Understanding Health Information Behaviors in Social Q&A: Text Mining of Health Questions and Answers in Yahoo! Answers

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

Final Report
2013 OCLC/ALISE Library and Information Science Research Grant Project

22 February 2015

Suggested citation:
Oh, Sanghee. 2015. "Understanding Health Information Behaviors in Social Q&A: Text Mining of Health Questions and Answers in Yahoo! Answers." 2013 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at: http://www.oclc.org/research/grants/reports/2013/oh2013.pdf

1. Introduction

The purpose of the current project is to investigate the health information needs people most likely seek and share in social media, by analyzing the content of health questions and answers 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. Social Q&A throws questions to the public, allowing everyone to benefit from the collective wisdom of many, called the Wisdom of Crowds (Surowiecki, 2004, p. 11). People may ask health questions in social Q&A because they think it is a minor problem or feel embarrassed to ask their doctors, may have not been satisfied with answers from doctors, or would like to learn from others' experiences (Oh & Zhang, Underdevelopment).

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. In social Q&A, people can elaborate their information needs in questions or describe sources of information in answers with their own words, explaining their diseases, medical histories, conditions, or resources with as many (or as few) details as they wish. All of the questions and answers in social Q&A are self-reported with real life experiences and the number of questions and answers collected in social Q&A services is substantial. For example, approximately 10 million health questions and associated answers are currently available for use in Yahoo! Answers[1], the most popular social Q&A service. In this report, therefore, the health questions from Yahoo! Answers were collected and analyzed using a method of a large-scale text mining, by extracting key terms and concepts from questions and answers and identifying relationships among the terms and concepts.

[1] This information is retrieved from the Health category of the Yahoo! Answers website on August 19, 2012.

2. Problem Statement

A significant amount of information in health has been produced and shared in social Q&A. Little is known, however, about what 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 by examining a small set of questions and answers (from hundreds to a couple of thousand) using the method of content analysis by reviewing them manually. Instead, the current project will be focused on observing health information behaviors from a large and complex collection of health questions mainly using a method of text mining with the following research questions:

RQ 1. What is the disease specific information (e.g., prevention, risk factors, symptoms, diagnosis, treatments) people would most likely discuss in health questions?

RQ 2. What are the personal experiences, expertise, and resources people share in health questions?

RQ 3. What are the social and emotional supports people would like to receive or share in health questions?

RQ 4. What are the relationships among the findings from the research questions above?

RQ 5. How have the findings from the research questions above evolved over time, from 2009 to 2012?

RQ 1 basically examines the major terms and concepts associated with certain diseases or health conditions. RQ2 investigates information people share in questions. People explain their medical or non-medical conditions in questions in order to obtain customized answers to their personal situations. RQ 3 indicates that people would like to seek and share social and emotional supports through exchanging questions and answers. RQs 1, 2, and 3 can be observed from the descriptive analysis of the major terms and concepts related to each. RQ 4 takes a further step from the descriptive analysis to investigating the relationships among the major terms and concepts indicating how they are closely related to one another to better interpret the complicated nature of health information behaviors presented in questions and answers.

Additionally, a longitudinal analysis was conducted in RQ 5 and provided an insight into changes in topics or trends of seeking and sharing health information in social Q&A over the past several years.

3. Significance of the Research

The current project could be a significant endeavor in promoting research into online health information behaviors in social contexts. The large-scale data collection and analysis could provide a more accurate picture of what and how people act in social contexts. From a methodological point of view, the proposed design of text mining in the current project could be applicable to analyzing the nature of questions and answers in other topic areas or the nature of the information shared or posted in other types of social media (e.g., wall messages in social networking sites, tweets, blogs, wikis). From a practical point of view, findings would be beneficial for health information professionals (e.g., doctors, nurses, health librarians, health researchers) to help them better understand the health information needs and behaviors of their patients or customers in real life. They could use the findings to design, evaluate, or improve their services or systems and guide their patients or customers to make health care decisions properly.

4. Background of Research

As the popularity of social Q&A has grown in recent years, so has the interest of researchers in trying to understand the knowledge and insights people share within questions and answers. Ignatova, Toprak, Bernhard, and Gurevych (2009) invited people to review about 800 questions across topics in Yahoo! Answers, using a taxonomy proposed by Graesser, McMahen, and Johnson (1994), which classified questions into concept completion, definition, procedural, comparison, causal, disjunctive, verification, quantification, and general information needs. In a similar approach, Harper, Weinberg, Logie, and Konstan (2010) applied a rhetorical framework for classifying questions into (1) deliberative, (2) epideictic,

and (3) forensic, and tested the framework with questions from Yahoo! Answers, AnswerBag, and Ask Metafilter (100 questions from each).

Most importantly, questions are the representations of information needs in real life. The methods of manual review/content analysis were often used to identify the meaningful features and trends in information needs. Lee (2010) analyzed about 2,000 music-related questions posted on Google Answers[2] using an iterative process of reviewing the questions and developed a taxonomy for music information needs and features. Yoon and Chung (2011) manually reviewed about 500 image-related questions posted on Yahoo! Answers and coded them into three categories: image needs, image attributes, and associated information. In health, Zhang (2010) randomly selected 276 questions and developed customized categories in health: (1) information about a disease, (2) information about drugs or supplements, (3) information about lifestyle, including diet and exercise, (4) information about people with similar conditions, and further analyzed contextual factors, such as user goals, motivations, emotions, and time. The PI of the current proposal also has investigated and reviewed the questions in health with a method of content analysis in order to identify the major topics and issues people discuss about cancer (Oh & Zhang, Underdevelopment) and sexually transmitted diseases (STDs) (Oh & Park, Underdevelopment). A thorough review of the questions in cancer and STDs led to develop a framework, composed of seven types of information described in health questions: (1) demographic information of patients or care givers, (2) disease-specific information provided (3) disease-specific information asked (4) socio-emotional supports asked, (5) daily-life information asked, (6) risk-factors related to certain type of diseases (Oh, Zhang, & Park, 2012). This framework was used for the basic guideline of conducting text mining in the current proposal.

Answers are the representations of information sources people share in everyday life. Information described in answers is comprehensive and resourceful. People provide information in answers based on

[2] It is a fee-based question asking and answering service. Google Answers were discontinued in 2009, but the questions and answers shared when the service was alive are available from http://answers.google.com/answers/.

their own knowledge and experiences. They indicate references of information (e.g., book titles, newspapers, URLs) to help others to locate the original sources (Oh, Oh, & Shah, 2008). They also express their emotional and social supports in answers (Kim & Oh, 2009). In spite of the variety and nature of the answers, there are few studies about the contents of answers. One of the major interests in research pertaining to answers in social Q&A was the answer evaluation in quality (Harper, Raban, Rafaeli, & Konstan, 2008; Oh, Yi, & Worrall, 2012), relevance (Kim & Oh, 2009), or credibility (Kim, 2010; Savolainen, 2011). Most of these previous studies also used the method of manual review/content analysis in evaluating answers.

There was a previous study which analyzed both questions and answers in social Q&A using a method of text mining. Kim, Pinkerton, and Ganesh (2011) examined about 5,500 Influenza A Virus (H1N1)-related questions and answers from Yahoo! Answers and identified flu-specific terms, medical and non-medical concerns, and sources of information presented in questions and answers.

5. Method

Test-bed: Yahoo! Answers

Yahoo! Answers is a representative, top ranking social Q&A service which has 17 million visitors per month in the U.S. alone [3]. People with a wide range of experiences and expertise, from lay people to health experts, ask and answer questions in Yahoo! Answers (Oh, 2012). The health-related topic coverage in Yahoo! Answers is comprehensive. Basically, people can ask and answer questions related to 18 health topics – Alternative Medicine, Diet & Fitness, Men's Health, Women's Health, Mental Health, Dental, Optical, Allergies, Cancer, Diabetes, Heart Disease, Infectious Diseases, Respiratory Diseases, Skin Conditions, Sexually Transmitted Diseases (STDs), First Aid, Injuries, and Pain & Pain Management. The topics excluded from the categories above are covered in two others categories - Other Diseases, and Other General Health Care. Users' information needs and experiences can vary depending on the topic in health.

[3] A 2012 report from Quantcast.com: http://www.quantcast.com/answers.yahoo.com

In the current report, STD was selected as a topic to analyze due to its sensitive nature; people may ask STD questions in social Q&A due to their own concerns about prevention and transmission of the diseases.

Data Collection: Health Questions in Yahoo! Answers

A web crawler, using Yahoo! Answers Application Programming Interface (API), was designed to collect health questions and associated answers posted in Yahoo! Answers. The API is freely available from the Yahoo! Developers website and it is allowed to collect about 5,000 sets of data per day. The API collects not only the contents of questions and answers but also data associated with them such as topical categories, an additional description of questions or answers, dates when questions or answers are posted, and star ratings given by Yahoo! Answer users. The PI has collected 69,363 STD questions and associated data using the API; the questions were posted from 2009 to 2012 and are stored in a SQL database.

Text Mining

IBM® SPSS® Modeler Premium (SPSS Modeler) is text mining software designed to analyze unstructured data, extracting words and concepts from texts and identifying the relationships between them using predictive models in data mining. In the current study, SPSS Modeler extracted terms from the texts of health questions, counted the unique number of frequency of the terms, and listed them by order of frequency. This first revealed the key terms that people used when discussing STD in health questions. And then, the key terms were classified by the types of information, proposed in Table 1, and further analyzed.

Table 1: Information Framework for Health Question Analysis in Social Q&A

Types of Information	Definitions & Descriptions	Sub-Types
Demographic Information	• Demographic information of patients and care givers	Sex, age, relationship between questions and patients or potential patients in questions
Diseases-specific Information	• Information asked or shared for by patients or potential patients about a disease that they are suspicious of having or for which they are already diagnosed	Prevention, symptoms, diagnoses, tests, treatment, prognoses

Socio-emotional Information	• Information asked for by patients or potential patients about the ways of handling their situations emotionally	Isolation, acceptance, social supports, coping
Daily Life Information	• Information asked for by patients or potential patients for maintaining healthy lives	Alcohol, smoking, environmental factors, medical records, family medical history, food/exercise.

The information framework in Table 1 has been developed from the PI's previous studies about the content analysis of health questions in social Q&A (Oh, Zhang, et al., 2012). Findings from the review were considered to revise the framework. With the framework, the relationships among key terms were defined and compared. SPSS Modeler was mainly used for the data analysis. In order to extract meaningful terms and concepts from text, MeSH (Medical Subject Headings) were employed as a main dictionary. In addition, a customized dictionary was developed to extract concepts from daily expressions in Yahoo Answers! users' discussions. The frequencies of the extracted words and concepts were counted, and were grouped into the categories of the information framework shown in Table 1.

6. Findings

From 69,363 STD questions, a total of 5,000 concepts (terms) were extracted. This report includes identification of the concepts and the sample concept maps. Table 2 shows the top 20 most popular concepts that are most frequently mentioned in STDs and the unique number of questions that are included in each concept. These concepts represent what people discuss regarding STDs. For example, people discuss disease names they suspect they may have, such as herpes, HIV, and AIDS. In the most common cases, people seem not to be able to specify what types of STDs they suspect, mentioning STDs in general. People described their symptoms to get others' opinions and advice on whether they may have STDs (e.g., "symptoms", "bumps", "itching", "sores", "burn"); they reveal the specific body part on which they had symptoms, as found in the questions (e.g., "vagina", "penis", "lips"). They also asked about suspected infection channels by discussing their relationships and sexual behaviors (e.g. "boyfriend," "girlfriend,"

"girl," "sex"). The prevention methods were also discussed by mentioning the concepts such as "condoms", "safe sex," and "unsafe sex" and how to consult with "doctors" or to take "tests."

Table 2. Top 20 most popular concepts

Rank	Major Concepts	No. of Questions
1	STDs	18,229
2	Sex	17,945
3	Herpes	15,432
4	help	15,029
5	HIV	11,739
6	Doctor	10,168
7	Vagina	7,866
8	Boyfriend	7,800
9	Condom	7,737
10	Test	7,543
11	Symptoms	7,259
12	Guy	7,052
13	Bumps	6,361
14	Question	6,211
15	Girl	6,014
16	AIDS	5,669
17	Feel	5,639
18	Need	5,435
19	Day	5,428
20	Penis	5,382

Disease-specific information

The extracted concepts were grouped into several categories according to their characteristics as shown in Figure 1. STD diseases appear to be the most popular topic people discuss in their questions, followed by risk factors, symptom description, relationships, and body parts. To illustrate, people appear to have discussed types of STDs they suspect they may have (e.g. HIV, Herpes, Gonorrhea, HPV). Risk factors, such as sexual relationships, sexual behaviors, and sharing foods with potential carriers, were discussed, regarding STDs. When seeking other's diagnostic opinions or advice, people described what symptoms they have (e.g. heavy discharges, changes in their skin colors, pains). The body parts on which those symptoms appear are also described. As STDs are transmitted thorough sexual relations, people's current, past, or casual relationships were also discussed in their questions.

Figure 2 illustrates types of STDs people are most commonly concerned about among STD-related diseases. STDs in general (18,229 questions) seem to be the most popular terms people mentioned when they discuss about STDs. When people specify types of STD-related disease in detail, "herpes" (15,432 questions) appears to be the greatest concern among other like diseases, followed by HIV(11,739 questions), AIDS (5,669 questions), HPV(4,173 questions), and chlamydia (4,548 questions).

Figure 1. Number of questions in categories

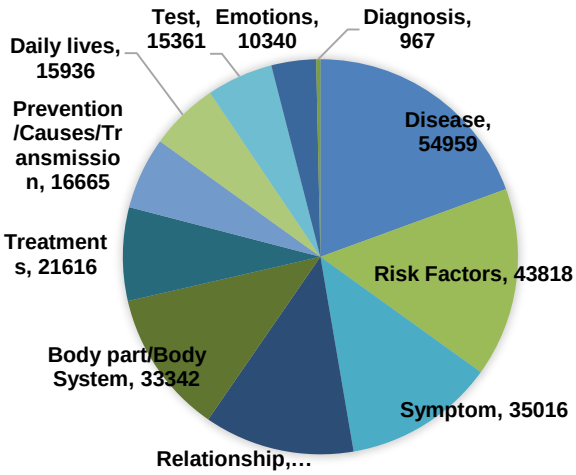
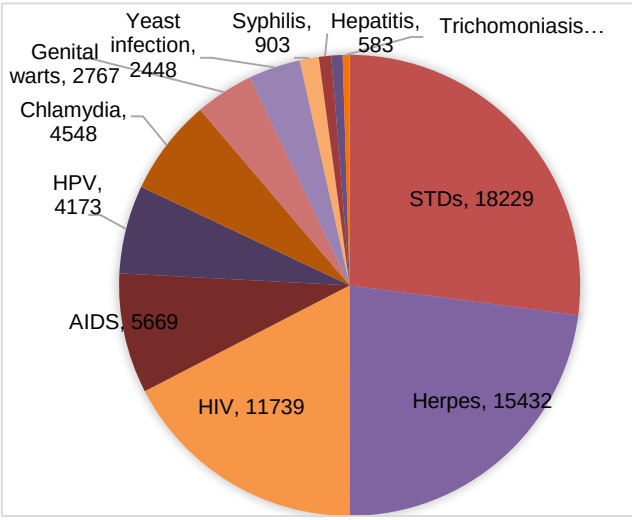


Figure 2. Types of STD



Each of the extracted concepts that people discussed in their questions can be visually displayed as shown in Figure 3 and Figure 4. These concept maps indicate the relationships among the major concepts by measuring the frequencies of the concepts commonly appearing in multiple data. To illustrate, Figure 3 shows a concept map related to "herpes." In the concept map, we can see a list of the 30 concepts highly 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 similarities are calculated based on the number of questions that have same concepts in them.

The concept map shows that people discuss herpes more when it appears 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. When people discuss treatments and tests, they seem to have mentioned "medicine," "research," and "urine [test]." The concept map of HIV in Figure 4 shows a sample map about "HIV." This map has 18 concepts that are related to HIV. "Test" and "Examination" are the most closely related to HIV. It appears that people discuss the test or examination using blood. When discussing methods for treatments or recovery, it seems people mentioned treatments using "medicine", "antibodies", "drugs." The terms such as "Sex, "Sexuality," and "infection" appear to represent people aware that sexual behaviors could be HIV infection channels.

Figure 3 A concept map of herpes

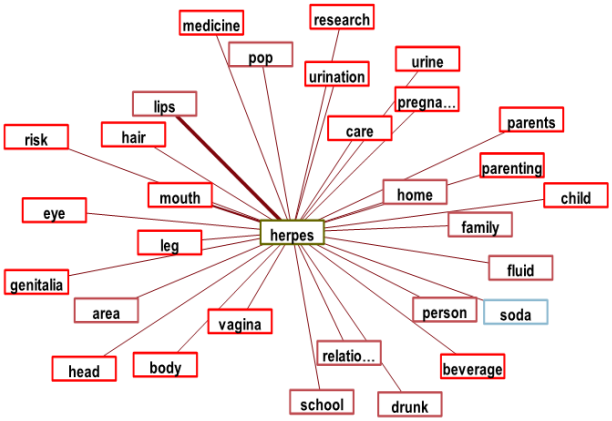
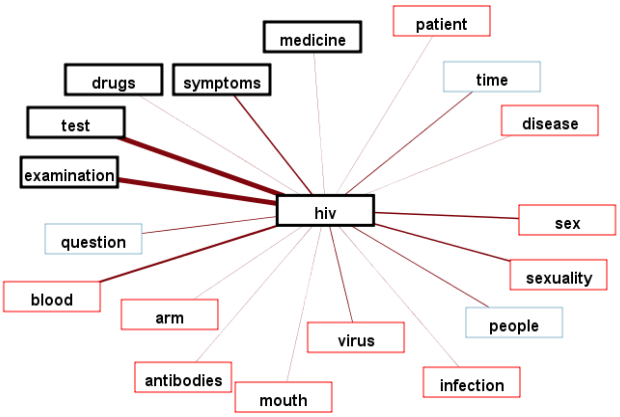


Figure 4. Concept map of HIV



Personal experiences, expertise and resources

Table 3 shows the top 20 most popular life related issues that people discussed regarding their STDs. By discussing with those who have had similar experiences, people seem to look for suggestions and advice on what possible impact the suspect disease may have on their lives. It seems that the concepts, "pregnancy", "baby," "kids," and "infertility' in Table 3 appeared since people are concerned their STDs might be transmitted to their current or unborn baby or children. They also seem to be concerned that having sexually transmitted related diseases may impact their current or future pregnancy, and in extreme cases it may lead to infertility. People's concerns regarding STDs were expressed with concepts such as "cost," "pay," and "health insurance." The possible impact on their current or future relationships also appears to

worry people having STDs. Their worries appeared within the terms such as “marriage,” “future,” and “relationships.”

Table 3. The top 20 most popular life issues

	Concept	No. of Questions
1	virginity	3,875
2	life	3,145
3	pregnancy	2,584
4	baby	976
5	kids	975
6	situation	872
7	health	871
8	birth control	766
9	health insurance	656
10	money	517
11	effect	471
12	planned parenthood	441
13	paperwork	416
14	lead	359
15	pay	321
16	cost	313
17	lie	307
18	future	286
19	infertility	271
20	marriage	220

Table 4. The top 20 most popular emotional issues

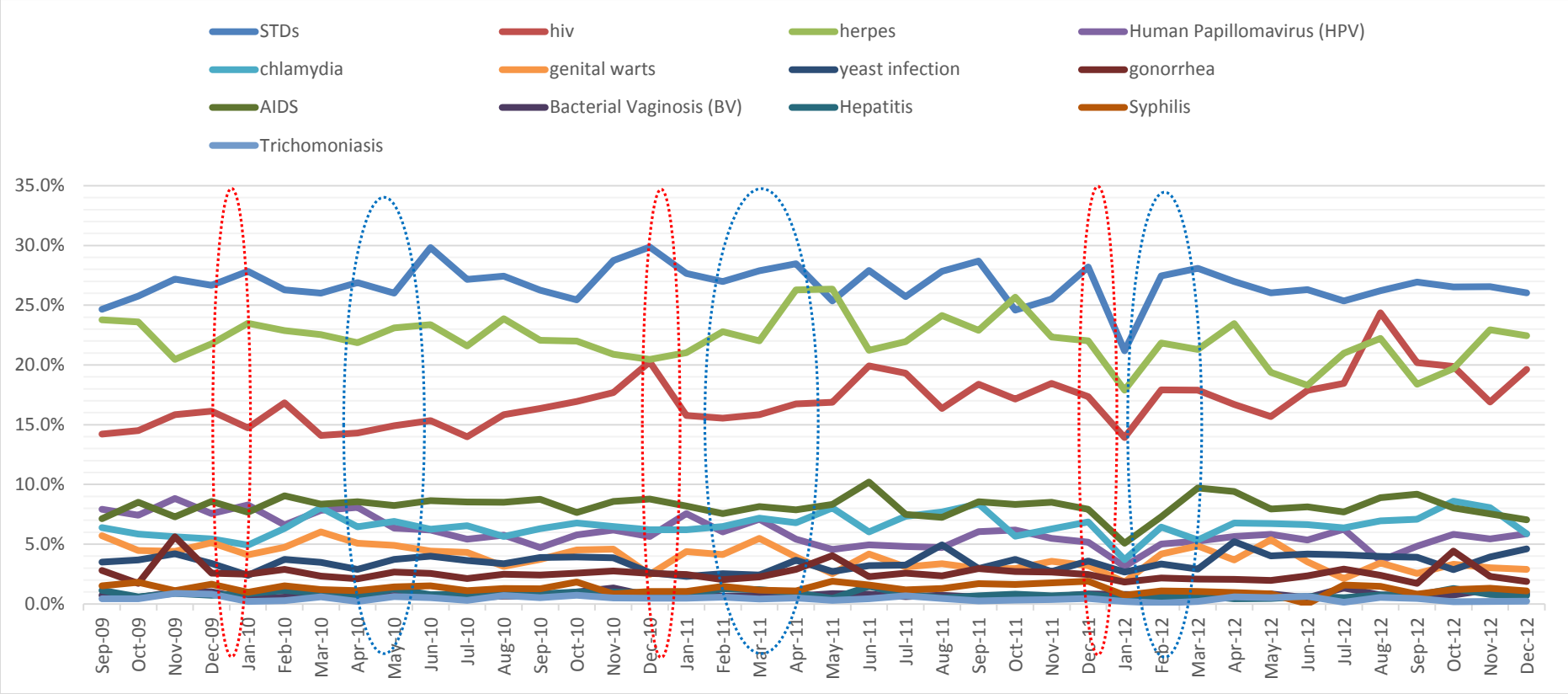
	Concepts	No. of Questions
1	freaking	1,213
2	worried	980
3	I don't know	718
4	love	640
5	fear	334
6	trust	329
7	anxiety	325
8	mistake	290
9	hate	287
10	doubt	278
11	nasty	243
12	concern	235
13	embarrassing	169
14	panic	149
15	ease	144
16	regret	138
17	hypochondria	107
18	fault	90
19	relief	86
20	pleasure	86

Emotional support from others was found as another factor that people want to receive and share in health questions in Social Q&A. People most commonly feel “freaking,” (1,213 questions) and “worried” (980 questions) when they suspect or they realize they have an STD. Some expressed their confusion saying “I don’t know.” They also described their fear and concerns by discussing the concepts, such as “fear”, “anxiety,” and “panic.” Some expressed their emotions about their partners or themselves, saying “hate,” “trust,” “doubt,” “regret,” and “mistake.” These terms appear to have been discussed since people suspect their STDs might have been transmitted from their partners, or regret their past sexual relations with others.

Longitudinal changes in topics

In order to examine if there are changes in the topics people discuss in social Q&A, topics regarding sexually transmitted related disease were analyzed. There seems to be no dramatic change in topics or trends of seeking and sharing health information in Social Q&A, regarding STDs issues, from 2009 to 2012. Slight decreases were identified during the winter seasons as shown in the red lines, and moderate increases were detected during spring seasons (See Figure 5).

Figure 5. The trends of STD topics presented in questions posted between 2009 and 2012



7. Evaluation & Future Research

Findings from the current project have been reviewed by STD healthcare providers in order to apply them to enhance STD clinicians' awareness of the public needs for STD information. The PI is currently working on developing publications reporting the findings and their applications to healthcare, evaluating how successfully the current health information services provide the kinds of information people want and guides them to use information in social media. STD was the topic focused on in the current project but the scope of health topics will be expanded to others, such as cancer. The PI is also collaborating with colleagues in information science to develop a theoretical foundation pertaining to the contexts in which people ask for health information online. In another stream of future research, the research design of the current project in text mining will be used to analyze other types of text messages about health in social media, such as tweets in Twitter and wall postings in Facebook. The nature and use of health information across these different types of social media will be compared and contrasted in future studies.

Acknowledgements

The PI would like to thank OCLC/ALISE for funding this project. Min Sook Park, doctoral student in School of Information at Florida State University and a research assistant of this project, participated in analyzing the data and contributed to developing this report.

References

Graesser, A., McMahen, C., & Johnson, B. (1994). Question asking and answering. *Handbook of psycholinguistics*, 517–538.

Harper, F. M., Raban, D. R., Rafaeli, S., & Konstan, J. K. (2008). *Predictors of Answer Quality in Online Q&A sites.* Paper presented at the CHI, Florence, Italy.

Harper, F. M., Weinberg, J., Logie, J., & Konstan, J. A. (2010). Question types in social Q&A sites. *First Monday, 15*(7).

Ignatova, K., Toprak, C., Bernhard, D., & Gurevych, I. (2009). Annotating question types in social Q&A sites. In W. Hoeppner (Ed.), *GSCL Symposium: Speech Technology and eHumanities* (pp. 44-49). Duisburg: Universität Duisburg-Essen.

Kim, S. (2010). Questioners' credibility judgments of answers in a social question and answer site. *Information Research, 15*(1), 432.

Kim, S., & Oh, S. (2009). Users' relevance criteria for evaluating answers in a social Q&A site. *Journal of the American Society for Information Science and Technology, 60*(4), 716-727.

Kim, S., Pinkerton, T., & Ganesh, N. (2011). Assessment of H1N1 questions and answers posted on the Web. *American Journal of Infection Control, 40*(3), 211-217.

Lee, J. H. (2010). Analysis of user needs and information features in natural language queries seeking music information. *Journal of the American Society for Information Science and Technology, 61*(5), 1025-1045.

Oh, S. (2012). Tha characteristics and motivations of health answers for sharing information, knowledge, and experiences in online environments. *Journal of the American Soceity for Information Science and Technology, 63*(3), 543-557.

Oh, S., Oh, J. S., & Shah, C. (2008). *The Use of Information Source by the Internet Users in Answer Questions.* Paper presented at the the 71st Annual Meeting of the American Society for Information Science and Technology.

Oh, S., & Park, M. (Underdevelopment). Information Needs on Sexually Transmitted Diseases (STDs)

Oh, S., Yi, Y., & Worrall, A. (2012). Quality of Health in Social Q&A *Proceedings of the 74th Annual Conference of the American Society for Information Science & Technology (ASIST' 12).* Silver Spring, MD: ASIST.

Oh, S., & Zhang, Y. (Underdevelopment). Cancer Information Needs in Social Q&A

Oh, S., Zhang, Y., & Park, M. (2012). Health information needs on diseases: A coding schema development for analyzing health questions in social Q&a *Proceedings of the 75th Annual Conference of the American Society for Information Science & Technology (ASIST' 12)*. Silver Spring, MD: ASIST

Savolainen, R. (2011). Judging the quality and credibility of information in Internet discussion forums. *Journal of the American Society for Information Science and Technology, 62*(7), 1243-1256.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Doubleday.

Yoon, J. W., & Chung, E. K. (2011). Understanding image needs in daily life by analyzing questions in a social Q&A site. *Journal of the American Society for Information Science and Technology*.

Zhang, Y. (2010). Contextualizing consumer health information searching: an analysis of questions in a social Q&A community *Proceedings of the ACM International Health Informatics Symposium* (pp. 210-219). New York, NY: ACM.