Health Information Needs on Diseases: A Coding Schema Development for Analyzing Health Questions in Social Q&A

Sanghee Oh
School of Library & Information Studies
Florida State University
shoh@cci.fsu.edu

Yan Zhang
School of Information
University of Texas at Austin
yanz@ischool.utexas.edu

Min Sook Park
School of Library & Information Studies
Florida State University
mp11j@my.fsu.edu

ABSTRACT
Health-related questions users posted on social Q&A sites are a representation of consumers’ real-life needs for health information. To tap into this source for a deeper understanding of general people’s health information needs and information searching behavior, this poster reports an attempt to develop a coding schema for analyzing disease-related questions in social Q&A sites. The developed schema could serve as a basis for analyzing a wide range of topics in health. It could also guide the implementation of automatic data mining approaches to analyze the vast amount of data generated on social media platforms.

Keywords
Social Q&A, health information, Online information seeking.

INTRODUCTION
The purpose of the current project is to investigate the health information needs that the general public most likely seeks, by analyzing the content of disease-related questions in social Q&A. Social Q&A is an online service that allows people to ask questions about any topic and to receive answers from anyone accessing the service. It encourages people to bring up their intimate issues, to seek solutions and suggestions, and to share personal experiences and knowledge. Social Q&A is easily accessible by those who have suspicions or symptoms of a certain disease. People can obtain information from others for free in an open and anonymous environment.

One of the greatest advantages of using social Q&A for information seeking is that people can elaborate on their information needs with their own words, explaining their situations and conditions with as much (or as little) detail as they wish. All of the disease-related questions in social Q&A are self-reported real life problems. People voluntarily post questions to obtain information and the number of questions collected in social Q&A services is substantial. For example, in Yahoo! Answers, the most frequently used social Q&A service in the U.S., as of May 2012, about 97,598,974 health-related questions have been resolved since the service was launched in 2005 and are available for browsing and searching. Therefore, social Q&A can be an essential venue for observing users’ natural behaviors of information seeking by making available an extensive collection of questions which represent authentic everyday-life information needs.

In the current study, four research questions were proposed in order to identify disease-related information needs from health questions in social Q&A:

1) What kinds of health information do people describe in their questions in order to obtain answers?
2) What are the conditions and symptoms pertaining to diseases that concern people?
3) What are the disease specific information, such as prevention, causes, diagnosis, and treatment that people want to know?
4) How do people’s information needs on a disease change based on the progression of their conditions?

People’s information inquiries expressed in questions are often complex, containing multiple concepts and problems that are correlated to one another in a personal situation. Therefore, content analysis was used as a research method in the current project, as the method allows human coders to understand the complex nature of the questions and review the questions thoroughly.
This poster reports on a coding schema that has been developed for analyzing the content of questions. The authors conducted a primary review of questions about STDs and cancer (200 each). From a thorough review of the questions about the two types of diseases, a set of coding schemas were developed with two variations pertaining to each type of disease. Findings from the current poster will be used for conducting the next stage of the studies which will investigate information needs on each type of disease by assigning the codes to questions, measuring the frequencies of coding results, and analyzing the nature and context of the questions.

LITERATURE REVIEW
Questions posted on social Q&A sites are representations of authentic user information needs. To uncover real-life user needs for music information, Lee (2010) analyzed more than 2000 questions posted on Google Answers, a fee-based service where users post questions and search experts hired by Google to answer the questions. A taxonomy of user needs and information features used to establish the needs were identified as a result of the analysis. Similarly, Yoon and Chung (2011) analyzed questions posted on Yahoo! Answers to identify ordinary people’s image needs in daily life. The authors suggested that the natural language questions contained rich information about the context of image needs, image attributes, and associated information or stories.

Researchers also analyzed health-related questions posted on social Q&A sites to identify health-related information needs and information searching behavior. For example, by analyzing questions posted on Yahoo! Answers, Zhang (2010) identified a set of contextual factors, including user goals, motivations, emotions, and time, to characterize consumer health information needs. Kim, Pinkerton, and Ganesh (2011) analyzed both questions and answers from Yahoo! Answers and identified major topics of interest to the general public during the H1N1 outbreak, including general health, flu-specific information, and nonmedical-related issues.

METHODS
A total of 400 questions about cancer and STDs (200 each) posted during 2011 were randomly selected from a pool of health questions in the STDs and cancer categories of Yahoo! Answers. The topics of STDs and cancer were chosen because of their comparative nature in health. STDs is one of the most common topics on which young adults seek information on the Internet (Gray & Klein, 2006), and people may seek due to their own concerns about prevention or transmission of the disease. On the other hand, cancer is a topic which could be a concern of anyone who is suspicious about or diagnosed with cancer. Not only patients but also caregivers would like to obtain information with which to help patients close to them, and they would seek on a wide range of health information, such as symptoms, treatments or prognoses of various kinds of cancers.

Since there was no appropriate coding schema applicable to the health information needs as described in the questions, a bottom-up approach was adopted. A manual review of the questions was conducted by reading them one by one and identifying the topics, concepts, inquiries, or expressions that people used in their questions. The authors reviewed the questions, identified potential categories of information, compared their categories with others, and developed the coding schema through an iterative process of discussion and then making decisions on the coding categories.

RESULTS & DISCUSSION
At the beginning of the coding schema development process, the authors used broad levels of categories to represent the health categories, since they had had previous experiences of research in social Q&A (Oh, Yi; Worrall, under review; Zhang, 2010). These categories included types of diseases, symptoms, treatments, and risk factors. Soon, however, it was realized from the in-depth review that the existing codes were not capable of uncovering the comprehensiveness of the information that people have shared in their questions in order to obtain decent answers to their questions. Thus, based on the review, a schema that better represents users’ health information needs and information seeking behavior on Social Q&A was developed, as shown in Table 1.

In social Q&A, not only patients or potential patients who are suspicious about their health conditions but also caregivers ask questions for their family or friends. Therefore, we specified the demographic information of patients (or potential patients) and caregivers in order to see if there are differences in the questions asked by the two groups. One of the main approaches that we considered for the development of the coding schema was that there are two types of information that people described in their questions, 1) information provided, and 2) information asked. In most cases, people provided a short description of their health conditions and problems and then asked for information about prevention, symptoms, diagnoses, or treatment of their diseases. They also expressed not only information-related concerns in their questions, but also revealed their emotional status in order to receive social support from those who would provide answers.

Overall, both groups of people who ask questions about STDs or cancer developed a similar pattern of questions as shown in the coding schema; but the nature of the information exposed in their questions was different. People who ask cancer questions provided information about health conditions, including food, diet, exercise, and smoking or drinking habits, or medical history of their own as well as their family members; while people who ask STD questions mainly described their sexual relationships or behavioral situations from direct or indirect contacts with
CONCLUSIONS
The current poster reports the results of a preliminary review of the disease-questions in social Q&A. There have been previous studies on analyzing questions about certain diseases or topics in social Q&A, but the current study has focused on developing a coding schema which would be applicable to a wide range of topics in health. As the next step of the study, we will apply the coding schema to manually review a comprehensive set of questions about STDs or cancer and conduct an in-depth review of the questions. Also, the coding schema will be used to develop category resources for an automatic approach, conducting data mining of questions. The wording or linguistic patterns of expressions in questions will be identified through the process of manual review and the results will be applied to data mining on a large scale of health questions in social Q&A.

As for the implications of the study, findings from the current study could be used by health-care professionals or information professionals to help them understand the health information needs that people seek out on the Internet, and to develop health-care or information services in order to provide or to guide them to use appropriate information to solve their problems. The coding schema was used to analyze the questions about STDs and cancer in the current study, but would be applicable to analyzing the nature of questions in other health topics.

REFERENCES
Oh, S., Yi, Y. J., & Worrall, A. (under review). Quality of...