit Statuses | Centralized: most users are in early action stage of quitting | Scattered: users are equally distributed in different groups |
3. Social Support Exchange Patterns and User Behaviors

In the Chapter 2, it is learned that users on QuitNet Forum are more motivated and active to take part in discussions. In this chapter, we focus on QuitNet Forum to carry out a thorough study of user behaviors and interaction patterns from the communication level. In health intervention programs, social support plays an important role in helping people to achieve better intervention outcomes. In QuitNet Forum, both smokers and ex-smokers participate in discussions of smoking cessation and exchange social support. It provides an ideal platform, in which smoking quitters can receive effective social support and achieve good intervention outcomes. However, existing studies of online smoking cessation programs did not analyze social support exchange from the perspective of user interactions. In this study, we will analyze social support exchange in the QuitStop forum online community. The types of social support will be extracted, and the patterns of social support exchange will be analyzed.

3.1 Problem Definitions

The exchange of social support is an interactive process. Although many studies analyzed the types of social support offered in different online communities, they ignored investigating user interactions (Cutrona & Suhr, 1992; Hwang et al., 2011; Mo & Coulson, 2008). Some studies tried to explore the “exchange of social support”, but they only counted and compared the numbers of messages offering and requesting social support (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Eichhorn, 2008). In this study, we analyze posts and comments extracted from QuitNet Forum. To understand the discussion content on this forum, we first investigate discussion themes of the messages from the perspective of social support. The research question is proposed:

RQ2.1: What are the discussion themes on QuitNet Forum in terms of social support exchange, and what types of social support are involved in the discussions?
Moreover, we are also interested in the interaction of users to exchange social support. Based on information seeking behavior in everyday life, Savolainen (1995) proposed two information patterns, which are practical information and orienting information. Similar to information exchange, social support exchange is a social phenomenon. It is important to study who exchanges information with whom, about what, and by which media (Haythornthwaite & Wellman, 1998). Motivated by information patterns defined by Savolainen (1995), we focus on the exchange patterns and user behaviors of social support. The research question is proposed:

RQ2.2: What are the exchange patterns of social support between user pairs on QuitStop forum?

Similar to information exchange, social support exchange is a social phenomenon. It is important to study who exchanges information with whom, about what, and by which media (Haythornthwaite & Wellman, 1998). To explore “who exchanges social support with whom”, we analyze social support exchanged among users at different quit stages. We propose the following research question:

RQ2.3: How do users at different quit stages interact with each other to exchange different types of social support?

In the followings of this section, we will solve these three questions respectively. To solve RQ2.1, we analyze the discussion themes and types of social support with a qualitative study (Section 3.2). For RQ2.2, we define patterns and user behaviors for social support exchange, and apply statistical methods to analyze the patterns (Section 3.3). RQ2.3 is addressed through analyzing user interactions between different quit stages with social network analysis (Section 3.4).

### 3.2 Discussion Themes and Types of Social Support Exchange (RQ2.1)

To solve RQ2.1, we analyze discussion messages on QuitNet forum from the perspective of social support exchange. Between 05/01/2011 and 05/31/2011, 3017 posts and 24,713 comments
were made on QuitStop forum. We randomly select 228 posts with their corresponding 1672 comments as a sample for the analysis. We use content analysis to investigate the content of posts and comments, and indentify the types of social support from the messages.

3.2.1 Coding Scheme
To attain a general and broad description of content, qualitative analysis is widely used to extract categories or codes of content. Inductive analysis and deductive analysis are two basic methods for content analysis (Elo & Kyngäs, 2008). Inductive analysis uses open coding to write all notes of content texts, and lists hierarchical categories. Deductive analysis is applying existed categories or codes in a new context.

For online healthcare communities, different coding schemes have been developed for discussion topics (Ahmed, Sullivan, Schneiders, & Mccrory, 2010; Bender, Jimenez-Marroquin, & Jadad, 2011; Eichhorn, 2008; Greene, Choudhry, Kilabuk, & Shrank, 2011) and types of social support (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Cutrona & Suhr, 1992; Eichhorn, 2008; Hwang et al., 2011; Mo & Coulson., 2008). In our approach, we first use deductive methods to apply existing codes on the sample data. Then inductive methods are used to revise, summarize and improve existing codes to develop a new scheme to better describe the data. Two coders independently code the messages (including all posts and comments) based on the new coding scheme, and the interrater reliability is calculated.

We identify five main categories for the content of posts and comments, including offering social support, requesting social support, receiving social support, other activities and irrelevant content. These categories describe the themes of posts and comments. The first three categories are related to user interactions with social support. Posts and comments which are not associated with direct exchange of social support are divided into the last two categories. Messages can be coded in more than one category, but most messages are only coded in one category. For each category,
there are different aspects and subcategories which describe the theme in detail. The descriptions and examples of the coding scheme are listed below.

(1) Offering Social Support

Users may offer social support for others in posts and comments. The types of social support are identified for messages in this category. Informational support and nurturant support are two main types of social support, which are described in (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Cutrona & Suhr, 1992). They also have specific subcategories.

*Informational Support*: offering specific information on treatment or coping. Subcategories include:

- **Advice**: offering suggestions to solve specific problems according to a recipient’s situation.
  
  e.g., “Too soon to go off the zyban. Stay on it until you feel like an ex-smoker than think about weaning off the zyban. See what the doctor says”.

- **Referral**: Referring recipients to other sources for further help.
  
  e.g., “The cognitivequitting.com website was very helpful”.

- **Fact**: reassessing the situation and presenting fact.
  
  e.g., “Vitamin B12 and D3 work on the energy stores of your body”.

- **Perceptual knowledge**: providing sensory information to reassess smoking cessation and help to establish confidence.
  
  e.g., “A crave is just a cramp. Not serious, it'll pass off soon”.

- **Personal experiences**: telling stories about personal experiences for suggestions.
  
  e.g., “when I first started on the patch, my BP was a bit shaky for while too, even though I'm on BP meds”.

- **Feedback/opinion**: a judgment about a recipient’s situation or idea.
e.g., “sounds like your coming along very nicely”.

*Nurturant Support:* expressing caring or concern. Subcategories include:

- **Esteem:** praising the achievement of support seekers.
  e.g., “Great job on your quit so far”.
- **Network:** broadening support seekers’ network to let them feel not alone.
  e.g., “Reach out and grab someone else... stay close to the Q these first few days.... stay busy helping others and let yourself be helped”.
- **Emotional:** providing sorrow and understanding, which helps support seekers building confidence.
  e.g., “When you get to that point, you will be fine”.

(2) **Requesting Social Support**

Some users try to seek for social support from others. Five strategies of soliciting social support were described in (Eichhorn, 2008). We summarized these strategies and found that messages requesting social support may include two aspects. First, users describe their personal situations to ask for others’ responses. Second, users explicitly ask for certain types of social support. Some posts and comments in this category may only include one aspect.

*Description*

Messages may describe the personal situations during the process of smoking abstinence. The purpose is to attract others’ attention and ask for their support. Users may describe their situations from three aspects.

- **Action:** describing the experiences of smoking or actions taken to quit smoking.
  e.g., “I have been unsuccessfully trying to quit for the past year”.
- **Feeling:** describing the emotional feeling or suffering during abstinence.
e.g., “I do sleep a lot and my mind receptors are ‘flickering’ to make me foggy ... and I do have those mood swings.”

- **Attitude:** expressing the positive or negative attitude (determination, desperation, etc.) to smoking cessation.
  e.g., “I make my own decisions. I govern my own future. And I'm not going fall for any of your dirty tricks. I realize that this declaration will anger you, tobacco, and in the next few months you will bombard me with everything you've got.”

These three descriptive subcategories are not exclusive. For each message describing the personal situation, there may be more than one type of description included.

### Types of social support

Some posts and comments explicitly indicate the types of social support to request. Some of them ask for informational support, e.g., “Hey all, today is my day one and I noticed that my chest hurts a little. It almost feels sore in there. Is this normal”? Others seek for nurturant support, e.g., “I'll be glad to take any prayers or positive thoughts you want to send my way”.

(3) Receiving Social Support

Users also express receiving social support from others in some messages. However, there is little research exploring how people express receiving social support. We construct three subcategories for this theme, which describe the response of users after receiving support.

**Expressing thanks:** simply expressing receiving support and thanks, without providing any additional information. e.g., “Thanks everyone! I really appreciate the support and will be checking in often”.

**Expressing feedbacks:** adding personal ideas or further information on received support. e.g., “Thanks everyone! My quit is FOR ME... To have more energy, more time, more money. It's definitely for me”.

**Requesting further support:** describing more details of personal situation to seek for further support. E.g., “Thanks for the help. But now that I have cut my patch in half and have been on half the 21 for 4 days. Do you think it will be ok to go just to the 14mg”?

(4) Other Activities

Social support is exchanged through posts and comments in the first three categories. But on QuitStop, there are also messages about smoking cessation which do not involve interactions with social support. Two subcategories are defined for these activities.

**Calling for supporting others:** calling for social support for others but not for self. E.g., “To those of us that have been blessed to know PetroJMan (Bruce) please take a moment to send him a big hug as he has always done for us...and your prayers and good thoughts”.

**Self disclosure** (White & Dorman, 2001): describing personal information, without exchange of social support. E.g., “7 days, 16 hours, 38 minutes and 59 seconds smoke free”.

(5) Irrelevant Content

Some posts and comments do not associate with smoking cessation directly. They are about other issues. E.g., “Iced mochas are my favorite form of caffeine”. For these posts, we regard them as irrelevant, and they are excluded in later analysis.

3.2.2 Result of Content Analysis

The 228 posts and 1672 comments are coded by two coders independently. The probability of random agreement is 95.0%, and the average Cohen’s Kappa of all categories is 58.0%, which indicates a “moderate agreement” (Viera & Garrett, 2005). When calculating Cohen’s Kappa, we ignore the hierarchical structure of categories and regard all of them at the same level. As a result, the probability of each message belonging to a certain category is low, which reduces the value of Cohen’s Kappa. However, the high rate of agreement still indicates a strong agreement between
two coders. The coding result is summarized in Table 3.1, which shows the numbers of posts and comments in each category, including main categories and all subcategories.

Table 3.1 Number of Messages in Each Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Posts</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Offering Support</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informational Support</td>
<td>51</td>
<td>371</td>
</tr>
<tr>
<td>Advice</td>
<td>2</td>
<td>124</td>
</tr>
<tr>
<td>Referral</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Fact</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>Perceptual Knowledge</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>Personal Experience</td>
<td>13</td>
<td>178</td>
</tr>
<tr>
<td>Feedback/opinion</td>
<td>0</td>
<td>112</td>
</tr>
<tr>
<td><strong>Nurturant Support</strong></td>
<td>8</td>
<td>525</td>
</tr>
<tr>
<td>Esteem</td>
<td>2</td>
<td>363</td>
</tr>
<tr>
<td>Network</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Emotional</td>
<td>6</td>
<td>182</td>
</tr>
<tr>
<td><strong>Requesting Support</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>110</td>
<td>9</td>
</tr>
<tr>
<td>Action</td>
<td>94</td>
<td>7</td>
</tr>
<tr>
<td>Feeling</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>Attitude</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informational Support</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>Nurturant Support</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td><strong>Receiving Support</strong></td>
<td>3</td>
<td>294</td>
</tr>
<tr>
<td>Expressing thanks</td>
<td>2</td>
<td>163</td>
</tr>
<tr>
<td>Additional Information</td>
<td>1</td>
<td>84</td>
</tr>
<tr>
<td>Requiring further support</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td><strong>Other Activities</strong></td>
<td>12</td>
<td>133</td>
</tr>
<tr>
<td>Calling for supporting others</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Self disclosure</td>
<td>10</td>
<td>135</td>
</tr>
<tr>
<td><strong>Irrelevant</strong></td>
<td>53</td>
<td>543</td>
</tr>
</tbody>
</table>

Figure 3.1 presents the distribution of five main themes in posts and comments. A few messages contain more than one theme. For example, three posts offer and request social support at the same time. Twenty-one comments both offer and receive support. Requesting social support is the most popular theme of posts (48%). However, it is the most unusual theme for comments (1%). Most comments are in the category of offering social support (42%).
For social support offered in posts, 86.4% is informational support and 13.6% is nurturant support. But 37.6% of social support offered in comments is informational, and 62.4% is nurturant. Figure 3.2 shows different types of informational support offered in posts and comments. Perceptual knowledge is offered much more frequently than other informational support in posts. It provides sensory information to help people reassess situations and establish determinations. For informational support offered in comments, personal experience, feedback and advice are more prevalent. These kinds of supports are targeted on specific individuals and cases.

The comparison of different types of nurturant support offered in posts and comments is shown in Figure 3.3. The most frequent nurturant support offered in posts is emotional support, and most frequent offered in comments is esteem. Network support is not common in either posts or comments.
Requesting Social Support and Receiving Social Support

All messages requesting social support describe the specific user situations, including action, feeling and attitude, to inspire response from others. The types of descriptions in posts and comments are shown in Figure 3.4. Most users describe their actions in messages, telling people about their experiences and actions adopted to quit smoking to seek for social support. Only a small number of messages explicitly indicate the types of social support to seek for.

Receiving social support is unusual in posts, but it is an important theme in comments, which covers 17% of comments. Figure 3.5 shows the distribution of its subcategories in comments. Expressing thanks is the most basic and frequent type of receiving social support. Some users also express further opinions based on information received. There are also users describing more details about their situations to solicit more support.
3.2.3 Discussions

Similar to other studies (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Eichhorn, 2008; Greene et al., 2011), requesting and offering social support are two important discussion themes in this online community of smoking cessation. Our result is consistent with former research that users offer informational and emotional support to help others achieve better health intervention outcomes (K. Chuang & Yang, 2010; K. Y. Chuang & Yang, 2011; Eichhorn, 2008; Finn, 1999; Schafer et al., 2007). In posts, users provide more informational support than nurturant support. But in comments, nurturant support is more prevalent. Nurturant support expresses caring and comforting. Usually, it is offered to a specific recipient. So, there is seldom nurturant support offered in posts, because posts usually direct at a wide range of users. Similar to findings of Caplan and Turner (2007), users on QuitNet Forum are comfortable expressing personal narratives and disclosing themselves, which supports the idea that computer-mediated communication can establish effective comforting conversations. Esteem and emotional support are common types of nurturant support. Consistent with studies of other online health communities (Caplan & Turner, 2007), many support providers show their understanding, encouragement and empathy for support receivers in the QuitNet Forum.
Users adopt different strategies to seek social support (Eichhorn, 2008). In the QuitStop community, most users describe individual actions to inspire replies from others. It is similar to Eichhorn’s finding that sharing experiences is the most frequent soliciting strategy for social support in online eating disorder support groups (Eichhorn, 2008). We also find that many users describe their feelings, including positive emotion and suffering, to ask for social support. It is similar to the finding that self-deprecating is an important strategy for soliciting social support (Eichhorn, 2008). But different from some other research (Eichhorn, 2008), there are only a few messages explicitly proposing types of support to request. When seeking social support on QuitStop, many people do not clearly express their intention of requesting support. Most of them just describe their personal situations, and other people would respond to the narratives and offer different social support. For users, QuitNet is not a platform for asking and answering questions about smoking cessation. It is a community where people share experiences and care for each other.

QuitNet Forum achieves conditions proposed in (Burleson & Goldsmith, 1988) to establish effective comforting conversations. When requesting social support in this community, many users are willing to talk about their personal experiences and upsetting matter. This is similar to Caplan and Turner’s finding (Caplan & Turner, 2007) that computer-mediated comforting communication creates an environment to encourage people to express personal narratives. Most support providers in this community show their understanding for support seekers. They appraise support seekers’ achievements and help them rebuild confidence. It is consistent with discussions in many other online support groups where empathy is a prominent feature (Caplan & Turner, 2007).

Besides offering, requesting and receiving social support, there are also other discussion topics about smoking cessation in this community. From time to time, some posts invite people to
“shout quit status” which encourage users to disclose themselves. Many users reply to the posts and disclose their status of smoking cessation. There are also messages calling for supporting others. But different from the theme of “requesting social support”, authors of these messages do not ask for supporting themselves. Instead, they call for supporting certain people who do not participate in the community at that time (for example, people who are ill or pass away). Or they call for mutual support among users in the community. So, exchange flow of social support cannot be extracted from these messages.

3.3 Patterns and User Behaviors of Social Support Exchange (RQ2.2)

To solve the question of “what are the exchange patterns of social support between user pairs on QuitNet Forum”, we define two patterns of social support exchange motive by the theory of user information behavior, and extract different patterns for each type of social support from the data set.

3.3.1 Measurement of Exchange Patterns

In this study, the definition and extraction of social support exchange are motivated by studies of information exchange. As mentioned above, practical and orienting information are two main types of information describing information exchange patterns (Savolainen, 1995). In this study, we focus on the exchange of social support, which is in a narrower scope than information exchange. The naming and definition for practical and orienting information are based on user seeking behavior. They reflect the information needs of information seekers (receivers). In this study, we define social support exchange patterns from both the perspectives of support givers and receivers. Similar to practical information seeking, some social media users explicitly express their need of social support (either informational support or nurturant support) and invite their peers to offer supports to them. On the other hand, similar to orienting information seeking, some social media users do not have specific questions in mind but they participate in the
community and interact with other users. Other users may initiate social supports to these users without any requests while interacting with them. Two basic patterns of support exchange are defined here in this study, which are initiated support exchange and invited support exchange as shown in Figure 3.6.

In the initiated-support exchange pattern, social support is offered initially and voluntarily without request. Usually, social support exchanged in this pattern is general and can benefit all users. For example, a user, marked as G, initiated a thread and wrote in the post that “…smoking is such an anti-life activity. There really are no redeeming qualities of this activity. It depletes our wallets, it makes us stink and it causes all sorts of health problems…” The theme of this post is offering support. Two users, R1 and R2, commented on this post to indicate receiving support, saying “very well said. This is going on the ‘My Good Words’ list I keep in my purse. Thanks” and “I love the way you make us think”. In this case, no user requested social support. The support was initially given by G and received by R1 and R2 respectively. In this pattern of initiated support exchange, social support flows from G to R1, and from G to R2.

In the invited-support exchange pattern, social support is requested before being offered. The support seeker actively requests support. Usually, social support exchanged in this pattern is offered to an individual user. For example, a user, marked as R, initiated a thread to request support, writing that “I no longer feel like I have to smoke and must say I look and feel much better, but I still have that desire. Does it ever go away, or is this something I just have to deal with?” This post is to request support. Three users, G1, G2 and G3, commented on this post to offer support, writing that “I'm only at day 22 but can sympathize and send you a few hugs and smiles and your quit is looking great at 60 days”, “…You are doing excellent and will continue to do so” and “You are doing great by posting here! So BRAVO to you for that”. In this case, the
user R actively requests support. G1, G2 and G3 offer support as R expected. In this invited-support exchange pattern, social support flows from G1 to R, from G2 to R and from G3 to R.

For each thread with social support interactions, a series of triples <Ug, Ur, P> are extracted, where Ug is the individual user who gives social support, Ur is the single receiver of social support, and P is the pattern of social support exchange (initiated-support exchange or invited-support exchange). For the two examples mentioned above, five triples could be extracted, which are <G, R1, Initiated>, <G, R2, Initiated>, <G1, R, Invited>, <G2, R, Invited> and <G3, R, Invited>. By examining the social support interactions through all of the collected threads, a triple set is generated, which indicates the frequency of each social-support exchange pattern. In this study, we also extract two subsets from the triplet set, one subset for each type of social support (i.e. informational support and nurturant support). In a triple set, a triple with the same elements could appear more than once, which indicates that different social support is exchanged with the same pattern between the same user pair several times.

Savolainen (1995) pointed out that seeking practically effective information is “active”, and seeking orienting information is “passive”. As we explore behaviors of both support givers and receivers, the active and passive behaviors of giving and receiving support are analyzed, and depicted in Figure 3.6 and Table 3.2. Concretely, four user behaviors of social support exchanges are identified, which are active giving, passive giving, active receiving and passive receiving. The behaviors of active giving and passive receiving are extracted from the initiated support exchange pattern, and the behaviors of passive giving and active receiving are extracted from the invited support exchange pattern. Usually, initiated support is offered in posts. The support giver initially starts a thread and offers support to all users. Other users comment on the posts and express receiving the support in the same thread. In this case, the support giver initiates the support exchange process and is regarded as active. The receivers passively accept the support
without requests. From the perspective of information seeking, initiated support is similar to
orienting information (Savolainen, 1995). The support receivers do not ask for particular social
support. They keep an eye on any information related to their general interests and concerns
(Burnett, 2000). In the invited-support exchange pattern, a support seeker usually starts a thread
and writes a post to request social support. Other users comment on the post and give
Corresponding support to the support seeker. In this case, the support seekers actively request and
receive social support, and the support givers passively offer social support in response. Invited
support is similar to practical information (Savolainen, 1995). With urgent and clear needs of
information, the support seekers ask questions to get specified answers (Burnett, 2000). For the
examples above, G actively gives social support; R1 and R2 passively receive social support; G1,
G2, and G3 passively give social support; and R actively receives social support.

Figure 3.6 Exchange Patterns and User Behaviors of Social Support Exchange
Table 3.2 Patterns and User Behaviors of Social Support Exchange

<table>
<thead>
<tr>
<th>Pattern of Social Support Exchange</th>
<th>User Behavior of Social Support Exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Giving Behavior</td>
</tr>
<tr>
<td>Initiated Support Exchange</td>
<td>Active Giving</td>
</tr>
<tr>
<td>Invited Support Exchange</td>
<td>Passive Giving</td>
</tr>
</tbody>
</table>

3.3.2 Analyzing Result of Exchange Patterns

From the data sample, 881 triples of <Ug, Ur, P> are extracted, which indicates 881 times of social support exchange between user pairs. 226 triples have the P as “initiated-support exchange”, and 655 triples have the P as “invited-support exchange”. Each of the triples is associated with informational support exchange or nurturant support exchange. It can also involve with both support exchange. 583 triples are associated with informational support exchange, and 497 triples are associated with nurturant support exchange. For informational support, there are 226 triples with the P of “initiated-support exchange” and 357 triples with the P as “invited-support exchange”. For nurturant support, there are 16 triples with the P as “initiated-support exchange” and 481 triples with the P as “invited-support exchange”.

Counting all posts and comments in our data sample, there are more messages offering nurturant support (533) than informational support (422). However, fewer triples of <Ug, Ur, P> are extracted for nurturant support (497) than informational support (522), which means that nurturant support is exchanged more frequently between user pairs. The reason is that most nurturant support targets one receiver, so each message offering nurturant support can only be received by one user. However, informational support usually targets multiple users, thus each message offering informational support may be received by more than one user. From the result, for both informational support and nurturant support, invited support exchange is more prevalent. However, for informational support, 38.8% is initiated; while for nurturant support, only 7.8% is initiated. So, comparatively, the pattern of initiated support exchange is more common for the dissemination of informational support. So, online health intervention programs may design
different channels to deliver social support. Informational support could be offered and disseminated in an open platform, where a large number of people could receive the support and benefit from it. Nurturant support could be delivered through private channels, where people can communicate more comfortable and receive one-to-one messages.

### 3.4 User Interactions of Social Support Exchange (RQ 2.3)

Some studies built social networks to analyze user interactions in online health communities (Chang, 2009; Cobb et al., 2010). However, they only explored static indicators to measure the whole social networks, such as network size, density, clique and centralization. They did not identify actor positions or analyzed the interactions between different user groups in the network. In this study, we apply social network analysis to investigate user interactions based on the exchange of different types of social support. We consider user quit statuses in the network structure, and explore interactions between users at different quit stages. The quit status and quit stage is introduced in Section 2.2.1.

As analyzed above, we have extracted 583 triples with informational support exchange and 497 triples with nurturant support exchange from the data sample (⟨Ug, Ur, P⟩). To analyze social support exchange of users at different quit stages, we remove users whose quit status cannot be acquired, and retain user triples with attainable quit status of Ug and Ur for social network analysis. For each type of the social support, a directed and valued social network is developed on support exchange between user pairs. In the social network, users are represented as nodes (actors). Each tie connects two actors between whom the specific type of social support is exchanged. The direction of the tie is from the support giver to the support receiver based on extracted triples. The value of the tie is the number of corresponding triples, which indicates the frequency of support exchange between the two actors in different threads. For informational support exchange, we built a social network of 204 nodes and 397 directed ties. For nurturant
support exchange, we built a social network of 171 nodes and 341 directed ties. By comparing these two social networks, the exchange patterns of different types of social support can be investigated. Network exposure and blockmodel based on quit stages are used for social network analysis in this study.

3.4.1 Network Exposure Model

The definition of network exposure model is introduced in Section 1.2.3. To apply formula 1.1 to our data, we employ continuous values of quit status as elements in the vector $y$. The weight matrix $W$ is built on values of ties in the network. As the social networks of informational and nurturant support are directed, we calculate two types of network exposure for the actors. For each actor $i$ in a social network, $E_i(G)$ is defined as $i$’s network exposure to its support givers. Only actors sending ties to $i$ are selected as neighbors to calculate $E_i(G)$. $E_i(R)$ is defined as $i$’s network exposure to its support receivers. Actors receiving ties from $i$ are selected as neighbors to calculate $E_i(R)$. So, $E_i(G)$ represents the quit status of support givers that $i$ is exposed to, and $E_i(R)$ represents the quit status of support receivers that $i$ is exposed to. For example, in the network of Figure 3.7, U1, U2, U3, U4 and U5 are neighbors who exchange social support with the user $i$. U1, U2 and U3 are support givers to $i$, and U4 and U5 are support receivers from $i$. The weights of ties between $i$ and U1, U2, U3, U4 and U5 are $w_1$, $w_2$, $w_3$, $w_4$ and $w_5$, respectively. The quit statuses of U1, U2, U3, U4 and U5 are $q_1$, $q_2$, $q_3$, $q_4$ and $q_5$, respectively. The network exposure to support givers of the user $i$ is calculated based on quit statuses of support givers U1, U2 and U3, as $E_i(G) = \frac{w_1q_1 + w_2q_2 + w_3q_3}{w_1 + w_2 + w_3}$. The network exposure to support receivers of user $i$ is calculated based on quit statuses of support receivers U4 and U5, which is $E_i(G) = \frac{w_4q_4 + w_5q_5}{w_4 + w_5}$. 
As mentioned above, two social networks are built, one for informational support exchange and one for nurturant support exchange, respectively. For actors in these two different social networks, we compare their network exposures. Specifically, for all actors in the informational-support network, we calculate their network exposures to givers, and represent them as a vector $\overline{E(G)}_I$. For all actors in the nurturant-support network, their network exposures to givers are denoted as $\overline{E(G)}_N$. The mean values of $\overline{E(G)}_I$ and $\overline{E(G)}_N$ are 945.02 and 575.71 respectively, and a T test indicates that they are significantly different ($p<.001$). Similarly, we represent network exposures to receivers as $\overline{E(R)}_I$ and $\overline{E(R)}_N$ for actors in informational-support network and nurturant-support network, respectively. The mean values of $\overline{E(R)}_I$ and $\overline{E(R)}_N$ are 105.74 and 207.43, and a T test shows that they are significantly different ($p=.011$).

So comparatively, the support givers of informational support have been abstinent for a longer time than support givers of nurturant support. The reason may be that informational support is based on knowledge, advice and personal experiences of support givers. Older smoking quitters are more knowledgeable at providing such information than new smoking quitters. Comparatively, nurturant support does not provide information to help solving any health problem. Nurturant support givers show their concern and caring for support recipients, and smoking quitters with any quit statuses can participate in this activity.
Generally, users receiving informational support have quitted smoking more recently than users receiving nurturant support. Comparatively, informational support may be more important for recent quitters, because information and suggestions are necessary for them to cope with problems. However, for people who have been quitted for longer time, nurturant support may be more important because they need encouragement and esteem to keep abstinence.

3.4.2 Blockmodel Based on Quit Stages

The introduction of Blockmodel can be found in section 1.2.3. In this study, we build blockmodels for user quit stages. For each social network, users are partitioned into different positions according to their quit stages. There are five positions for actors in each social network, which are marked by $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, $\mathcal{B}_4$ and $\mathcal{B}_5$. Corresponding to user groups of different quit stages, actors in $\mathcal{B}_1$ are at early action stage of quitting smoking, actors in $\mathcal{B}_2$ are at late action stage, actors in $\mathcal{B}_3$ are at early maintenance stage, actors in $\mathcal{B}_4$ are at late maintenance stage, and actors in $\mathcal{B}_5$ have successfully quit smoking.

The densities of the five blocks are calculated by formula 1.2. To decide a one-block or zero-block with formula 1.3, we designate $\alpha = \varphi \times \Delta_{\text{max}}$, where $\varphi$ is the golden ratio, which approximate 0.618, and $\Delta_{\text{max}}$ denotes the maximum value of elements in the density matrix $\Delta$. For the social networks of informational support and nurturant support, the $\alpha$ value is set as 0.0110 and 0.0064 respectively. According to the density matrix $\Delta$ and $\alpha$, the block matrix $\mathcal{B}$ is built for each social network. Table 3.3 and Table 3.4 present the density matrixes $\Delta$ and block matrixes $\mathcal{B}$ for the social networks of informational support and nurturant support.

For social networks of informational support and nurturant support, we draw reduced graphs to present ties within and between positions as shown in Figure 3.8 and Figure 3.9. In a reduced
graph, positions are represented as nodes and ties between positions are represented as arcs. There is an arc between positions of $v_k$ and $v_l$ if $b_{kl}=1$, and there is no arc between $v_k$ and $v_l$ if $b_{kl}=0$.

Table 3.3 Density Matrix and Block Matrix of Informational Support Network

<table>
<thead>
<tr>
<th>Density Matrix</th>
<th>Block Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>$B_1$</td>
</tr>
<tr>
<td>$v_2$</td>
<td>$B_2$</td>
</tr>
<tr>
<td>$v_3$</td>
<td>$B_3$</td>
</tr>
<tr>
<td>$v_4$</td>
<td>$B_4$</td>
</tr>
<tr>
<td>$v_5$</td>
<td>$B_5$</td>
</tr>
</tbody>
</table>

$B_1$ | 0 | 0 | 0 | 0 | 0 |

$B_2$ | 0 | 0 | 0 | 0 | 0 |

$B_3$ | 1 | 0 | 0 | 0 | 0 |

$B_4$ | 1 | 1 | 0 | 0 | 0 |

$B_5$ | 1 | 1 | 1 | 0 | 0 |

Table 3.4 Density Matrix and Block Matrix of Nurturant Support Network

<table>
<thead>
<tr>
<th>Density Matrix</th>
<th>Block Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>$B_1$</td>
</tr>
<tr>
<td>$v_2$</td>
<td>$B_2$</td>
</tr>
<tr>
<td>$v_3$</td>
<td>$B_3$</td>
</tr>
<tr>
<td>$v_4$</td>
<td>$B_4$</td>
</tr>
<tr>
<td>$v_5$</td>
<td>$B_5$</td>
</tr>
</tbody>
</table>

$B_1$ | 1 | 0 | 0 | 0 | 0 |

$B_2$ | 1 | 1 | 0 | 0 | 0 |

$B_3$ | 1 | 0 | 1 | 0 | 0 |

$B_4$ | 1 | 0 | 0 | 1 | 1 |

$B_5$ | 1 | 0 | 0 | 0 | 1 |

Figure 3.8 Reduced Graph of Informational Support Network

Figure 3.9 Reduced Graph of Nurturant Support Network
From the reduced graph of informational support network in Figure 3.8, informational support usually flows from users at late quit stages to users at early quit stages. Users at early and late action stages (group 1 and group 2) only receive informational support from users at later quit stages. Users at late maintenance stage (group 4) and after the maintenance stage (group 5) only offer informational support to others. Users at early maintenance stage (group 3) give information to users at the early action stage (group 1), and also receive information from users who have completed smoking cessation (group 5). Informational support is usually offered by users at late quit stages that have plenty of experiences and knowledge, and received by users who have been abstinent for shorter time.

Figure 3.9 indicates that similar to informational support, nurturant support is offered by users at late quit stages (group 2, 3, 4 and 5) to those at the early action stage (group 1). But differently, nurturant support is also frequently exchanged between users within the same quit stage. Offering nurturant support is showing understanding and caring. Users with similar quit statuses are likely to better understand each other, because they share similar experiences. Moreover, users at the same quit stage have been supporting each other since they start quitting at the same time. As a result, they maintain a supporting group of the same quit status, and other users who quit smoking earlier and later seem to be less active in interacting with them but find support from their own group.

### 3.5 Conclusion

In this section, we analyzed discussion themes and types of social support exchanged on QuitNet Forum. The exchange patterns of social support and user exchange behaviors were extracted and analyzed. User interactions with informational and nurturant support exchange were explored within and between users in different groups of quit stages.
Our results show that informational support is exchanged more frequently between user pairs, although there are a larger number of messages giving nurturant support in the QuitStop forum. Initiated-support exchange and invited-support exchange are defined as two basic exchange patterns of social support. Correspondingly, four exchange behaviors are extracted for users, including active giving, passive giving, active receiving and passive receiving. When comparing social networks of informational support and nurturant support, it is found that support givers of informational support have been abstinent for a longer time than support givers of nurturant support, and support receivers of informational support have been abstinent for a shorter time than support receivers of nurturant support. Usually, informational support is offered by users at late quit stages to users at early quit stages. Nurturant support is also exchanged among users within the same quit stage.

Studies of social support and support exchange behavior in online health communities could help us better understand user motivations, actions and outcomes of smoking cessation in online intervention programs. By studying exchange patterns and user interactions of social support in health community, we could build different prediction models based on message content and user health features. In the next chapter, we will scale the analytics to a large dataset with limited manual work. Different machine learning approaches are applied to build prediction models for thread classification and user recommendation.
4. Thread Classification and User Recommendation

In previous chapters, we analyze user behaviors and interaction patterns of social support exchange. The extracted knowledge could be applied to develop better services to improve smoking cessation intervention outcomes. On QuitNet Forum and other health social media, there are too many threads and peers on QuitNet Forum. It is difficult for users to find proper discussion topics and peers to interact with. In this section, we develop classifiers to categorize user discussion messages, and build recommendation systems for QuitNet Forum to recommend threads for users to participate in.

Figure 4.1 (a) and (b) represent two different scenarios of threads on QuitStop forum. The users who initiate the two threads publish posts with similar text content (zyban, craving...). However, they post for different purposes, and ask for different types of social support. In (a), the user is asking for informational support to deal with problems. But in (b), the person is asking for emotional encouragement. This illustration shows that classifying the user intentions and the social support types will be helpful in making recommendations. This chapter includes two parts. In 4.1, we develop classification algorithms to automatically categorize threads in QuitNet Forum. In 4.2, we apply different recommendation techniques to recommend threads for users, and utilize classification result to improve recommendation efficiency.
4.1 Classification of Forum Messages

Usually, text features are used for classifications of online messages. For a thread on QuitStop forum, text could be extracted from title, post and comments. We choose different text features to build classifiers separately, and combine them with certain weights to optimize the classification result. Moreover, we extract health information from user profiles, and use their health status as features to improve text classification. Experiments are designed to determine the features that achieve the best performance.

4.1.1 Problem Description

In Chapter 3, we develop classification schemes to describe discussion topics and social support types on QuitNet Forum. Understanding user generated content in QuitNet Forum can help us build recommendation systems. However, it requires a lot of manual work to analyze each message with Qualitative Analysis. It would be helpful to apply machine learning techniques and develop classifiers to categorize posts and comments on QuitNet Forum. Based on user interaction patterns learned in Chapter 3, text and health feature sets are built.
For a classification task, there are k classes denoted as \( C = (c_1, c_2, \ldots, c_k) \). A post or comment could be represented as a set of features \( X = (x_1, x_2, \ldots, x_n) \). The classification goal is to assign \( X \) to a specific class in \( C \).

Task Description

In this study, we classify posts and comments based on different features. Two tasks are proposed for post classification and comment classification, respectively: classification of user intentions, and classification of social support types.

With qualitative analysis in our previous study, five themes were extracted from messages of QuitStop forum, which are offering social support, requesting social support, receiving social support, other activities and irrelevant content. These themes reflect user intentions to publish corresponding posts or comments. These five categories are used for the classification of user intention in this study.

For classification of social support types, two categories are developed, which are informational support and nurturant support. According to previous definitions, informational support is specific information about the disease, treatment or coping (K. Y. Chuang & Yang, 2011; Cutrona & Suhr, 1992), including subcategories of advice, referral, fact, perceptual knowledge, personal experiences and feedback/opinion. Nurturant support is expressing caring or concern, and expressing the importance of relationship, including esteem, network and emotional support. These two types of social support are not exclusive. A message could be assigned to both informational support and nurturant support.

Feature Description

In this study, we build classifiers with different types of feature sets, and combine various classifiers to achieve the best results. Two types of feature sets are used for classification: text feature sets and health feature sets of users.
I. Text Feature Set

Usually, message classification is based on text features. On QuitNet Forum, a thread is consisted of texts at different positions. We built classifiers with text features at different thread positions, i.e. title, post, and comments, and combine different features linearly. Title refers to the title of the corresponding thread that is created by the post author. Post refers to content of the post in the corresponding thread created by the post author. In post classification, comment refers all comments in the corresponding thread, which may be published by different comment authors. In comment classification, comment is the message content to be classified, which is created by one comment author.

To build text feature sets, we first preprocess raw text in different positions by discarding non-alphabetic content, removing general stop words, stemming and lemmatizing with WordNet database. Then, term features are extracted and transformed to term vectors. As a result, each of the text feature sets is composed of a bag-of-words. On average, there are 2.6 words in a thread title, 72.6 words in a post message, and 21.4 words in a comment.

II. Health Feature Set

In the previous sections, we find that there are certain relations between user quit status and their behavior to publish different types of messages. So, in this study, we use quit status and quit stage as health features to improve text classifications. Based on quit status and quit stage, different health feature sets are developed as shown in Table 4.1. PA status, PA stage, CA status and CA stage are the four health feature sets. But the details of their meanings vary in post classification and comment classification.
III. Evaluation Metrics

For a classification task, precision, recall and F1 score could be calculated for each class. They can evaluate the result of classifications. For a certain class, let $n$ be the number of records that are predicted in the class, $N$ denotes the number of records belonging to the class in ground truth, and $tp$ be the number of records that are predicted in the class correctly, the precision could be calculated as $P = \frac{tp}{n}$, the recall could be calculated as $R = \frac{tp}{N}$, and the F1 score is calculated as $F1 = \frac{2\cdot P \cdot R}{(P+R)}$.

For different tasks of post and comment classifications, we focus on the precision, recall and F1 score of certain classes for evaluation. For the task of intention classification, the classes of offering social support, requesting social support and receiving social support are important. In practice, people may not concern the precisions and recalls of the other two classes – other activities and irrelevant content because they are not relevant to the healthcare application. In the dataset, there is no post receiving social support, while there is no comment requesting social support. So, in this study, for the intention classification of posts, we only focus on classes of offering social support and requesting social support, and the average precision, recall and F1 score of the two classes are used for evaluation. For the intention classification of comments, the...
average precision, recall and F1 score of offering social support and receiving social support are selected for evaluation.

To classify the types of social support for both posts and comments, the classes of informational support and nurturant support are both important. We calculate the average values of precision, recall and F1 score respectively, and use them to evaluate the results of classifications.

4.1.2 Approaches
In this study, we randomly collect 375 threads on QuitStop in periods of 05/01/2011 - 05/31/2011 and 07/01/2013 – 07/31/2013, which include 375 posts and 1365 comments. The messages are manually classified as gold standard. The approaches of manual coding is introduced Chapter 3.

We classify posts and comments respectively, and experiments are conducted for both post classification and comment classification. For each of classifications, 80% of all the records were extracted as training data and others are used as test data. The training data and test data are randomly generated five times, and the results of their average performances in the experiments are reported.

(1) Main Process to Develop Classifiers
For post and comment classifications, we build classifiers on different feature sets, and combine different classifiers to get the optimized classification results. Separate experiments are designed to optimize precision, recall and F1 score of certain classes as mentioned above.

Figure 4.2 summarizes our process to develop and combine different classifiers. For each of the task, the posts and comments are classified separately. Precision, recall and F1 score are used as evaluation indicators, and they are regarded as optimization goals in different experiments. For each experiment, we first develop different text classifiers on title, post and comment feature sets, respectively. Then, the three classifiers are linearly combined, and the combination weights are
calculated to optimize the goal indicator (precision, recall or F1 score). From the four text classifiers, which are built on title, post, comment and combination of the three, we select the best classifier with the highest evaluation indicator value. Based on this best text classifier, different health feature sets are added to boost the text classification. In the dataset, we cannot get the health features of all users. Messages without proper health features would not be boosted. For these messages, only the text classification is used. The text classification is boosted by adding four different feature sets respectively, and the best result is selected. To combine text and health classifications, we also develop algorithms to look for the efficient combination weights to optimize the goal evaluation indicator.

Figure 4.2 Main Process to Develop Classifiers
(2) Classification Approaches

In this study, all classifiers are built on Naïve Bayes classification. For a classification task, there are k classes denoted as c₁, c₂, ..., cₖ. A massage could be represented as a set of features X = (x₁, x₂, ..., xₙ). The probability of message X in class cᵢ is calculated as:

\[
P(cᵢ|X) = \frac{P(X|cᵢ)P(cᵢ)}{P(X)} \tag{4.1}
\]

Assuming that features in the set are independent, the probability could be calculated by:

\[
P(cᵢ|X) = \prod_{j=1}^{n} P(x_j|cᵢ) \tag{4.2}
\]

The likelihood \(P(x_j|cᵢ)\) is estimated with training data. For text feature sets and the feature sets of quit stages, \(x_j\) is a categorical value in multinomial distribution. In this case, the likelihood is calculated as:

\[
P(x_j|cᵢ) = \frac{N_{ji}}{N_j} \tag{4.3}
\]

where \(N_{ji}\) is the number of messages with feature \(x_j\) in class \(cᵢ\) in the training data, and \(N_j\) is the number of all messages with feature \(x_j\) in the training data.

For the feature of quit status, \(x_j\) a continuous value, which is assumed in Gaussian distribution. The likelihood is calculated as:

\[
P(x_j|cᵢ) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( -\frac{(x_j-\mu_i)^2}{2\sigma_i^2} \right) \tag{4.4}
\]
where $\mu_i$ and $\sigma_i$ are the mean and the standard deviation of the feature value for all messages in the training set.

To classify a message $X$ in the test data, $P(c_i|X)$ is calculated according to formula 4.2 for all the classes.

For the task of intention classification, different classes are exclusive. We develop one classifier for the five intention classes, and each message in the test data is assigned to the class with the highest probability of $P(c_i|X)$.

However, for the task of classification of social support types, the classes of informational support and nurturant support are not exclusive. For this task, we develop binary classifiers for the two classes separately. For a certain type of social support (informational support or nurturant support), let $c_0$ represents that the message is not related to corresponding type of social support, and $c_1$ represents that the message is related to the social support type, a message $X$ would be assigned to the class of social support type if $P(c_1|X) > P(c_0|X)$.

(3) Classifier Combination

As shown in Figure 4.2, some classifiers are combination of other classifiers. For example, Classifier 4 is the combination of three text classifiers, which are Classifier 1, Classifier 2 and Classifier 3. Each of Classifier 9 to Classifier 12 combines a text classifier and a health classifier.

Concretely, to combine $n$ classifiers, let $P_j(c_i|X)$ denotes the probability of message $X$ belonging to class $c_i$ calculated by the $j$th classifier, the newly combined classifier computes the logarithms of probabilities, and linearly combine them by:
\[ L(c_i|X) = \sum_{j=1}^{n} w_j \times \log(P_j(c_i|X)) \quad (4.5) \]

where \( w_j \) is the weight for the \( j \)th classifier for combination, and \( \log(V) \) is the logarithm value of \( V \). For each of the \( k \) classes, \( L(c_i|X) \) is calculated by formula 4.5, and the message \( X \) is assigned to the class with the highest value of \( L(c_i|X) \).

In this study, we specify \( w_j \) as a float between 0 and 1. To look for the best combination weights for the classifiers, we develop a genetic algorithm to optimize goal indicators of precision, recall and F1 score separately in experiments. The weight vector \( W = (w_1, w_2, ..., w_n) \) is calculated in each experiment. For each process of genetic algorithm, 10 different weight vectors of \( W \) are randomly generated as the initial population. Then, the population is expanded iteratively toward the goal of increasing the goal evaluation indicator. There are 100 generations, and the weight vector that achieves the highest goal indicator is selected. In the genetic algorithm, BLX-\( \alpha \) technique is used for crossover with the crossover probability of 0.75. Gaussian mutation technique is used with the mutation probability of 0.015.

4.1.3 Results and Discussions

(1) Classification of Posts

I. Classification Task of Intention

As mentioned above, we only focus on the classes of offering social support and requesting social support in this task for post classification. All the precisions, recalls and F1 scores reported in this task are the average values of that of the two classes. We first develop classifiers with text feature sets of title, post and comment separately. The classifier using title feature set reaches the highest precision, recall and F1 score, which are 0.554, 0.560 and 0.506 respectively. Then, we combine different text feature sets, and utilize health feature sets to boost the classifications. Experiments are carried out with goals to increase precision, recall and F1 score, respectively.
Figure 4.3 (a) depicts the result of experiments with the goal of improving precision. For the text-only classification, the classifier that combines title, post and comment reaches the highest precision. Through an optimization process, the combination weights of title, post and comment feature sets are 0.700, 0.602 and 0.085, respectively. Then, we use different health feature sets to boost the text classification, and found that adding CA status (the mean of quit statuses of comment authors) can improve the precision. In the experiments, the highest precision is achieved when combining classifiers of title, post, comment and CA status. However, the recall and F1 score cannot be improved during the process.

Figure 4.3 (b) is the result of experiments that try to improve recall. For the text-only classification, the classifier only with title feature set achieves the highest recall. Combining different text feature sets could not outperform the title-only classification. To improve recall of the title-based classifier, adding PA status (the quit status of the post author) reaches the highest recall. From the bars in Figure 4.3 (b), it is indicated that during the process of optimizing recall, the precision and F1 score are also improved accordantly.

Figure 4.3 (c) is the result of experiments with the goal of optimizing F1 score. Among all the text-only classifiers, the one with the title feature set reaches the highest F1 score. We use different health feature sets to boost the title-based classifier, and finds that the classifier that combines PA status and title text performs the best, which is the same as that of experiments of optimizing recall.

Analyzing all classifiers developed in our experiments for this task, the classifier that reaches the highest precision (0.715) is the one that combines feature sets of title, post, comment and CA status with weights of 0.678, 0.634, 0.362 and 0.711, respectively. The highest recall and F1 score are 0.719 and 0.636, and they are achieved by the classifier that combines features of title
and PA status with weights of 0.922 and 0.304. So, for post classification of intention, title is the most important feature that reaches the highest recall in text-only classification. Combining title with post content can achieve high precision. Adding health feature sets, including quit statuses of post author and comment authors, can help improving precision, recall and F1 score.

![Figure 4.3 Results of Post Classification of Intentions](image)

II. Classification Task of Social Support Types

Similarly, we first built classifiers on texts of title, post and comment separately. The precision, and F1 score can reach the highest value with post text, which are 0.894 and 0.589. The classifier built on title has the highest recall, which is 0.579. With goals of optimizing precision, recall and
F1 score respectively, a series of experiments are carried out to combine different text features and health features.

Taking precision as the optimization goal, the result of experiments is shown in Figure 4.4 (a). The classifier only with post text has the highest precision. Adding other text features or health features could not greatly improve the precision.

Figure 4.4 (b) is the result of experiments that optimize recall. For the text-only classification, the classifier combining the three text feature sets performs better than the other classifiers. The optimized combination weights for title, post, comment are 0.975, 0.038 and 0.108, respectively. Using health features can improve the text-based classification, and adding PA status reaches the highest recall in the experiments.

Figure 4.4 (c) is the result of experiments with the goal of optimizing F1 score. For text-only classification, using only post feature achieves the highest F1 score. Different health feature sets are added to post classification. It is found that the classifier with PA stage performs the best.

For all the classifiers, the one only with the post feature set reaches the highest precision of 0.894. The classifier that combines post and PA stage has the highest recall (0.645) and F1 score (0.692). The combination weights for post and PA stage are 0.155 and 0.997.

For post classification of social support types, post content is the most important text feature to reach high precision in classifications. Title is more helpful to achieve high recall. Adding quit stage of post author can improve recall and F1 score.
For the intention classification of comments, we only concern the classes of requesting social support and receiving social support. All the precisions, recalls and F1 scores reported are calculated on these two classes. We started from developing classifiers with text feature sets of title, post and comment respectively. The classifier using post feature has the highest precision and F1 score, which are 0.834 and 0.735, respectively. The classifier with title feature set reaches the highest recall of 0.711.
Figure 4.5 (a), (b) and (c) depict the results of experiments to optimize precision, recall and F1 score, respectively. Combining the text feature sets of title, post and comment can achieve the highest precision, recall and F1 score. But none of the health feature sets can outperform the combined text classifier. The highest precision is 0.859, and it is reached by the classifier combining title, post and comment with the weights of 0.030, 0.675 and 0.626. The highest recall is 0.721. It is achieved by the classifier that combines title, post and comment with the weights of 0.491, 0.620 and 0.014, respectively. The classifier with combination weights of 0.267, 0.360 and 0.083 achieves the highest F1 score of 0.743.

Analyzing the combination weights of different text feature sets, it could be concluded that the post content is important to indicate the intention of comments. Adding comment content is helpful to improve precision, while adding title can help to improve recall and F1 score.
II. Classification Task of Social Support Types

Text-only classifiers are built on feature sets of title, post and comment, respectively. The classifier with title feature has the highest precision of 0.810, and the classifier using comment feature reaches the highest recall and F1 score of 0.971 and 0.864 respectively.

Figure 4.6 (a), (b) and (c) show results of experiments that optimize precision, recall and F1 score. Combining text features of title, post and comment can improve precision and F1 score, but it cannot greatly improve recall. Adding neither of the health feature sets could outperform text-only classifiers. From our experiments, the highest precision and F1 score are is 0.816 and 0.874 respectively, and they are achieved by the classifier that combines title, post and comment with weights of 0.896, 0.063 and 0.600. The highest recall is 0.971, which is reached by the classifier with comment feature.

Observing text classifiers that reach high precision, recall and F1 score, we can conclude that comment content is important for the classification. Adding title with a high weight could greatly improve recall as well as F1 score, but it cannot improve precision.
Techniques of classification and optimization are applied to classify posts and comments on QuitNet Forum. Different text feature sets and health feature sets are used to build classifiers. The classification results are analyzed and compared. Summarizing all experiments in this study, it is observed that:

1. For a thread on QuitNet Forum, the content created by the post author (title and post text) can indicate the intention of both post and comments in that thread. For intention tasks of both post and comment classifications, the optimization process assigns high weights to the feature sets of title and post.

4.1.4 Conclusion
Techniques of classification and optimization are applied to classify posts and comments on QuitNet Forum. Different text feature sets and health feature sets are used to build classifiers. The classification results are analyzed and compared. Summarizing all experiments in this study, it is observed that:
(2) For the classification of social support types, the message content is an important feature set. Concretely, post content is effective for post classification, and comment content is effective for comment classification.

(3) Health feature sets are effective to improve post classifications of both intentions and social support types. However, none of the health features sets can improve comment classifications substantially.

In the next section, we develop recommender systems to help users find proper discussion topics and threads. The classification results are applied in recommendation models to achieve better prediction result.

### 4.2 Thread Recommendation for Users

Due to a large amount of user generated content, information overload is a problem for online communities. To help smoking quitters find useful information and reach helpful peers, it is necessary to recommend them proper threads or discussion topics to participate in. As introduced above, recommendation systems are widely used in E-commerce and other areas. In this section, we apply recommendation techniques to recommend threads for users in QuitNet Forum.

#### 4.2.1 Problem Description

In a dataset of $N_T$ threads and $N_U$ users, our goal is to recommend threads for each user to participate in (comment on). For a user $u$ commenting on the thread $t$, we record the pair $< t, u >$ in the dataset. A set of thread-user pairs could be constructed, which is denoted as $TU$. $TU$ is randomly divided into a training set $\overline{TU}$ and a test set $\overline{TU}$, such that $\overline{TU} \cap \overline{TU} = \emptyset$ and $\overline{TU} \cup \overline{TU} = TU$. Given $TU$ known, for a user $i$ in the dataset, we recommend him/her a thread set $T_i$ with the size of $k$. For a thread $j \in T_i$, $j \notin TU$. The Top-K recall in $\overline{TU}$ is used to evaluate the result. The recall of $T_i$ for user $i$ is calculated as

$$\frac{|\{< j, i > | j \in T_i \text{ and } < j, i > \notin \overline{TU}\}|}{k}$$
4.2.2 Approaches

We collect all posts and comments on QuitStop during 05/01/2011 - 05/31/2011 and 07/01/2013 – 07/31/2013. There are 5061 threads, 34269 comments and 1327 users collected. Let \( p = \frac{|TU|}{|TU|} \), which is the percentage of test set. We set \( p \) as 30%, 50% and 70% respectively, and for each \( p \) value, we randomly generate five training and test sets for the experiments. For each experiment, the average result of the five sets is reported.

(1) Collaborative Filtering and Content-based Method

Collaborative Filtering (CF) and content-based method are traditional approaches for recommendation systems. They are widely developed and applied in different areas. In our study, we first apply the traditional approaches for thread recommendation. The basic idea is that we calculate the similarities of each pair of user and thread in the dataset. For each of the user, the top \( t \) threads with the highest similarities with the user are recommended to him/her. First, the similarities between users and threads are calculated by neighbor-based collaborative filtering. A probability matrix is built for each pair of users and threads for thread recommendation. Then, the title texts of threads are used as content-based methods to improve the similarity matrix of users and threads.

Neighbor-based collaborative filtering is the most basic technique for recommendation. In our experiment, let \( A \) be the a matrix with the size of \( N_U \times N_T \), where \( N_U \) is the number of users in the dataset, and \( N_T \) is the number of threads in the dataset. \( A \) is a binary matrix indicating the participating activities in the training set.

\[
A_{ut} = \begin{cases} 
1, & \text{user } u \text{ comment on thread } t \\
0, & \text{otherwise} 
\end{cases}
\]  

(4.6)
ST is a matrix indicating similarities between every two threads. It could be inferred from A. The similarity between threads \( t_1 \) and \( t_2 \) is calculated by

\[
ST_{t_1 t_2} = \frac{\sum_{i=1}^{NU} A_{i t_1} \times A_{i t_2}}{\sum_{i=1}^{NU} A_{i t_1} + \sum_{i=1}^{NU} A_{i t_2} - \sum_{i=1}^{NU} A_{i t_1} \times A_{i t_2}} \quad (4.7)
\]

For all users in the dataset, we calculate the probability matrix PC based on the CF information. It indicates similarities between users and threads, which could be used to predict the probabilities of users’ participation in threads. The probability of user \( u \) participating in thread \( t \) is:

\[
PC_{ut} = \frac{\sum_{j=1}^{NT} A_{uj} \times ST_{tj}}{\sum_{j=1}^{NT} A_{uj}} \quad (4.8)
\]

The probability matrix PC can be used for recommendation and prediction. To recommend threads for a user \( u \), we select top K threads from PC with the highest probabilities.

For content-based approaches, it is assumed that user preference can be captured by a content profile and a user might be interested in a thread if this user’s content profile is similar enough to the content of a thread. According to Section 4.1, title text is helpful for post classifications of both tasks of user intentions and social support types. It can indicate the content of threads in terms of social support. Thus, to apply content-based methods in the recommendation system, we use terms in thread titles. The raw title text is preprocessed by discarding non-alphabetic content, removing general stop words, stemming and lemmatizing with WordNet database.

To construct user \( u \)’s content profile, we extract all threads in the training set that \( u \) participates in, and aggregate them into a term vector \( C_u = < c_{1u}^u, c_{2u}^u, ..., c_{nu}^u > \), where \( c_{i}^u \) denotes the frequency percentage of term \( w_i \) in all thread titles in the training set that user \( u \) has participated in. Given a
thread \( t \), it is also represented by a term frequency vector \( C^t = \langle c^t_1, c^t_2, ..., c^t_n \rangle \), where \( c^t_i \) is the frequency percentage of term \( c_i \) in the title of thread \( t \). Note that \( C^u \) and \( C^t \) are normalized, and each item in the vectors is a percentage value. The Cosine similarity between user \( u \) and thread \( t \) can be calculated as:

\[
\text{CSim}(C^u, C^t) = \frac{\sum_{i=1}^{n} c^u_i \times c^t_i}{\sqrt{\sum_{i=1}^{n} (c^u_i)^2} \times \sqrt{\sum_{i=1}^{n} (c^t_i)^2}} \tag{4.9}
\]

Considering the content-based method to improve the probability matrix in formula 4.8, the probability of user \( u \) participating in thread \( t \) is:

\[
P_{\text{Cut}} = W_{\text{CF}} \times \frac{\sum_{j=1}^{N_T} A_{uj} \times ST_{tj}}{\sum_{j=1}^{N_T} A_{uj}} + W_{\text{Content}} \times \text{CSim}(C^u, C^t) \tag{4.10}
\]

where \( W_{\text{CF}} \) and \( W_{\text{Content}} \) are weights to combine CF and content-based methods. \( W_{\text{CF}} \geq 0, W_{\text{Content}} \geq 0 \) and \( W_{\text{CF}} + W_{\text{Content}} = 1 \). \( PC \) is a basic prediction matrix that could be calculated by formula 4.8 with CF information or formula 4.10 with both CF and text content.

Recommendation based on the prediction matrices \( PC \) is used as baselines in this study. \( PC \) can be calculated by formula 4.8 that is based on pure CF information. It can be also calculated by formula 4.10 that is a hybrid approach combining CF and content information.

(2) Integration of Classification Result

In 4.1, we propose algorithms to classify messages on QuitNet Forum from the perspectives of user intentions and social support types. It is supposed that each user has a preference to different categories of posts, which decides whether he/she will take part in the discussion of the thread.
According to formula 4.5, the logarithm of probabilities of the post belonging to each category could be calculated.

There are five categories of user intentions, including offering social support, requesting social support, receiving social support, other activities of smoking cessation and irrelevant content. For a thread $t$, we can construct a normalized probability vector $L^t = [l^t_1, l^t_2, ..., l^t_5]$, where $l^t_j$ denotes the normalized logarithm of probability that $t$ is in the $j^{th}$ category of user intentions. From the previous section, it is found that title text and the poster’s quit stage are important features for user intention classification of posts. We choose these two features sets in formula 4.10 with weights of 0.8 and 0.2 to calculate $l^t_j$. Note that $L^t$ is normalized that the sum of all elements in $L^t$ is equal to 1.

For a user $u$, we extract all threads in the training set that $u$ participates in, and construct a normalized vector $L^u = [l^u_1, l^u_2, ..., l^u_5]$, where $L^u = \frac{\sum^{N_T}_{j=1} A_{uj} \times L^t_j}{\sum^{N_T}_{j=1} A_{uj}}$. The preference of $u$ to the $j^{th}$ category of user intentions is calculated by the average value of the probabilities of all posts that $u$ participates in. If $u$ does not participate in any threads in the training set, all elements in $L^u$ are set 0.

The classification of user intentions is integrated to improve the prediction matrix $PC$ in formula 3 and formula 5, which leads to a comprehensive prediction matrix $P$.

$$P_{ut} = W_{pc} \times PC_{ut} + W_{li} \times CSim(L^u, L^t) \quad (4.11)$$

Where $W_{pc}$ and $W_{li}$ are weights above 0 to integrate classification result with the baseline method.
Towards types of social support, there are two categories of the posts, which are informational support and nurturant support. These two types of social support are not exclusive. In the classification section, we develop binary classifiers for them respectively. For the thread $t$ and a certain type of social support, we can calculate two logarithms of probabilities, which indicates whether $t$ contains that social support or not. The calculation of the logarithms is based on formula 4.10. We choose features sets of the post title and the poster’s quit stage with weights of 0.9 and 0.1. The binary logarithms are normalized to make the sum of them equal to 1. For the thread $t$, we construct a logarithm vector $LS^t = \langle ls^t_i, ls^t_n \rangle$, where $ls^t_i$ is the normalized logarithm probability of $t$ containing informational support, and $ls^t_n$ is the normalized logarithm probability of $t$ containing nurturant support. Note that the sum of elements in $LS^t$ is not necessary to be 1, because the two logarithm values are normalized for different types of social support respectively.

For a user $u$, we extract all threads in the training set that $u$ participates in, and construct a vector $LS^u = \langle ls^u_i, ls^u_n \rangle$, where $LS^u = \frac{\sum_{j=1}^{NT} A_{uj} \times LS^j}{\sum_{j=1}^{NT} A_{uj}}$. The preference of $u$ to each type of social support elements is calculated by the average value of probabilities of all posts that $u$ participates in. If $u$ does not participate in any threads in the training set, all in $LS^u$ are set 0.

Similarly, the classification result of social support types is used to improve $PC$. In this case, the prediction matrix is proposed:

$$P_{ut} = W_{pc} \times PC_{ut} + W_{ls} \times CSim(LS^u, LS^t) \quad (4.12)$$

Where $W_{pc}$ and $W_{ls}$ are weights above 0.

(3) Integrating Users’ Quit Stages
As analyzed in Section 4.1, users’ quit stages could be used to improve classifications of posts. In this approach, we directly integrate users’ quit stages to the prediction matrix for recommendation, instead of using classification results. For a user \( u \), we suppose that he/she has different preferences to threads with authors at different quit stages. For a thread \( t \), it is supposed that users at the same quit stage have the similar probability to participate in it. Thus, the quit stages of authors and participants of the threads in the training data are used for recommendation.

For a user \( u \), and a thread \( t \) whose author is at the quit stage of \( Q_t \), the probability of \( u \) participating in \( t \) is calculated by observing the user’s participating to threads with authors at the same quit stage. The prediction matrix based on thread authors’ quit stages \( PA \) is calculated by:

\[
PA_{ut} = \frac{\sum_{i \in Q_t} A_{ui}}{\sum_{j=1}^{N_t} A_{uj}} \quad (4.13)
\]

Similarly, a thread may attract participants at the similar quit stages to participate in it. For a user \( u \) at the quit stage of \( Q_u \), the probability of \( u \) participating in thread \( t \) is calculated by observing the thread’s existing participants that can be learned from the training set. The prediction matrix based on participants’ quit stages \( PP \) is proposed as:

\[
PP_{ut} = \frac{\sum_{i \in Q_u} A_{it}}{\sum_{j=1}^{N_t} A_{jt}} \quad (4.14)
\]

PA and PP are prediction matrix developed from users’ health status. They are integrated into prediction matrix \( PC \) in formula 4.8 and 4.10, and the final prediction matrix \( P \) can be calculated as:

\[
P_{ut} = w_1 \times PC_{ut} + w_2 \times PA_{ut} + w_3 \times PP_{ut} \quad (4.15)
\]
4.2.3 Experiments and Results

CF and content-based approaches are widely used in different areas. They are used as two baselines in our experiments. The pure CF method is applied based on formula 4.8, and the hybrid approach of CF and content is based on formula 4.10. In this section, we design experiments to investigate whether classifications could be used to improve baseline methods. Also, we propose experiments to integrate user quit stages for recommendations.

(1) Integrating Classification Result

First, we use classification result to boost the CF method. To implement formula 4.11 and 4.12, PCut is calculated by formula 4.8, which is Collaborative Filtering, not considering content. Classifications of intentions and social support types are integrated with formula 4.11 and 4.12, respectively. For formula 4.11, $W_{pc}$ is set as 15, and $W_{li}$ is set as 1. Similarly for formula 4.12, $W_{pc}$ is set as 5, and $W_{ls}$ is set as 1. As mentioned earlier, we generate the test set and training set at different percentages. Figure 4.7 shows the Top-K Recalls with the baseline of basic collaborative filtering.
From Figure 4.7, when the test set includes a small percentage of data (Figure 4.7 (a), p=30%), integrating classification results cannot improve CF method. However, with a higher percentage of test data (Figure 4.7 (c), p=70%), integrating classification results could apparently improve
the recommendation. With the bigger test data, there is less information in the training data set. It indicates that when a thread is initiated, and there are few participants known, we can utilize classification results to improve recommendation. After more people participate in the thread and we get enough information, the basic CF can work effectively for the recommendation.

Besides basic CF, the hybrid method that combines CF and content-based approach is used as the baseline. Formula 4.10 is used to implement formula 4.11 and 4.12, considering both collaborative filtering and content. The Top-K Recalls are shown in Figure 4.8.
Figure 4.8 Improving Hybrid Method (CF and content-based) with Classification Results - Top-N Recall with Different Percentages
Figure 4.8 presents the results with the hybrid baseline of CF and content-based approach. The classification results are used to boost the hybrid method. It can be learned that when the test set is small (Figure 4.8 (a), p=30%), content information cannot improve CF method, and the hybrid method is not advantageous. From Figure 4.7 (a), using only classification result cannot improve the basic CF either. However, integrating classification result in the hybrid method can greatly improve the recommendation result as shown in Figure 4.8 (a). It indicates that combining both content information and classification information can greatly boost the recommendation.

When less information is known and there is a higher percentage of test data (Figure 4.8 (a), p=70%), the hybrid method of CF and content can greatly improve the basic CF approach. Integrating classification of social support types can slightly improve the hybrid baseline, but their results are similar.

From Figure 4.7 and Figure 4.8, integrating classification results in the hybrid method can always achieve the best recommendation recall with different percentages of test data. However, its vantage differs with different percentages of test data. When we have known most information and the test data is small, the combination of classification and the hybrid method can greatly outperform other approaches. Classifications of intentions and social support types achieve similar results. In practice, the integrated approach could be directly applied, no matter how much information is known. It could always achieve the best result.

(2) Integrating Users’ Quit Stages

In our approaches, classifications are based on title content and quit stages of thread authors. Here, quit stages of thread authors and participants are directly used to improve the baselines. To implement Formula 4.15, we choose the basic CF method as the baseline, and different weights are set to compare the results. There are three weights in Formula 4.15. \( W_1 \) represents the weight
of the prediction matrix based on CF method, $W_2$ is the weight of the prediction matrix based on thread authors’ quit stages, and $W_3$ is the weight of the matrix based on thread commenters’ quit stages. We set $W = (W_1, W_2, W_3)$, and the results for different percentages of test sets are shown in Figure 4.9.
Figure 4.9 Improving basic CF with Users’ Quit Stages - Top-N Recall with Different Percentages (p) of Test Sets
When \( W = (1,0,0) \), it is the basic CF model for prediction. From Figure 4.9, assigning a positive value to \( W_2 \) can improve the baseline. Different values of \( W_2 \) (1 or 5) lead to similar results. However, assigning a positive value to \( W_3 \) will lead to a worse result. It means that the prediction matrix based on thread authors’ quit stages can boost the recommendation. But on the contrary, the prediction matrix based on other participants’ quit stages cannot improve the recommendation.

Integrating quit stages of thread authors is effective to improve the basic CF methods. We also apply prediction matrix of authors’ quit stages to improve the hybrid baseline. Concretely, to implement formula 4.15, we use formula 4.10 to calculate \( PC_{ut} \), with \( W = (2,1,0) \). To compare the result with that of integrating classification result, we also use classifications of intentions and social support types to improve the hybrid baseline. All the results are shown in Figure 4.10.
It can be learned from Figure 4.10 that directly integrating authors’ quit stages can outperform
other methods and achieves the best recommendation result. It leads to similar recommendation result with integrating classification results, and performs even better to some degree. Integrating quit stages to the CF matrices is efficient, and it is easy to apply.

4.2.4 Conclusion
In this section, we build recommendation systems to recommend proper threads for users to participate in. Collaborative filtering and the hybrid method of CF and title content are used as baselines. The post classifications of user intentions and social support types are integrated to boost the baselines. It is found that integrating classification results could improve the baselines, but the rates of the improvement differ with different amounts of known information. Generally, combining classification results with the hybrid method can reach the highest recalls in the experiments with different percentages of test. Moreover, directly integrating user quit stages to the prediction matrices can reach similar results, or even perform better. This method is easy to apply, and it works well with different unknown information. It could be directly applied in practice with threads initiated at different times, not matter how many participants have participated in.

However, our research is carried out on a relatively small data set. A lot of potential information of the participating might be missing, so it is difficult for the evaluation. In the future, we could apply the methods to a large scale of data, and develop better services for the online health communities.
5. Conclusion

Online health communities are emerging communication channels that have great potentials to benefit patients and health consumers. This dissertation studies smoking cessation communities from different aspects, and the framework and approaches could be applied to many other health issues. For online health communities, most of the participants have health problems or concerns, and it is important to take their health statuses into consideration. The health statuses are important features that may influence users’ interests, behaviors and interactions in online communities. Also, the changes of health statuses are factors reflecting the outcomes of health intervention, which can be used to evaluate online health intervention programs.

Currently, there are many studies on online health communities, however, most of them are preliminary and discontinuous. In this dissertation, we focus on QuitNet organization, and investigate different research problems based on the same dataset. To carry out a comprehensive study on this emerging area, our research includes three parts: a preliminary study, an in-depth analytics, and an application development.

Chapter 2 compares smoking cessation communities that are built on different social media. The technical development of different social media are geared for healthcare communications, but it is not clear from related literature what makes for a better website in terms of social support in smoking cessation. In this chapter, we compare user characteristics and behaviors of QuitNet Forum and QuitNet Facebook. It is found that most users of QuitNet Forum are at early quitting stage. They participate in the discussions actively and frequently with great passions. But on the contrary, QuitNet Facebook attracts various users with different quit statuses, and their discussions and performances are not that active. We discuss possible reasons for the differences of these two communities. To our knowledge, it is the first study that compares health communities through different social media channels. With this knowledge, we can provide
different smoking cessation services on different types of social media, and achieve better intervention outcomes (Zhang, Yang, & Chuang, 2014; Zhang, Yang, & Li, 2012).

Through the preliminary study of different smoking cessation communities, we realize that QuitNet Forum is a centralized platform that attracts a lot of active smoking quitters with highly-qualified discussions. Thus we focus on QuitNet Forum to analyze the discussion content and user interactions of social support exchange. In Chapter 3, we use different approaches to analyze the discussion content on QuitNet Forum, and analyze user behaviors and interaction patterns. Social support is an important factor to help smokers to quit smoking. Many studies use content analysis to extract the topics and types of social support in different online health communities. Besides content analysis in our study, we also apply and develop the theory of user information behavior to identify user behavior patterns. Social networks are constructed to analyze interactions between people at different quit stages for different types of social support exchange. Thus, the analytics carried out in this dissertation investigate user interactions from a deeper level, which is beyond other studies of online health communities (Zhang & Yang, 2014c; Zhang, Yang, & Gong, 2013). It can help us get the all-sided knowledge to understand user interactions in QuitNet Forum. The proposed approaches can be easily applied to other health issues.

There are a large amount of data on QuitNet Forum and other online health communities. In most online communities, 90% of users are lurkers who never contribute, 9% of users contribute a little, and 1% of users account for almost all the actions (Nielsen, 2006). For QuitNet website, although more than 800,000 individuals had registered, only 7569 users participated in interactions during March 1, 2007 – April 30, 2007 (Cobb et al., 2010). It is important for an online community to deploy an effective recommender system to encourage users’ contributions and to make its users to stay longer. In Chapter 4, we apply data mining techniques to categorized threads on QuitNet Forum, and build recommendation systems for this community. Text feature sets and health
feature sets are constructed for thread classifications, and the most efficient classifiers are
developed with optimization approaches. The classification results and user quit stages are
integrated to Collaborative Filtering and content-based approaches to reach the best
recommendation efficiency (Zhang & Yang, 2014a, 2014b).

Compared to traditional face-to-face intervention programs, online smoking cessation programs
target at a large population, and the intervention outcomes are less effective for individuals. It is
important to improve the design and services of online intervention smoking cessation programs
to reach better intervention outcomes. In Chapter 2, we compare user health features and
behaviors on QuitNet Forum and QuitNet Facebook. Based on the result, different services could
be developed and provided on different social media channels. Smoking quitters at different quit
stages could be introduced to distinct channels to achieve better intervention outcomes. In
Chapter 3, we investigate user behavior of different quit stages, and analyze how people at
different quit stages support each other. We utilize this knowledge in Chapter 4 to develop
machine learning predictions to help people find proper topics and peers to interact with. Health
features and user behaviors are used as features to improve traditional machine learning
techniques. Classification and recommendation are carried out in our study. It helps us
automatically detect users’ discussion topics on QuitNet Forum, and recommend proper topics
and peers to different individuals. In this way, we can suggest and provide personalized services
for different individuals in online intervention programs of smoking cessation that could lead to
better outcomes.

In this dissertation, we find that quit status and quit stages play significant roles in different
aspects. They are important features to effect user behaviors and interactions. They are also
associated with the topics of user generated content in the communities. In the application part,
the health statuses are proved to be effective to improve classifications and recommendations. For
online communities of other health issues, users’ health statuses may be measured in different ways. But it is still important to involve health statuses to analytics and predictions. In the future, we will apply data mining techniques to a large data set or other health areas. The possible future research includes:

(1) Integrating language processing techniques. In the data mining applications, text is used as an important feature for both classification and recommendation. However, our approaches do not include advanced language processing procedure, which may reduce the accuracy for the predications. In online communities, languages used are usually loose and free. It is especially import to preprocess the language to get better information. For health issues, some medical vocabulary may be included. In the future work, we will include general and medical language processing techniques, and include emotion analysis to preprocess the raw text. It is supposed that better results would be achieved.

(2) Developing different predictions for online health communities. In this dissertation, we realize classification and recommendation for the online health communities. There are many other data mining approaches that could be applied to get useful information and make predictions for health communities. For example, using clustering and graph mining techniques, we could detect communities among users of the online health programs, and identify different user roles in the communities. In addition, association rule mining can be applied to look for relations between users, diseases, medicines or other entities. For all the prediction tasks, we can utilize patterns found in this dissertation, including user health statuses, discussion topics, types of social support, user behaviors and so on.

(3) Scaling to larger data sets. For the analytics and different applications, our ultimate goal is to improve online services and intervention programs. As there is a large amount of data in online communities, our analytics and applications should be applied to bigger
data. Also, it is important to develop user-end services to present the analytics results and provide better services. Visualization techniques and user interface design can be applied to deploy different applications.
List of References


Hong, L., & Davison, B. D. (2009). A Classification-based Approach to Question Answering in Discussion Boards. Paper presented at the the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09), Boston, Massachusetts, USA.


Intervention as Adjuvant Treatment in a Smoking Cessation Intervention. Nicotine Tob Res, 8(1), 59-67.


Zhang, M., & Yang, C. C. (2014c). Using content and network analysis to understand the social support exchange patterns and user behaviors of an online smoking cessation intervention program. Journal of the Association for Information Science and Technology.


VITA

Contact
Mi Zhang, mi.zhang999@gmail.com

Education
- Ph.D.: 2010/09 – 2015/03
  Information Studies, Drexel University
- Master of Science: 2008/09 – 2010/06
  Information Science, Nanjing University, P. R. China
- Bachelor of Science: 2004/09 – 2008/06
  Information Management and System, Wuhan University, P. R. China

Research Areas
- Data Mining, Machine Learning, Data Analysis
- Health Information
- Information Visualization

Work Experiences
- Research Assistant, Drexel University, 2010-2014
- Research Associate Intern, Analytics Lab, HP Labs, 2014/06 – 2014/10

Publications
- Zhang, M. & Yang, C.C. Using Content and Network Analysis to Understand the Social Support Exchange Patterns and User Behaviors of Online Smoking Cessation Intervention Program. Journal of the American Society for Information Science and Technology (JASIST). (Accepted)
- Zhang, M. & Yang, C. C. The Effectiveness of Smoking Cessation Intervention on Facebook: A Preliminary Study of Post and Users. Proceedings of International Conference of Smart Health (ICSH), Beijing, China, August 3-4, 2013.

Pending Patents:

Services
- Program Committee of SBP 2013, SBP 2014 and SBP 2015 (International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction)
- Web Chair of ICEC 2014 (The 16th International Conference on Electronic Commerce)