Text Mining as a Method of Analyzing Health Questions in Social Q&A

Sanghee Oh
School of Library & Information Studies
Florida State University
P.O. Box 3062100
Tallahassee FL 32306-2100
shoh@cci.fsu.edu

Min Sook Park
School of Library & Information Studies
Florida State University
P.O. Box 3062100
Tallahassee FL 32306-2100
mp11j@my.fsu.edu

ABSTRACT
This poster introduces a work-in-progress project focusing on text mining of health-related questions that people post in social Q&A which they obtain and share information, advice, and experiences. 69,363 of health questions about Sexually Transmitted Diseases (STDs) posted from 2009 to 2012 were randomly collected from Yahoo! Answers and used for text mining and analysis in this study. A preliminary finding of the data analysis is described in this poster. The use of text mining as a method of analyzing STD questions is discussed, as well.

Keywords
Social Q&A, health Information, text mining, online information seeking.

INTRODUCTION
Thanks to Web 2.0 technology, people can easily reach anonymous others who have varying levels of expertise and experiences through a number of channels of social media, such as blogs, wikis, and social networking services. Health is one of the popular topics people discuss in everyday life; in fact, 41% of patients who participated in the Pew Internet Research survey answered that they have experiences of ratings or reading reviews or comments related to health issues given by others in online news groups, websites, or blogs (Fox, 2009).

In social media, people have produced a significant amount of text-based health information, such as, posting a wall message in Facebook, writing an article in a blog or a wiki site, or leaving comments on someone else’s reviews or articles online. Social Q&A (a social questioning and answering service which allows people to ask and answer questions on any topic in everyday life) is one of the essential venues for an extensive volume of text-based health information. In social Q&A, people can obtain information from others for free in an open and anonymous environment. This encourages people to bring up their intimate issues, to actively seek solutions and suggestions, and to share personal experiences as well as to give and receive social and emotional support. A large number of health-related questions and answers have been exchanged in social Q&A, and all of the contents are available for the public in the website so that those who have similar health concerns can search and browse the questions and answers in which they are interested. For example, in Yahoo! Answers, one of the most popular social Q&A sites, approximately 10 million health-related questions and answers have been posted and stored in the website since its launch in 2005.

A substantial amount of information in the category of health has been produced and shared in social Q&A. Little is known, however, about what, exactly, people have discussed or what kinds of health information people have shared in social Q&A. There were previous studies about information behaviors in social Q&A, but most of their approaches were limited to examining a small set of questions and answers (from hundreds to a couple of thousand) and to using the method of content analysis by reviewing questions and answers manually. Given the volume and the growth rate of unstructured text information in social media settings, efficient methods are required to assimilate and interpret this massive amount of information with efficiency and minimal human intervention.

Therefore, the current project is focused on observing health information behaviors by mining a large and complex collection of health questions mainly using a method of text mining with a research question, “What are the STD-related information (e.g., prevention, risk factors, 

\[1\text{This information is retrieved from the Health category of the Yahoo! Answers website on June 20, 2013.}\]
symptoms, diagnosis, treatments) and daily life issues associated with STDs that people would most likely discuss in health questions?"

The large-scale data collection and analysis may provide a more accurate picture of what people do and how they do it in social contexts than is currently available. From a methodological point of view, the use of text mining in the current project is an approach that is applicable to analyzing the nature of questions and answers in other topic areas as well as in examining the nature of information shared or posted in other types of social media. From a practical point of view, findings will be beneficial for health information professionals in that it will help them better understand the health information needs and behaviors of people in real life.

This poster reports on a preliminary finding after analyzing health questions, especially the questions related to Sexually Transmitted Diseases (STDs) posted from 2009 to 2012. Prior to this project, the authors conducted a study of 200 questions on STDs using content analysis in order to identify the major topics and issues people discuss about STDs in social Q&A and a coding schema for manual review of the questions was developed (Oh, Zhang, & Park, 2012). Findings from the previous study were used to interpret the data, which was automatically generated from the text mining in the current project. In discussion, the use of text mining as a method of analyzing STD questions is described.

LITERATURE REVIEW
As the popularity of social media has grown rapidly, so has the interest of researchers in trying to discover the knowledge and insights people share in their postings. Data mining is one of the methods, which is frequently used to uncover authentic information needs. Hu, Downie, Wext, and Ehmann (2005) conducted a preliminary study with 1,800 user-generated reviews in order to understand how to automatically mine music review information.

Text mining, which is one of data-mining techniques, is often considered as an effective tool for extracting knowledge from unstructured text datasets drawn from social media environments (Mottr, Huan, & Johnson, 2008; Rajman & Besancon, 1998). Kim (2009) mined 485 cancer blog posts, and analyzed descriptive characteristics and relevant subject tags. Postings in the cancer blogs were mined for relevant subject concepts using Text Analysis Portal for Research (TAPoR) software and a frequency list of the individual words in each blog posting was generated. Kim further identified essential subject topics for each posting using SPSS text mining software, Clementine. Some studies utilized the tools associated with text mining that can deal with a massive amount of text data. Coreley, Cook, Mikler, and Singh (2010) explored 158,497,700 web and social media (WSM) items collected during an arbitrary time period of 24 weeks to identify outbreaks and increases of Influenza infection shown in WSM. Complemented by a graph-based data mining technique, they applied the text mining approach to detect mentions of influenza in WSM which was then correlated to real-world the Centers for Disease Control and Prevention (CDC) influenza-like illness (ILI) surveillance data. In a social Q&A setting, Kim, Pinkerton, and Ganesh (2011) investigated 5,500 Influenza A Virus (H1N1)-related questions and answers from Yahoo! Answers. Using SPSS Clementine text mining software, they extracted the major topical categories from the collected dataset such as flu-specific terms, medical and non-medical concerns, and sources of information.

METHOD
For analysis of the STD-related questions in this poster, 69,363 questions posted from 2009 to 2012 were randomly collected from the STD subject category in Yahoo! Answers. In order to conduct a comprehensive but thorough review of the questions, special data-mining software was used, namely, IBM SPSS Modeler Premium (SPSS Modeler), since it provides an automatic way to analyze data, using the predictive models contained in the software. In SPSS Modeler, Text Analytics is designed to analyze unstructured data in particular (e.g., web documents, blog posts, customer feedback, emails) by extracting words and concepts from texts and identifying the relationships between them using predictive models in data mining. One of the advantages of using SPSS Modeler is that a variety of control vocabularies are available, by default, to analyze the text data. MeSH (Medical Subject Headings) was mainly used for this project for analyzing the health-related terms. Additionally, the idioms or expressions people use in describing STDs in everyday life were manually added to the vocabulary list and included in the analysis.

In the current study, SPSS Modeler Text Analytics extracted major concepts from the texts of STD questions, counted the frequency of the concepts shown in one question, and listed them by order of frequency in order to identify the popular concepts discussed in the data set. Once all of the major concepts were extracted, they were manually reviewed in order to correct misspelled words, to group similar concepts and to clean noise. Additionally, the software generated concept maps so that we could identify the relationships among the concepts.

RESULTS
From 69,363 questions, a total of 4,072 concepts (terms) were extracted and used for the analysis. A preliminary finding of the concepts and the sample concept maps are reported in this poster.

Table 1 shows the top 20 most frequently mentioned concepts related to STDs and the unique number of questions that included the concepts. The extracted concepts can be grouped into several categories according to their characteristics. For example, “herpes” seemed to be the greatest concern among STD-related diseases that people discussed in their questions, followed by “HIV”, [Table 1 content here]
“AIDS,” “chlamydia,” and “HPV.” People described symptoms and asked if they had STDs (e.g., “symptoms,” “bumps,” “itching,” “sores,” “burn”) and the specific body part on which they had symptoms, as found in the questions (e.g., “vagina,” “penis,” “lips”). They also asked about how to prevent STDs by discussing the concepts, such as “condoms”, “safe sex,” and “unsafe sex” and how to consult with “doctors” or to take “tests.” There were also questions about their partners (e.g., “boyfriend,” “girlfriend,” “girl,” “guy”).

A number of concept maps can be generated to explain the relationships among the concepts. Figure 1 shows a concept map related to “herpes.”

<table>
<thead>
<tr>
<th>Rank</th>
<th>Major Concepts</th>
<th>No. of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sexually Transmitted Disease</td>
<td>15,187</td>
</tr>
<tr>
<td>2</td>
<td>Herpes</td>
<td>13,935</td>
</tr>
<tr>
<td>3</td>
<td>Sex</td>
<td>11,693</td>
</tr>
<tr>
<td>4</td>
<td>Help</td>
<td>10,903</td>
</tr>
<tr>
<td>5</td>
<td>HIV</td>
<td>10,506</td>
</tr>
<tr>
<td>6</td>
<td>Vagina</td>
<td>5,929</td>
</tr>
<tr>
<td>7</td>
<td>Doctor</td>
<td>5,784</td>
</tr>
<tr>
<td>8</td>
<td>Test</td>
<td>5,138</td>
</tr>
<tr>
<td>9</td>
<td>Condom</td>
<td>5,102</td>
</tr>
<tr>
<td>10</td>
<td>Acquired Immunodeficiency Syndrome (AIDS)</td>
<td>5,101</td>
</tr>
<tr>
<td>11</td>
<td>Boyfriend</td>
<td>4,946</td>
</tr>
<tr>
<td>12</td>
<td>Symptoms</td>
<td>4,586</td>
</tr>
<tr>
<td>13</td>
<td>Question</td>
<td>4,447</td>
</tr>
<tr>
<td>14</td>
<td>Guy</td>
<td>4,136</td>
</tr>
<tr>
<td>15</td>
<td>Bumps</td>
<td>4,077</td>
</tr>
<tr>
<td>16</td>
<td>Chlamydia</td>
<td>3,773</td>
</tr>
<tr>
<td>17</td>
<td>HPV</td>
<td>3,692</td>
</tr>
<tr>
<td>18</td>
<td>Girl</td>
<td>3,642</td>
</tr>
<tr>
<td>19</td>
<td>Penis</td>
<td>3,542</td>
</tr>
<tr>
<td>20</td>
<td>Need</td>
<td>3,421</td>
</tr>
</tbody>
</table>

Table 1. The top 20 most popular concepts in STD questions.

The concept map in Figure 1 shows a list of the concepts highly related to “herpes.” Figure 1 shows the 13 concepts that are most related to herpes. The thickness of the lines connecting the terms indicates the degree of similarity. The thicker the line, the closer the concepts are related. The concept map shows that people discuss herpes more when it exists in the area of the mouth than when it is in the vagina or genital areas. People also mentioned other body parts, such as “hair” and “head.” It seems that the concepts, “soda”, “drinking,” and ‘beverage’ were shown because they are considered as indirect channels of herpes infection.

Figure 2 shows a sample of the concept map about “HIV.” Eighteen concepts related to HIV were generated in the concept map. “Tests” and “examinations” were the most closely related concepts to HIV. It seems that people discussed those tests using “blood.” People were interested not only in HIV “symptoms” but also in treatments because they ask about “medicine”, “drugs,” and “antibiotics.” Additionally, they discussed HIV using “virus” and “infection.”

**DISCUSSION**

This project proposes that text mining is an effective method of analyzing massive, unstructured data; it can be used for identifying health-related issues and problems that
people share in social Q&A. It was found that text mining is useful for identifying the key terms or concepts that arise frequently in STD-related questions in the current data set. Also, the concept maps generated meaningful data with which to identify the strong or weak relationships among the concepts.

There are, however, limitations when using text mining as the sole method for understanding what people discuss in social Q&A. Interpretation of the results from text mining is mostly based on terms without considering the contexts in which the questions are shared. For example, the case by-case situations in which people find themselves when asking for someone else’s opinions on STDs in the anonymous environment should be observed in order to better interpret the results from text mining. As noted earlier, the authors of this poster manually reviewed about 200 STD questions and conducted content analysis of them in order to understand the contexts of the questions. An additional review of the content analysis with about 1,000 STD questions had been conducted prior to this study (Findings will be reported in a forthcoming journal article) and it helped with the analysis and interpretation of the data obtained from text mining in this current project.

Another important thing to consider in text mining is that it is necessary to conduct iterative processes of manual review of the data prior to the data analysis. SPSS Modeler is a useful tool to extract major terms using a set of control vocabularies automatically, but it was found that there were terms and phrases that have been either missed or added incorrectly during the process of concept extraction. A series of data extraction and manual reviews of the data had to be done in order to clean the data prior to the data analysis.

CONCLUSION
Based on the preliminary findings from this poster, an in-depth analysis of the topics and issues people discuss when posting health questions about STDs in social Q&A using text mining will be continued. Additionally, both manual and automatic approaches with which to analyze and interpret data will be further investigated by comparing findings from the current study using text mining and findings from the previous study that used content analysis. Although the current study focuses on STD questions only, the scope of the topics will be expanded to other important health topics. Not only health questions but also health answers will be thoroughly analyzed, as well.

REFERENCES


