HIV/AIDS Question Analysis with Text Mining: Using Concept Maps for Data Analysis and Interpretation

Sanghee Oh, Florida State University
Min Sook Park, Florida State University

Abstract
This poster reports preliminary findings of a work-in progress project focusing on examining questions regarding HIV/AIDS that people generate in social knowledge spaces, social Q&A. A total of 15,574 HIV/AIDS questions out of 74,665 STD questions posted in Yahoo! Answers were randomly selected and analyzed using text mining. Category maps and concept maps that have been used for interpreting the data from text mining are introduced in this poster.
Keywords: Health information behavior; Social Q&A; Text mining; Content analysis
Citation: Oh, S., Park, M.S. (2015). HIV/AIDS Question Analysis with Text Mining: Using Concept Maps for Data Analysis and Interpretation. In *iConference 2015 Proceedings*.

Contact: shoh@cci.fsu.edu, mp11j@my.fsu.edu

1 Introduction

Thanks to the rapid advances and explosive popularity of web 2.0 technologies, the web has quickly grown into a convenient platform for lay people to communicate and share information (Andersen & Söderqvist, 2012). Health is one of the most popular topics in variety of social media platforms. In fact, 72% of Internet users say they looked online for health information (Fox, & Duggan, 2014). Lay people often ask health questions using social media when they think they might have a sensitive health problem, and are embarrassed to speak openly.

Social Q&A is an online service that allows people to ask and answer questions about many topics in everyday life, such as health issues. Social Q&A enables lay people to interact with those who have similar concerns or issues on the internet and to obtain knowledge, information, personal experiences, advice, suggestions, or social and emotional support from them. This service aims to benefit everyone through the collective wisdom of many, called the Wisdom of Crowds (Surowiecki, 2004). In this sense, social Q&A can be an essential venue for observing the natural behaviors of information seeking with an extensive collection of questions and answers which represent information needs and behaviors in real life.

A significant quantity of text-based health information has been produced in social Q&A. Little is known, however, about what people have tended to discuss 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 those approaches were limited to examining a small set of questions and answers (from hundreds to a couple of thousand) using the content analysis method by reviewing content manually. Instead, the current project is focused on observing health information behaviors from a large and complex collection of health questions, mainly using a method of text mining.

Among diverse health issues discussed in social Q&A, this poster focuses on human immunodeficiency virus (HIV) and acquired immune deficiency syndrome (AIDS) since it is one of the most serious public health challenges. The estimated incidences of HIV/AIDS in the United States are over 1.1 million among people aged 13 years or older. 1 in 6 people with HIV may not be aware of their infection (CDC, 2013).To observe people's information needs regarding HIV/AIDS, this study mined 15,574 health questions with a research question: "What are the HIV/AIDS related information and daily issues associated with HIV/AIDS that people would most likely discuss in health questions?" This project is on-going, and preliminary findings of the text analysis are reported in this poster.

2 Methods

Yahoo! Answers, the test-bed of this study, is one of the most popular social Q&A services with approximately 6.3 million monthly visitors in the United States (quantcast, 2014). For the current study, 15,574 questions which contain keywords, either "HIV" or "AIDS", are randomly selected from a set of 74,665 questions posted between 2009-2014 in the sexually transmitted diseases (STDs) category in Yahoo! Answers. IBM SPSS Modeler Premium (SPSS Modeler) was used to analyze the collected questions. SPSS Modeler extracts concepts (or terms) from the text data and identifies the relationships

between the concepts, using predictive models in data mining. For this study, MeSH was used to extract concepts that people discuss about HIV/AIDS and to detect the relationships between relevant concepts. Prior to running text mining, variations of natural language (e.g., slangs, idioms, typos) were normalized through preprocessing. Terms with different wording or misspelling were cleaned to reduce the noise in the original collection of questions.

Prior to this study, authors performed content analysis of 188 questions on HIV/AIDS to capture contextual meanings of terms. Findings from the content analysis were used to interpret the data generated from text mining. Once the major concepts regarding HIV/AIDS were extracted, the major concepts were further analyzed by a set of health information categories developed from content analysis, such as 1) Body Part/System, 2) Daily Lives, 3) Disease, 4) Prevention, 5) Relationships, 6) Symptoms, 7) Tests, 8) Treatments, and 9) Emotion. The categories were used for grouping the associated concepts and to generate concept maps to identify the relationships among the concepts.

3 Results & Discussion

3.1 Results

A total of 3,027 terms were extracted from 15,574 questions. Among them, 237 terms that are occurred in 50 or more unique number of questions were manually reviewed and identified as concepts related to HIV/AIDS. These concepts were classified into one of the nine categories in order to analyze the relationships between categories and terms. This poster mainly presents the findings of two types of concept maps generated by SPSS Modeler during the process of text mining -- 1) concept maps of categories (named category maps), and 2) concept maps of terms.

Category maps present the distribution of the extracted concepts assigned to the nine categories and their relationships (See Figures 1 and 2). The size of a node (a blue circle next to the names of categories) indicates the number of concepts classified to that category. The lines that connect the nodes present the number of co-occurring questions in the connected nodes. Thicker lines indicate that more questions are related to the connected categories.

The category maps of HIV and AIDS from Figures 1 and 2 show a similar pattern to one another in that the greatest number of HIV/AIDS related concepts are about Disease and this was followed by Prevention, Relationship, Body Part/Body System, and Emotion. The lines that start from Disease are thicker than others and are connected to Prevention, Relationship, and Body Part/Body System. People ask questions about HIV/AIDS related diseases such as chlamydia in questions, in relation to how to prevent the diseases or what to do with symptoms of a disease that people observe from their body parts (The node, Body Part/Body System, is connected to Symptoms with a thick line). People are also concerned about family or social relationships with others, asking how they can discuss their diseases with their family members or close friends.

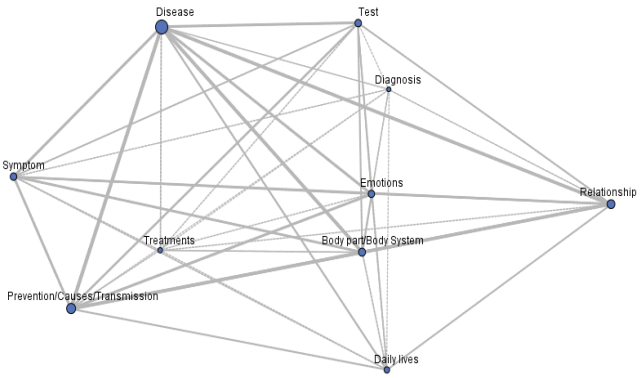

Figure 1. Category Map of HIV

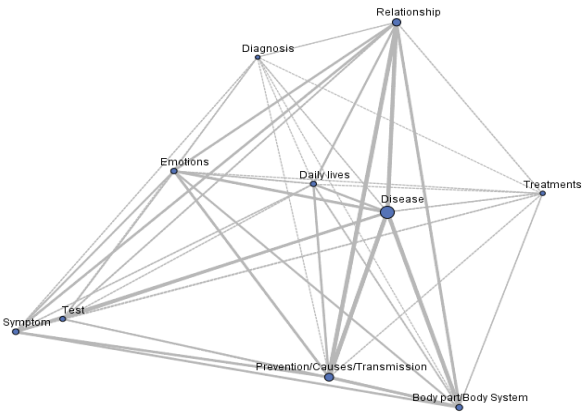

Figure 2. Category Map of AIDS

Compared to category maps, concept maps show the relationship between HIV/AIDS and other extracted concepts assigned to each category (See Figures 3 and 4) The color of each box with the concepts indicates one of the nine categories; 1) purple is for Body System, 2) brown is for Daily Lives, 3) grey is for Disease, 4) green is for Prevention, 5) dark blue is for Relationship, 6) pink is for Symptom, 7) light green is for Test, 8) olive green is for Treatment, and 9) blue is for Emotion. The lines indicate the

number of questions that contains the connected concepts. Thus, thicker lines indicate that the more the concepts are related.

Figures 3 and 4 show the top 30 concepts that are highly related to HIV and AIDS, respectively. People often discuss “blood” and “unsafe sex” in relation to HIV, as a channel of infection more seriously than others such as “fingering”, “homosexuality”, and “food”. People may believe that they have symptoms of HIV, such as fever, fatigue, flu, and soreness in many places of their body parts (See the terms in purple box). They also express negative emotions such as “worried” and “scared” regarding HIV. The concepts “blood” and “condom” were the most highly discussed in relation to AIDS, considering “blood” as a channel of infection whereas “condom” as a preventive method. The concepts regarding relationship such as “man,” “guy,” and “girl” appear more often than those in the map of HIV. Social relationships or interactions with others are also frequent concerns when people discuss AIDS.

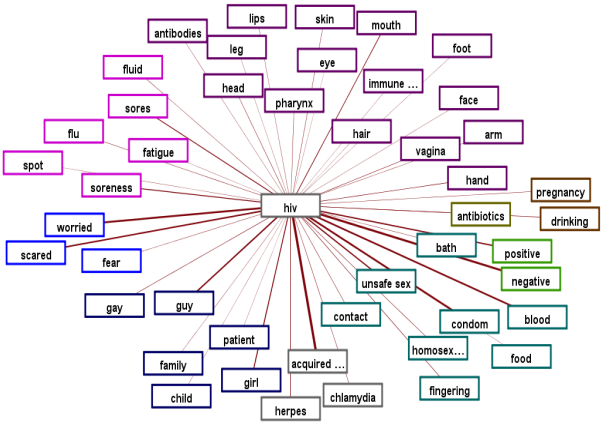

Figure 3. A Concept Map of HIV

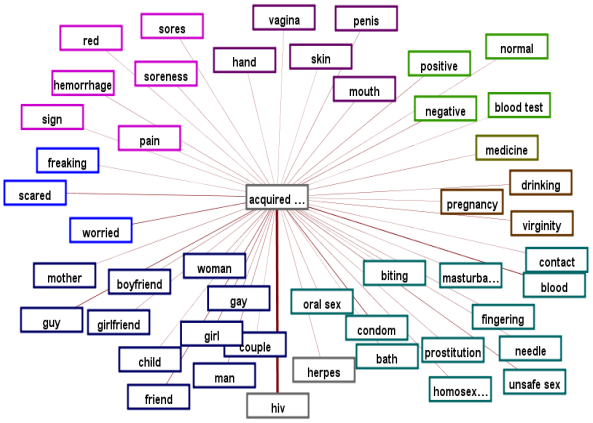

Figure 4. A Concept Map of AIDS

3.2 Discussion & Conclusion

Text mining with a large set of HIV/AIDS questions can be performed in several different ways to interpret data. Category maps and concept maps are useful tools to identify hidden relationships between categories and concepts. The presence and frequency of terms shown in each category are transformed into circles or colored boxes in concept maps. This facilitates reviewing the nature and trends of concepts at large. The lines connecting terms (nodes) show the strength of the relationships between concepts (either strongly related or weakly related). This aids the interpretation of data by examining how the concepts and categories are related to one another. A handful category and concept maps are presented in this poster with a preliminary review of the data, but an in-depth analysis of the major concepts pertaining to HIV/AIDS will be further analyzed.

Another approach that we plan to carry out is analyzing the semantic relations between concepts. Basically, semantic relations are defined as meaningful associations between two or more concepts or entities expressed in text and then identifying the meaning connecting those concepts (Khoo & Na, 2006; Slaughter, 2002). Once the semantic meaning between the concepts is identified, the relationships between concepts could be interpreted more thoroughly. HIV/AIDS questions that we obtained from Yahoo! Answers are a collection of representative expressions generated from a number of people who have common concerns on HIV/AIDS. The semantics that we identify from the collection of questions could represent the semantic structure of expressions that people have shared in the HIV/AIDS community.

A large scale data collection and analysis could provide a comprehensive picture of the nature of health information needs within health topics. Findings would be beneficial for health information professionals to help them better understand the needs for health information, as well as understand what types of information and concepts lay people conceptually associate with their health issues in real life.

References

Abbas, J. (2010). *Structures for Organizing Knowledge: Exploring Taxonomies, Ontologies, and Other Schemas*. New York, NY: Neal-Schuman Publishers, Inc.

Andersen, N., & Söderqvist, T. (2012). *Social media and public health research* (Technical report).Centers for Disease Control and Prevention (CDC) (2013). *HIV/ADIS* Retrieved from http://www.cdc.gov/hiv/

Khoo, C., & Na, J.-C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–228.

Peters, I. (2009). *Folksonomies: Indexing and Retrieval in Web 2.0*. Berlin, German: Deutsche Nationalbiblothek.

Fox, S., & Duggan, M.(2014) *Health fact sheet*. Retrieved from http://www.pewinternet.org/fact-sheets/health-fact-sheet/

Slaughter, L. (2002). *Semantic relationships in health consumer questions and physicians' answers: A basis for representing medical knowledge and for concept exploration interfaces*. University of Maryland, College Park.

Surowiecki, James. (2004). *The Wisdom of Crowds*. New York: Doubleday.Quantcast, (2014). *Yahoo! Answers.* Retrieved from https://www.quantcast.com/answers.yahoo.com#!traffic

Quantcast (2014). Answers.yahoo.com Retrieved from https://www.quantcast.com/answers.yahoo.com