Thomson Reuters Doctoral Dissertation Proposal Scholarship

Exploring Social Semantic Relationships for Knowledge Representation in Health Through Mining Social Media

Min Sook Park
PhD Candidate
Information Studies
Florida State University

142 Collegiate Loop Tallahassee, FL 32306-2100
Phone: 850-363-1196
Fax: 850-644-9763
Email: mp11j@my.fsu.edu

I. Description of the Study

Thanks to the explosive growth of Web 2.0 technologies, searching for health information online has become one of the most popular activities of information users. In fact, 72% of Internet users looked for health information, and half of caregivers (52%) have participated in an online social activities related to health (Fox & Duggan, 2013). One of the factors that has contributed to the explosive popularity of second-generation web-based technologies, or the Social Web, as a healthcare resource is the ease of publishing and sharing user-generated information directly to a worldwide audience. This has led to a massive volume of health information for public consumption, accumulating many people's experiences, ideas, and wisdom on the Web (Andersen & Söderqvist, 2012).

The massive amount of health information available online, paradoxically, places limitations on the ability to organize and find relevant information. There were self-organization efforts using information resource applications such as bulletin boards, blogs and podcasts. However, the previous organization efforts faced many challenges in overcoming the scalability of user-generated content (Li & Lu, 2008). One of main reason is that knowledge representation (KR) such as metadata has been traditionally created by information professionals, using strict rules and reflecting experts' views (Greenberg, 2003; Mathes, 2004a; Mikroyannidis, 2007b). This top-down approach, governed by strict rules, is usually incongruent with the rich, dynamic, and flexible nature of human conceptual system and innate human intelligence of natural language processing (Assefa, 2007).

In addition, health is a complex domain, which requires advanced knowledge so that it often causes difficulty for health information users who are not very familiar with medical concepts. Also, the amount of resources presented in an unstructured manner easily confuses

online health information users (Slaughter, 2002). Consequently, the information-seeking process for laypeople can be particularly complex often because they lack of understanding medical concepts and how the concepts are related to each other (Slaughter, 2002).

To address these issues, socially generated metadata (i.e., tags) in the Social Web is being considered a form of KR (Weller, 2010; Yoon, 2010). Users attach keywords to resources, thus reflecting users' needs, interests, vocabularies, and conceptual association (Mika, 2007; Shirky, 2005). However, the lack of vocabulary control in tags impairs precision and recall (Hunter, 2011; Mathes, 2004; Taylor & Joudrey, 2009). Another proposed idea is ontologies in the Semantic Web, which is similar in terms of focusing on uniquely labeling individual knowledge resources (Weller, 2010). This approach involves enabling automatic processing of resources on the Web and combining them into new meaningful units rather than relying on human understanding (Weller, 2010). This new approach is also not without limitations. Most of all, machines are unaware of the actual *context* and *meaning* of different web resources (Boulos, Roudsari, & Carson, 2002).

To maximize the merits of those two approaches, researchers have recently begun the discussion of harvesting socially created rich vocabularies from users so as to bridge the richness by creating or enhancing the existing structures for the next generation of the Web (Abbas, 2010; Peters, 2009; Sharif, 2009; Weller, 2010). The expectation is that Social Web elements will lead to the creation of semantically rich and handier KR (Gruber, 2008; Hunter, 2011; Kamel Boulos & Wheeler, 2007; Mikroyannidis, 2007; Pileggi et al., 2012; Stuart, 2012). The discussions have been centered around how to integrate two different KR mechanisms for information resources on the Web: tags in the Social Web and ontologies in the Semantic Web.

Significance of the Study

The proposed study has its significance in three aspects. First, it examines a new approach, the Social Semantic Web, to explore the possibility of creating socially enriched KR in the medical domain. This new approach to KR marks an important new effort to overcome the limitations of existing KR for the Web. Mining the agreed concepts from a large amount of tag data and associated data (e.g., rating data, users, resources) will enable the extraction of ontological structures (Gruber, 2007; Mika, 2007; Mikroyannidis, 2007).

Secondly, detecting collectively imposed semantic structures from social media, which the health consumers actually use and recognize, may aid in improving the semantic networks in the existing medical subject headings, such as the UMLS. This ultimately may contribute to reducing the gap between legitimate medical knowledge organization system (KOS) and health consumers' concepts. It would also help health information users find relevant information, and eventually let users engage in serendipitous reuse and discovery of related information (Shadbolt, Hall, & Berners-Lee, 2006).

Thirdly, this study employs a mixed methods approach that integrates technologies and human intellectual power to maximize the generalizability of the study while harnessing a large scale of unstructured data. A majority of previous related studies took either a quantitative approach (e.g., statistical analysis, linguistic parsing, and data mining) without including linguistic analyses or a qualitative methodology (e.g., interview) rather than attempting to combine the advantages of different approaches. It should also be noted that except a few studies (e.g., Ding et al., 2009), many of the studies reported here are small in scale. These limitations often lead to incomplete descriptions or limited generalizability of studies to the issues the studies aimed to understand.

Research Design and Methodology

Purpose and Objectives

This study aims to explore semantic relationships in tags, in associated data, and between tags and associated data on the Social Web with the primary research question "what are the semantic structures in socially generated health resources and associated user tags?" Identifying semantic relationships in user-generated content and associated tags serves a number of purposes:

a. The frequency of relationship instances within socially generated resources details which semantic relationships are expressed in the resources generated by health information users;
b. The frequency of relationship instances within the content and among associated tags may indicates usage patterns (Slaughter, 2002); and
c. The semantic relationships that are implied between concepts within the content and tags demonstrate the inferences required to represent the contents and the hierarchical relationships between the KRs and the identified concepts.

Although socially generated tags have often been discussed as a potential user-centered approach for metadata or knowledge description (Shadbolt et al., 2006), words such as tags do not stand alone when capturing the semantic content of documents, since understanding the content of document is necessary to express the format of Web content in more machine-understandable forms (Assefa, 2007; Jacob, 2003; Shadbolt et al., 2006; Soergel, 1999). Besides, for many studies on information retrieval, understanding the content of documents is considered one of critical functions of a system which in turn requires understanding natural language (Assefa, 2007). Natural language processing and understanding is inherent to human intelligence

(Assefa, 2007). Thus, it is plausible to learn how humans use, represent, and relate natural language.

Methods

Text mining and content analysis are integrated in this study to amplify the synergy between numerical techniques and qualitative expectations in analyzing textual data, Text mining is an effective and automatic way to explore and find unknown patterns from a massive amount of unstructured textual data at high speeds, which is beyond human capability (Gupta & Lehal, 2009). Automatic tag analysis and clustering does not generate full ontologies (Weller, 2010). However, co-occurrence of tags, similarity measures for deriving semantics, and tag clusters may provide the basis for harvesting richer semantics for a more fine-grained KOS. In this study, text mining will be used to identify concepts and capture hidden and implicit semantic relationships between concepts presented in health-related texts, in associated tags, and between text and tags. However, the biggest challenge in eliciting ontological structure from text is capturing the true semantics of the content (Stavrianou, Andritsos, & Nicoloyannis, 2007). In this sense, intellectual refinement is also required for identify ontological structures (Weller, 2010).

For qualitative inquiry exploring the latent content and contextual relationships of textual data, content analysis will be used in two ways: First it aims to capture the most proper meaning of a concept through manual reviewing of actual sentences the concepts are embedded in. Secondly, content analysis will be used to analyze the characteristics of relationships between frequently co-occurring concepts that are identified using text mining. Manual reviewing will help identify the semantics of a term with the highest possible specification and detect the inherent meaning of a term, both of which are required when a KR needs to be content-specific and accurately represent a Web resource based on the inherent meaning of a term (Jacob, 2003;

Weller, 2010). In addition, in order to develop explicit semantic modeling from implicit relationships between concepts, there is a need to develop an in-depth understanding about the nature of the relationships between paired concepts.

Test Bed and Data Collection

Data will be collected from Tumblr (www.tumblr.com). Tumblr is one of the most popular a micro-blogging and social networking service and hosts over 221.4 million blogs with 102.5 billion posts as of February 2015 (Tumblr, 2015).

User-generated health information on the Social media such as Tumblr has been traditionally harnessed as a live resource for understanding consumer health vocabulary and their behavior, with the purpose of bridging the gap between laypeople and medical professionals (Doing-Harris & Zeng-Treitler, 2011; Hu & Liu, 2012; Pentland et al., 2013).

Among many different social media platforms, Tumblr has been chosen as a test bed for the current study. Firstly, Tumblr is designed to support users producing their own content as much as they want by not limiting the number of words that can be posted in a blog. In this sense, a user is able to describe related issues and thoughts that express his/her concepts in a blog, which may represent his/her concepts in a richer way. Furthermore, Tumblr allows its users to create their own metadata, with as many tags as they want attached to a specific piece of information based on their subjective experience or ideas. This allows users to categorize or index pieces of information in ways that may be more meaningful to them compared to tags generated in an automatic or recommendation tagging system.

Data will be collected from the Tumblr website using its application program interface (API), which refers to a set of routines, protocols, and tools for building software applications. In particular, data collection will be concentrated on content regarding health issues and associated

tags. Particularly, blog posts that include health-related tags (e.g., medicine and healthcare, cancer, diabetes, health, fitness, and flu) will be collected. Although Tumblr allows its users to post text, images, and multimedia, the data collection is concentrated on text posts, since textual data may better for identifying explicit concepts regarding health rather than images or multimedia.

Data Analysis

Text Mining and content analysis will be used by turn to maximize the strengths of each method. The combination of a quantitative approach (i.e., text mining) and qualitative approach (i.e., content analysis) aims to deal with a massive amount of textual data as well as capture accurate semantics and characteristics of relationships between concepts mined from natural language which are contextual and background dependent. The two methods will be used by turn according to the sequence described below:

In the first phase, a text mining program, IBM® SPSS® Modeler 17.0 (SPSS Modeler hereafter), will extract the major concepts or tokenized representations from a large size of textual data generated by laypeople about health in the test bed. First, a set of tags and a set of associated blog content from the test bed, respectively, will be imported to SPSS Modeler. In order to identify meaningful medical concepts from natural language, Medical Subject Headings (MeSH) will be applied. Secondly, the texts from the tags and the blog text content, respectively, are broken into discrete words, and the set of tokenized representations will be preprocessed using the automatic stop word filtering and stemming of SPSS Modeler. Then the researcher will manually review the extracted terms to further remove any unfiltered stop words and natural language variations so as to have better refined topical content. Through these preprocessing, a list of concepts from each of tag data and blog content data will be prepared and the frequencies

of each concept will be counted for further analysis. The most frequently used tags will be used to identify major meta-concepts. The identified concepts from tags will be utilized as central concepts when generating concepts in the following analysis phase. Once the sets of concepts in tags and in blog content are extracted, the concept maps will be generated to identify the latent semantic relationships between the most frequently co-occurring concepts, using the predictive models in the software.

In the second phase, a content analysis method will be used to interpret the type of concept and the relationships between concepts. For the interpretation, the researcher will extract, from each set of tags and blog content, a small set of data that contains the set of concepts and the associated concepts that are identified in the first phase. The researcher will manually review the selected original content of the blog and associated tags, and then code the relationships between concepts. The coding will be guided by the predefined Semantic Type and Semantic Network of the UMLS. On the concept level, each concept will be matched to the Semantic types of the UMLS based on their definitions. On the between concept level, each identified relationships will be interpreted and coded according to the predefined the Semantic Relationships of the UMLS. Semantic types and relationships will be coded using the most specific type or relationship possible. If there are semantic types or relationships that are not found in the predefined UMLS Semantic Network, the semantic types and the relationships will be added and interpreted through manual reviews.

In the last phase, the identified concepts and their relationships will be classified using the classification techniques of the text mining software. The identified semantic types and the semantic relationships in the second phase will be used as categories in the current phase. Once the categories are set up in the software, each concept and the paired concepts will be assigned to

a specific type or relationship and counted their occurrences in the collected dataset. This phase aims to detect usage patterns by identifying the frequency of semantic types and relationship instances within the content and among associated tags. Using the categorization results, a semantic map will be also created which will show the perceived semantic roadmap of health information users.

As for the implicit semantic relationships between tags and blog content, a separate text mining node will be created to learn how the extracted concepts from a set of tags are related to those from a set of blog content. The semantic relationships that are implied between concepts within the content and tags demonstrate the inferences required to represent the contents and the hierarchical relationships between the KRs and the identified concepts (Slaughter, 2002). In doing so, the second and third phases described above will be repeated to analyze and interpret the relationships between extracted concepts.

Reliability and Validity

To ensure the reliability of the study, a fellow doctoral student will be recruited to code 10% of the sample data to allow the researcher to assess the reliability of analysis. The researcher will provide the list of the UMLS Semantic Network and its definitions, and the inter-coder will be allowed to add if he/she notices new types of semantic relationships from the dataset. Once the coding is completed, the inter-coder reliability agreement between the researcher and the recruited doctoral student will be computed using Cohen's κ.

Both text mining and content analysis are similar in dealing with document collections that have already been collected. Developing a valid coding procedure needs attentions in content analysis (Huck, 2011; Schutt, 2009) Since this study employ the UMLS as a coding scheme, the validity of the coding frame of this study can be regarded as valid.

Summary

The proposed study falls under the broad context of knowledge representation systems in the health domain, adopting the Social Semantic approach. This new approach to knowledge representation (KR) would provide a framework to create and maintain semantically richer KR for online health resources.

Motivation behind this study is to look at support for user-oriented health KR systems and ultimately improve the users' access to health information. To do so, the study harnesses user-generated metadata (i.e., tags) and large amounts of associated unstructured data regarding health issues on the Social Web, and investigate he identified semantic relationships based on the existing ontology structure of the UMLS. In this sense, this study can be characterized as a user-centered bottom-up approach to structured KR, investigating semantic relationships between concepts that are generated by many health information users in the Social Web environments.

In order to examine relationships in textual data, text mining and content analysis will be used. The combination of humans' language capability and computers' speedy processing ability enables the researcher of the current study to effectively identify semantic relationships within socially generated textual data and KRs on the Web by better comprehending natural language and dealing with large volumes of text. Automatic detection of semantic relationships based on text mining algorithms, such as co-occurrence and similarity measures, may provide the basis for deriving richer semantics. However, intellectual refinement is also required for a more fine-grained KOS (Weller, 2010). In this sense, the integration of text mining and content analysis will offset the weakness inherent within each method and produce well-validated, and well-substantiated, and more legitimate results (Creswell, 2009). Reliability and validity of the study will be secured computing inter-coder reliability.

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).

Assefa, S. (2007). *Human concept cognition and semantic relations in the Unified Medical Language System: A coherence analysis*. University of North Texas.

Boulos, M., Roudsari, A., & Carson, E. (2002). Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. *Med Sci Monit*, *8*(7), 24–36.

Creswell, J. (2009). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage publications.

Ding, Y., Jacob, E. K., Zhang, Z., Foo, S., Yan, E., George, N., & Guo, L. (2009). Perspectives on social tagging. *Journal of the American Society for Information Science*, *60*(12), 2388–2401.

Doing-Harris, K., & Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, *13*(2). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221384/

Fox, S., & Duggan, M. (2013). *Health online 2013*. Washington, DC: Pew Internet & American Life Project.

Greenberg, J. (2003). Metadata and the World Wide Web. In *Encyclopedia of Library and Information Science* (pp. 1876–1888). New York, NY: Marcel Dekker, Inc.

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *Int'l Journal on Semantic Web & Information Systems*, *3*(2).

Gruber, T. (2008). Collective knowledge systems: Where the Social Web meets the Semantic Web. *Semantic Web and Web2.0*, *6*(1), 4–13.

Gupta, V., & Lehal, G. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76.

Huck, S. (2011). Reliability and validity. In *Reading statistics and research* (4th ed., pp. 68–89). New York, NY: Pearson.

Hunter, J. (2011). Collaborative semantic tagging and annotation systems. *Annual Review of Information Science and Technology*, *43*(1), 1–84.

Hu, X., & Liu, H. (2012). Mining text data. In *Text analytics in social media* (pp. 385–414). New York, NY: Springer.

Jacob, E. K. (2003). Ontologies and the Semantic Web. *Bulletin of the American Society of Information Science and Technology*, *4/5*(19-22). Retrieved from http://www.asis.org/Bulletin/Apr-03/jacob.html

Kamel Boulos, M., & Wheeler, S. (2007). The emerging Web 2.0 social software: An enabling suite of sociable technologies in health and health care education. *Health Information & Libraries Journal*, *24*(1), 2–23. http://doi.org/10.1111/j.1471-1842.2007.00701.x

Li, Q., & Lu, S. C.-Y. (2008). Collaborative tagging applications and approaches. *Institute of Electrical and Electronics Engineers*, *15*(3), 14–21.

Mathes, A. (2004a). Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication*, *47*(10).

Mathes, A. (2004b). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Retrieved from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

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.

Pentland, A., Reid, T., & Heibeck, T. (2013). *Big data and health: Revolutionizing medicine and public health*. World Innovation Summit for Health.

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

Pileggi, S. F., Fernandez-Llatas, C., & Traver, V. (2012). When the social meets the semantic: Social Semantic Web or Web2.5. *Future Internet*, *4*, 852–864.

Schutt, R. (2009). *Investigating the social world: The process and practice of research* (6th ed.). thousand Oaks, CA: Pine forge press.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web revisited. *Intelligent Systems, IEEE*, *21*(3), 96–101. http://doi.org/10.1109/MIS.2006.62

Sharif, A. (2009). Combining ontology and folksonomy: An integrated approach to knowledge representation. In *The emerging trends in technology: Libraries between Web 2.0, semantic web and search technology*.

Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. Retrieved from http://www.shirky.com/writings/ontology_overrated.html

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.

Soergel, D. (1999). The rise of ontology or the reinvention of classification. *Journal of the American Society for Information Science*, *50*(12), 1119–1120.

Stavrianou, A., Andritsos, P., & Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *ACM SIGMOD Record*, *36*(3), 23–34.

Stuart, D. (2012). FOAF within UK academic Web space: A Webometric analysis of the Semantic Web. In G. Widen & K. Holmberg (Eds.), *Social information research* (Vol. 5, pp. 173–191). Emerald Group Publishing Limited.

Taylor, A., & Joudrey, D. (2009). *The organization of information*. Westport, CT: Libraries Unlimited.

Tumblr. (2015). Tumblr. Retrieved from https://www.tumblr.com/about

Weller, K. (2010). *Knowledge representation in the Social Semantic Web*. Berlin, German: Walter de Gruyter GmbH & Co.

Yoon, J. (2010). Indexing. In M. Norton (Ed.), *Introductory concepts in information science* (2nd ed., pp. 67–86). Medford, NJ: American Society for Information Science and Technology.

II. Schedule of Completion

Tasks		2014							2015												2016			
		Summer			Fall				Spring				Summer				Fall				Spring			
		6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
Project Proposal Preparation																								
Pilot study/ IRB approval																								
Dissertation Proposal Defense																								
Data Collection	Crawler Development																							
	Database Development																							
	Data crawling																							
Data Analysis	Text Mining																							
	Content Analysis																							
	Text Mining																							
Writing Dissertation	Writing Findings																							
	Writing Discussion																							
	Feedback & Review																							
Dissertation Defense																								

Notes: dark gray shading = completed, light gray dot shading = projected

III. Budget Justification

A. Wage

One programmer at masters-level is required for 50 hours at $15/hour ($600). This student will assist in developing programing scripts for data collection and database development for data storage.

B. Equipment

An external backup hard drive (5TB, 199.99) will be required.

C. Materials and Supplies

Funds totaling 573.98 are required during year 1 to purchase two data analysis programs: IBM® SPSS® Modeler 17.0 and IBM®SPSS® Statistics Pack 23 ($286.99.99 each)

IV. Other support: N/A