Research Design: Understanding Semantic Relationships in Health Question-Answering Behavior in Social Context

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

Abstract
This poster introduces a research design focusing on understanding the semantic relationships in socially generated health information in social Q&A. A total of 164,279 questions and 413,900 answers posted during 2013 will be used for text mining and content analysis in this study. This poster explains the process of using the mixed methods for identifying the semantic relationships between major concepts in the questions and the answers.
Keywords: Semantic relationships; health information; social Q&A; text mining
Citation: Park, M.S., Oh, S. (2015). Research Design: Understanding Semantic Relationships in Health Question-Answering Behavior in Social Context. In *iConference 2015 Proceedings*.

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

1 Introduction

Thanks to the explosive popularity of web 2.0 technologies, looking for health information became one of the most popular online activities for lay people. In fact, 78% of the US adult population says that they looked online for health information (HINTS, 2013). Lay people are, however, not necessarily successful in finding the health information they want online, mainly due to their lack of understanding of medical concepts and their unfamiliarity with effective search strategies. Furthermore, the massive amount of an unstructured data on the Web 2.0 has resulted in difficulties in finding relevant information and has caused users to experience information overload.

In this respect, understanding the semantic relationships of unstructured data distributed through social media could be helpful for users to identify and evaluate health information online. Semantic relationships are meaningful associations between two or more concepts expressed in text and conveys the meaning of connecting those concepts (Khoo & Na, 2006; Slaughter, 2002). Recent studies, including work by Ding, Jacob, Caverlee, Fried, & Zhixiong (2009) and Mika (2007) suggest that there is a potential for the profitable convergence of formal semantic structures with the large volume of social Web users' intelligence. Mining the agreed concepts from a large amount of socially generated knowledge will enable the extraction of semantic structures of expressions that the social community uses in common (Gruber, 2007; Mika, 2007; Mikroyannidis, 2007). To illustrate, one can detect hidden, implicit semantic relations from the aggregation of user-driven data and the associated entities (e.g., resources, users, time, rating etc.) and then can make the relations explicit.

However, there has been little efforts to understand the semantic relationships of concepts in questions and answers generated in social contexts. This proposal, therefore, concentrates on the methods with which to investigate and discover semantic relationships observed in health information in the social Q&A (social questioning and answering) settings. In social Q&A, people seek and share an extensive volume of text-based health information in the form of question and answers for their real life health issues (Oh, Zhang, & Park, 2012). Social Q&A is one of the essential venues where lay people share and discuss their health issues with others who may have similar concerns (Oh & Park, 2013). In questions, people express their information needs (Oh et al., 2012). Taking advantage of sharing by many of those who are willing to share their knowledge, experience, and support, people obtain information and social support from others in the form of answers.

To examine semantic relationships that occur in the social context, this study examines the types and the frequencies of each of the relationships in social Q&A setting with the following research questions:

a) What semantic relationships exist in health questions in social Q&A? How frequently do the relationships occur?

b) What semantic relationships exist in health answers in social Q&A? How frequently do the relationships occur?

c) What semantic relationships are implicated between questions and answers in social Q&A? How frequently the relationships do occur?

Many previous studies about semantic relationships in information are small in scale, or used a single approach. Given the volume and growth rate of unstructured text information generated in the web 2.0, more efficient methods are required to examine the massive amount of information. Therefore, given the social context of question-answering behavior, we propose a mixed method of content analysis and text mining as an effective means for examining semantic relationships of question-answering behavior in health. We believe that the two research methods can be combined in a way to maximize the ability of humans and computers in data analysis.

2 Methods

In this study, 164,279 health questions and 413,900 associated answers have been randomly collected from the health category of Yahoo! Answers posted during 2013.

Content analysis is useful for exploring latent and contextual relationships between text messages. In social Q&A settings, people describe their issues in natural language, which includes linguistic variations such as slang, spelling variations, and contextual meaning. The process of identifying semantic relationships in questions and answers would involve humans' linguistic ability to cope with the variety of natural language and to understand the context of the expressions used (Slaughter, 2003). In this study, content analysis will be mainly used to capture social and contextual meanings in major concepts identified in social Q&A. Two human coders will identify propositions among major concepts in questions and answers, and then detect and code the semantic relationships in the propositions.

While content analysis involves an iterative and manual review process of texts, , text mining is an automatic tool to explore unknown patterns from unstructured text documents or natural languages in large quantities (Blake, 2011; Kumar & Bhatia, 2013). Text mining is useful when extracting relevant and meaningful information efficiently, revealing hidden relationships between concepts, instances, and their attributes. It enables us to construct the extension and intention of formal concepts. Text mining will be used to capture hidden, implicit semantic relationships between concepts presented in health questions and answers in this study. IBM SPSS Modeler Premium (SPSS Modeler) will be used to extract and analyze the concepts (or the tokenized representation) in the datasets.

An overview of the process of data analysis using the two methods of content analysis and text mining is shown in Figure 1. We will use the two methods by turns according to the sequence in Figure 1.

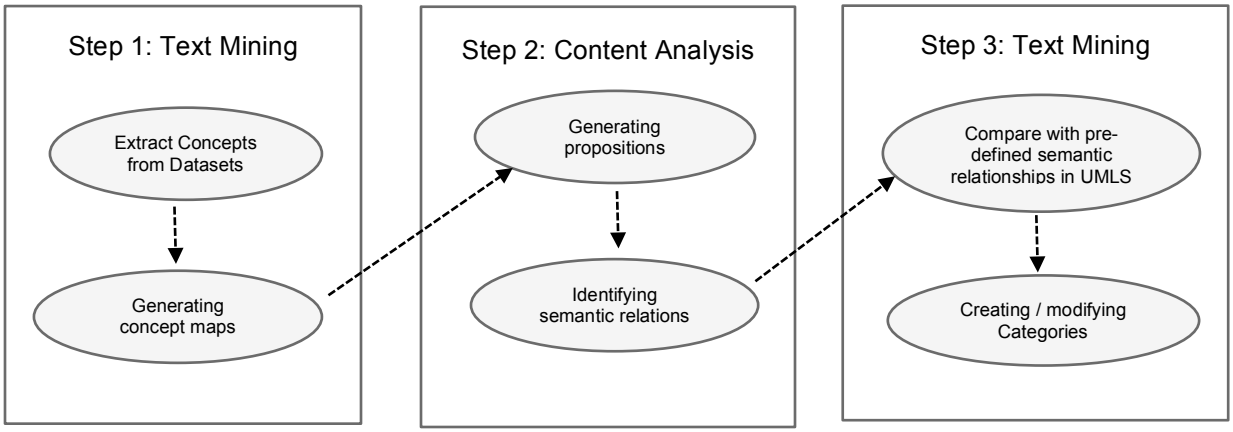

Figure 1. Process of the Study

Step 1 (Text mining): The SPSS Modeler will first extract major concepts based on the Unified Medical Language System (UMLS). Once the set of concepts in questions and that of answers are extracted, concept maps will be generated to detect the latent semantic relationships for the most frequently appearing concepts, using the predictive models contained in the software.

Step 2 (Content analysis): The relationships between concepts will be interpreted based on semantic associations from UMLS. We will also refer to and review the original postings in Yahoo! Answers to confirm and finalize a set of propositions which contain the identified concepts. For example, if a concept "chlamydia" is found to be strongly related to a concept "sexual behavior," the authors look up the original posting in Yahoo! Answers, and then generate a proposition such as 'Chlamydia is a type of sexually transmitted disease,' which has the 'is a' of semantic relationship. The number of similar semantic relationships generated for each major concept will be counted.

Step 3 (Text mining): The extracted concepts and their semantic relationships will be further analyzed and classified by their contexts and then a set of categories of major concepts will be developed. The pre-identified semantic relationships and class in UMLS will guide the category generation.

For the analysis of semantic relationships implicated between questions and answers, we will carry out data analysis, following the three steps in Figure 1, as well.

Step 1 (Text Mining): The SPSS Modeler will generate two separate text mining nodes, one for each of the concepts extracted from questions and the other for each of the concepts extracted from answers. . The two nodes will be combined using the inner joining function of text mining software. The combined node will be run to learn how a specific concept in a question is related to the associated answers. The semantic relationships between the concepts in the questions and the answers will be automatically assigned to the categorization system that is developed from the previous step.

Step 2 (Content Analysis): Since the semantic relationships between the questions and the associated answers are implicit, the authors will extract a small set of sample data and will code the semantic relationships between a question and the associated answers. The pre-identified semantic relationships and class in UMLS will guide the coding.

Step 3: (Text mining): The categories that were generated from the previous step will be modified based on the coding results.

3 Implication & Conclusion

The use of semantic relationships can guide lay people by suggesting concepts not overtly expressed in initial queries (Slaughter, 2002), helping both indexers and searchers to identify various kinds of related terms (Khoo & Na, 2006). In this sense, semantic relationships are important to improve greater effectiveness and refinement in information seeking techniques such as information retrieval, information extraction, questions and answering, and text summarization (Khoo & Na, 2006). Particularly, socially generated concepts and relationships may reflect lay people's own needs, interests, and conceptual associations in their own vocabularies. These semantic structures may emerge through communities developed within the socially-constructed knowledge spaces (Abbas, 2010; Peters, 2009). The detected semantic relationships in questions and answers in this study may help to determine the expected answer and employment of information retrieval methods to retrieve documents for passages likely to contain answers (Voorhees, 2003), overcoming current approaches of term matching and passage extraction (Khoo & Na, 2006)

References

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

Blake, C. (2011). Text Mining. *Annual Review of Information Science and Technology*, *45*, 123–155.

Ding, Y., Jacob, E. K., Caverlee, J., Fried, M., & Zhixiong, Z. (2009). Profiling social networks: A social tagging perspective. *D-Lib Magazine*, *15*(3/4). Retrieved from http://www.dlib.org/dlib/march09/ding/03ding.html

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, *3*(1), 1–11.

Health Information National Trends Survey (2013). *In the past 12 months, have you done the following things while using the Internet? Looked for health or medical information for yourself?* Retrieved from http://hints.cancer.gov/question-details.aspx?qid=757

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

Kumar, L., & Bhatia, P. K. (2013). Text mining: Concepts, process, and applications. *Journal of Global Research in Computer Science*, *4*(3), 36–39.

Mika, P. (2007). *Social Networks and the Semantic Web*. New York, NY: Springer.

Mikroyannidis, A. (2007). Toward a Social Semantic Web. *Computer*, *40*(11), 113–115.

Oh, S., & Park, M. S. (2013). Text Mining as a Method of Analyzing Health Questions in Social Q&A. Presented at the he 76th Annual Conference of the American Society for Information Science & Technology (ASIST' 13), Montreal, Quebec. Retrieved from https://www.asis.org/asist2013/proceedings/submissions/posters/83poster.pdf

Oh, S., Zhang, Y., & Park, M. S. (2012). Health information needs on disease: A coding schema development for analyzing health questions in social Q&A. American Society for Information Science and Technology.

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

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.