Classification of Online Health Discussions with Text and Health Feature Sets

Mi Zhang and Christopher C. Yang
College of Computing and Informatics, Drexel University
mi.zhang@drexel.edu, chris.yang@drexel.edu

Abstract

Nowadays, many health groups and forums are established on the Internet, where health consumers discuss health issues and interact with each other. Although there is a large amount of user generated content about healthcare on different social media sites, few studies have applied data mining or artificial intelligence techniques for knowledge discovery on a large scale of data in this particular emerging area. In online health forums, it is difficult for users to find relevant topics or peers due to the large amount of information. Traditional recommendation systems may not work well for health online forums, because health consumers have different intentions of participation or may be interest in different types of supports even if the content matches their interest. To help solving this problem, we apply Naïve Bayes methods in this study to classify posts and comments on QuitStop forum, which is an online community for smoking cessation intervention. Classifiers are built on different text features and health features of user quit status. Two different classification tasks are investigated: (1) classification of user intentions, and (2) classification of types of social support exchanged in interactions. We developed classifiers for posts and comments separately, and conducted experiments to compare classifiers with different text and health feature sets. It is found that using thread title or post content can achieve the highest classification accuracy on both posts and comments for user intention classification with text features. On the other hand, using the content of post or comment itself performs the best for the classification of social support types. In particular for the post, integrating health features of the post author can boost the text classifications of user intention and support type. However, user health features cannot help in improving text classifiers for the comments.

Introduction

With the development of Web 2.0, the concept of Health 2.0 emerges with a variety of features, including social networking, participation, apomediation, collaboration and openness (Eysenbach 2008; Belt, Engelen et al. 2010). Many online communities and social networking platforms are developed for people to discuss health issues and interact with each other. Fox et al. reported that "the social life of health information is robust", with the fact that 52% online health inquires involved interaction with others (Fox and Jones 2009).

In Web 2.0 era, user generated content in various social media sites provides a rich resource for knowledge discovery. Data mining techniques are applied to extract knowledge from the unstructured data (Rajman and Besançon 1998). In medical and healthcare areas, data mining is applied to formal biomedical records in many studies (Cohen and Hersh 2005; Saeys, Inza et al. 2007). Although a lot of online communities, including forums and discussion groups, are built for health discussions and user interactions, few studies focus on this emerging field for knowledge extraction and discovery. In this study, we extract user discussion content from a smoking cessation forum, QuitStop, and apply classification technology to classify messages according to user intentions and social support exchange types in interactions.

QuitStop is a forum on QuitNet website, which is one of the most popular websites for smoking cessation (http://www.quitnet.com/qnhomepage.aspx), where different intervention services are provided (An, Schillo et al. 2008). QuitNet has developed 11 Web forums. QuitStop (http://forums.quitnet.com/aspBanjo/Message_List.asp?Conference_ID=10&Forum_ID=8&r=100777) is the most popular one among them, on which users can discuss the tobacco quitting process, ask questions, and give or receive social support.

QuitStop arouses a large number of discussions every day. Usually, a thread can only stay on the first page for a very short time. So it is difficult for users to look for relevant topics to discuss, or identify proper peers to communicate with. It would be helpful if we could recommend interesting topics or predict potential users for QuitStop