Supporting better treatments for meeting health consumers' needs: extracting semantics in social data for representing a consumer health ontology

Yunseon Choi

Abstract

Introduction. The purpose of this paper is to provide a framework for building a consumer health ontology using social tags. This would assist health users when they are accessing health information and increase the number of documents relevant to their needs.

Methods. In order to extract concepts from social tags, this study conducted an empirical study on terms collected from a social networking site. The semantics of tags were analyzed and a concept list was developed by using the middle-out strategy.

Analysis. This study analysed the semantic values of tags by employing Latent Semantic Analysis (LSA). This is a method for extracting and representing the contextual-usage meaning of words by analyzing relationships between documents and the terms they contain and word semantics.

Results. The process of building an ontology using social tags shows how using this consumer health ontology could improve user access and retrieval. It demonstrates how terms extracted from tags are related to each other with similarity and relationships within hierarchies in the ontology.

Conclusion. The study has implications for better design of ontology applications that support the search for health-related resources. This will enhance the communication between health consumers and professionals.
Introduction

As a large number of online health resources have become available, there has been a great increase of the number of health consumers replying on online health resources available on the World Wide Web (Andreassen, Bujnowska-Fedak, Chronaki, Dumitru, and Pudule, 2007; Fox, 2011; Rice, 2006; MacLean and Heer, 2013). It has been reported that health consumers should be able to have effective access and utilise relevant health information to meet their needs (Nutbeam, 2008; World Health Organisation, 2011). A Pew Research Center survey indicates that 72% of U.S. adult Internet users have looked for health information online (Fox and Duggan, 2013). Studies also show that most consumers lack the skills to access and use effectively online health resources (Friel, Bond, and Lahoz, 2015; Gray, 2005; Jain and Bickham, 2014; Ratzan and Parker, 2000; Rowlands et al., 2013).

There have been efforts to provide access to reliable health information on the World Wide Web, and MedlinePlus and InformedHealthOnline are such examples. MedlinePlus is maintained by the National Library of Medicine and it is a Web-based consumer health information service (Miller, Lacroix, and Joyce, 2000). InformedHealthOnline is published by the German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (or IQWiG) and is the English-language version of the German website which provides health information to the public and patients.

Information in health or medical domains is critical and should be provided to health consumers without difficulty. However, the growing amount of health information on the web has increased concern about effective access to quality health information because terminology, currently used for organising health or medical information, is generated by professionals and may not be familiar to users. The terminology gap between users' and professionals' vocabulary in describing medical-related web documents was also uncovered by a study on indexing consistency of social tagging in comparison with professional indexing (Choi, 2014). Health consumers and healthcare professionals tend to use different terms to describe health-related concepts, for example, dry mouth vs. xerostomia and flu vs. influenza (Vydiswaran, Vinod,
Hanauer, and Zheng, 2014). This terminology gap in the health domain prevents health consumers from accessing health information relevant to their information needs. For example, when a health consumer tries to find information related to nosebleed symptoms, she/he may not find the resources including only the term epistaxis in the meta tags, title and text (Zielstorff, 2003). In large medical health consumer websites, it has been reported that when a consumer's terms are different from physician-defined terms, the search returned no results, for example, heart attack vs. myocardial infarction (Zeng, Kogan, Ash, and Greenes, 2001) and shakes vs. tremor (Zielstorff, 2003).

On the other hand, as networked information resources on the web continue to grow rapidly, digital information environments have led librarians and information professionals to manage digital resources on the web. Thus, this trend has required new tools for organizing and providing more effective access to the web. Subject directories or Web directories are such tools for internet resource discovery since subject directories organise Web documents by subject areas. Yet, studies have shown that subject directories based on traditional organisation schemes are not sufficient for the web (Golub, 2006; Nowick and Mering, 2003; Macgregor and McCulloch, 2006). This is because they were developed using traditional library schemes which have been developed with a focus on physical library collection. Web documents, however, were originally organized and indexed by professionally-generated keywords. This means they do not reflect intuitively and instantaneously expressed users' current needs (Macgregor and McCulloch, 2006).

Although there have been efforts to involve users in developing information organization systems, they are not necessarily based on users' real languages. Accordingly, social tagging has received significant attention as a promising way to solve this challenge since users' tags reflect their interests and their languages. Social tags are good sources for identifying users' terms. Several researchers have discussed the impact of tagging on retrieval performance on the web (Bao, 2007; Choi, 2009;
Choy and Lui, 2006; Golder and Huberman, 2006; Heymann, Koutrika, and Garcia-Molina, 2008; Sen et al., 2013; Yanbe, Jatowt, Nakamura, and Tanaka, 2006). Although social tags have been discussed regarding its usefulness as additional access points for classification and retrieval (Trant, 2009; Choi, 2014), there has been little research conducted on the use of social tags to improve practices in information organization. Since social tags provide additional access points as user-generated terms, using them would improve information access and promote effective reasoning for retrieval.

In terms of information organization, ontologies have been used for information organization and information integration. Ontology is a shared understanding of a domain that can be communicated between people and computers (Ding, 2001). Especially, in the medical and health services, information systems should be able to communicate difficult and complex concepts. However, analysing the structure and concepts of medical terminologies cannot be easily achieved.

There have been very few studies conducted on building health or medical ontologies which features concepts and vocabularies familiar to health consumers. Mayo consumer vocabulary, a taxonomy of consumer health terms and concepts, was developed and maintained by Mayo Clinic (Sedorff et al., 2013). The Consumer Health Vocabulary Initiative resulted in the creation of the Open access collaborative consumer health vocabulary, which was designed to complement the existing framework of the Unified medical language system and to aid the needs of consumer health applications (US. National Library of Medicine, 2012). However, this vocabulary is not implemented using a knowledge representation language such as Web Ontology Language which supports semantic search and knowledge reasoning.

The aforementioned important components of effective health information organization are applied in this study:

- Due to the unfamiliarity of health consumers to current terminology used for organizing health or medical information, medical information systems need to include user-friendly vocabulary.
Considering the characteristics and quality of social tags in representing users' views, social tags should be utilised to improve practices in information organization.

- To establish a closer link between health consumers' information needs and professionals' responses, a powerful semantic-based ontology needs to be built.

This paper is part of a larger research project which aims to answer questions about how we can assist users when they are accessing health information in order to increase the number of documents they find relevant to their needs. The ultimate goal of the project is to build a consumer health ontology by utilising social tags assigned to health-related documents. The main objective of this paper is, therefore, to provide the framework for a consumer health ontology by discussing the process of building an ontology featuring social tags. This paper intends to show how social tags can be utilised for developing class hierarchies in the ontology in order to identify unambiguously implicit relations among social tags.

**Ontologies for information organization and information integration**

**Definitions of ontologies**

The term *ontology* has been used in several disciplines, from philosophy to computer science. As a branch of philosophy, ontology studies the structures of the objects, properties and relations of reality (Smith, 1997). In computer science, into which the term came from artificial intelligence, the ontology is a model of the representation of objects in the world with properties and relationships (Garshol, 2004). An ontology is defined as a formal, explicit specification of a conceptualisation (Gruber, 1993; Studer, Benjamins, & Fensel, 1998):

- **Conceptualisation** refers to 'an abstract, simplified view of the world that we wish to represent for some purpose' (Gruber, 1993, p. 1).
- **Explicit** refers to the 'type of concepts used, and the constraints on their use are explicitly defined' (Studer, et al., 1998, p. 25).
Formal refers to the fact that 'the ontology should be machine readable' (Studer, et al., 1998, p. 25).

- **Shared** means that 'an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group (Studer, et al., 1998, p. 25).

Other researchers describe ontologies as *taxonomic hierarchies* (Baeza-Yates & Ribeiro-Neto, 1999; Vickery, 1997). Vickery notes the aspect of taxonomic hierarchies of classes, with class definitions and the subsumption relations. Baeza-Yates and Ribeiro-Neto describe ontologies as hierarchical taxonomies of terms representing topics.

All above definitions show that there may be different views or several interpretations concerning the concept of the ontology. In this study, we take the view of taxonomy defined by Vickery and Baeza-Yates and Ribeiro-Neto as above. The benefit of this approach is that it allows us to understand that ontologies are closely related to conventional information organization and access tools, such as classification schemes or thesauri, in that they all organize concepts according to a certain rule in a hierarchical structure. In thesauri, however, the semantic differences of hierarchical relations have occurred, because BT/NT (broader term/narrower term) relations were differently defined in different thesauri. In some thesauri it means subsumption (subclass and subproperty), while in other thesauri it can mean BTI (broader term instance) or BTP (broader term partitive). The discussion on subsumption in hierarchies has been a well-known issue in the area of knowledge representation. Brachman (1983) has discussed semantics of the subsumption to provide some clarity in organizing taxonomies. Ontologies are more expressive than classifications or thesauri, because ontologies allow more explicit semantics and relationships between concepts in formal, machine understandable languages. Accordingly, ontology-based, semantic searches retrieve the results by analysing the context and semantics of the query.

**Types of ontologies**
Ontologies exist at several levels of abstraction and are described as three types: upper, mid-level, and domain. An upper ontology, sometimes referred to as universal ontology (Colomb, 2002), provides a framework for a common knowledge base which consists of basic and universal concepts that can be applied to a wide range of specific domains (Semy, Pulvermacher, and Obrst, 2004; Singh and Singh, 2014). An upper ontology is a high-level, domain-independent ontology and there are several standardised upper ontologies including Dublin core, Suggested upper merged ontology (SUMO) and Unified medical language system (UMLS), etc. The Dublin Core element set defines elements for cataloguing library items and other electronic resources. The Suggested upper merged ontology was developed by merging a number of existing upper-level ontologies (Niles and Pease, 2001). The Unified medical language system was developed by the US National Library of Medicine to provide integrated access to biomedical resources.

A mid-level ontology serves as a bridge between abstract concepts defined in the upper ontology and low-level domain specific concepts specified in a domain ontology' (Semy, Pulvermacher and Obrst, 2004, p. 2-3). For example, the Gellish ontology is a combination of both an upper and a domain ontology. A domain ontology specifies concepts particular to a domain of interest and represents those concepts and their relationships from a specific domain (Semy et al., 2004). Domain ontologies can be driven from mid-level or upper ontologies by using or extending concepts and vocabulary expressed in mid-level or upper ontologies.

**Ontologies in medical or health domains**

In the field of health and medical services, ontologies have been built as knowledge bases for health professionals as a way of representing and organizing medical terminologies. Specialised medical ontologies and terminologies include:

- GENIA ontology for the microbiology domain, Medical Entities Dictionary as a large repository of medical concepts.
- Gene Ontology providing a common language to describe aspects of a gene product's biology.
SNOMED CT (Systematised nomenclature of medicine—clinical terms) is a comprehensive clinical terminology, originally created by the College of American Pathologists (CAP). (U.S. National Library of Medicine, 2016)

RxNorm provides normalised names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software. (U.S. National Library of Medicine, 2014)

Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine. (Bodenreider, 2004)

Additionally, there have been several research efforts focusing on developing frameworks to help health consumers search for information (Puustjarvi and Puustjarvi, 2011; Dong and Hussain, 2011). The Personal Health Server was developed for helping patients obtain and understand health information, and make appropriate health decisions (Puustjarvi and Puustjarvi, 2011). Also, ontology-based, semantic-Web technology was applied to develop the health semantic search engine, specifically to describe service domain knowledge in digital health ecosystems (Dong and Hussain, 2011).

Methods

Overview

The framework for building a consumer health ontology is depicted in Figure 1. This diagram outlines the phases of the research on how a consumer health ontology can be built by using social tags.
Phase 1 focuses on collecting the concepts from social tags. The extracted concepts were used to define classes of the ontology. In order to extract concepts from social tags, this study conducted an empirical study on terms collected from a social networking site. This study analysed the semantic values of tags by employing latent semantic analysis, which is used for extracting latent semantics of words by statistical computation. There is no other study using this method to develop health-related ontologies. Latent semantic analysis uses natural language processing, and analyses relationships between documents, the terms they contain, and word semantics (Deerwester, 1990).

The focal point in this research is not to criticise the quality of professionals' keywords but to point out the lack of additional access points or complementary terms in controlled vocabularies which are used by professionals. Since the keywords provided by professionals are regarded as accurate terms when describing topics within documents, it is worthwhile to see whether there are semantic relations between tags and professionals' keywords for the documents which are described by both tags and keywords. If tags are conceptually similar to professionals' keywords, those tags are also regarded as key terms or good descriptors in describing the document.

Accordingly, latent semantic analysis was conducted to investigate to what extent tags are conceptually related to professionals' keywords. The basic idea of the method is
that if two terms tend to occur in similar documents, the terms are similar. Thus, this study computed semantic relatedness between tags and professionals' keywords in terms of a specific document, and higher values of latent semantics between tags and professionals' keywords would demonstrate that those tags can be considered to be good index terms. Since the keywords provided by professionals are regarded as accurate terms describing topics of documents, if tags are conceptually similar to professionals' keywords, those tags are also regarded as good terms in describing the document.

Table 1 shows the examples of semantic analysis cosine values between two vectors. It shows that the semantic similarity (0.74) between two terms, which are *library* and *book*, is higher than the semantic similarity (0.02) between *library* and *beach*.

<table>
<thead>
<tr>
<th>Vector 1</th>
<th>Vector 2</th>
<th>Cosine values</th>
</tr>
</thead>
<tbody>
<tr>
<td>library</td>
<td>book</td>
<td>0.74</td>
</tr>
<tr>
<td>library</td>
<td>beach</td>
<td>0.02</td>
</tr>
<tr>
<td>library</td>
<td>information</td>
<td>0.30</td>
</tr>
<tr>
<td>library</td>
<td>skirt</td>
<td>0.11</td>
</tr>
<tr>
<td>library</td>
<td>catalog</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 1: Examples of latent semantic analysis values between two vectors

Latent semantic analysis was performed by using a Web-based tool, **LSA@CU** with the semantic space 'general reading up to 1st year college (300 factors)' Touchstone Applied Science Associates corpus with one-to-many comparison (comparing a particular text against many other texts, i.e., how associated are a target text and all other texts), term-to-term comparison (comparing two terms, i.e., how semantically similar are two terms). This corpus contains approximately ten million words and is a set of short English documents, extracted from novels, newspaper articles, and other sources. The corpus was collected to develop The Educator's Word Frequency Guide (Turney and Littman, 2003).

Phase 2 (Figure 1) leads to extending the concepts using the existing categories. In this step, the study consults the following three reference tools which are standard vocabularies for health and diseases:

- *International classification of functioning, disability, and health*, a classification of health and health-related
domains and also include a list of environmental factors (World Health Organization, 2001).

*International classification of diseases* (World Health Organization, 1999), the standard diagnostic tool for epidemiology, health management and clinical purposes and is used to classify diseases and other health problems including death certificates and health records (World Health Organization, 1999), and

*Medical subject headings*, a controlled vocabulary thesaurus, which is provided by the National Library Medicine and is used for indexing articles for the PubMed medical journal (National Library Medicine, 1999).

In this study, health-related terms listed in these reference tools are used for extending concepts extracted from social tags in order to build a class hierarchy.

Phase 3 is designed for analysing ontological relations among concepts in a class hierarchy. This study uses the middle-out strategy (Uschold and Gruninger, 1996) which is the combination of the top-down and bottom-up approaches.

**The strategy for building an ontology**

There are three common strategies for building ontologies: *top-down*, *bottom-up*, and *middle-out*. In a top-down approach, core terms or relevant concepts are identified and organized into a high-level taxonomy, and then more specific terms and axioms are identified from there. A top-down approach results in "a structure which represents a bird's eye view of the world and which should make the task of defining domain-specific content relatively trivial" (Niles and Pease, 2001, p.2). Ontologies built using a top-down approach can be reused for developing domain-specific ontologies in different applications. In a *bottom-up* approach, domain-specific concepts are identified and then extended or developed more from there. While a bottom-up approach identifies from the most concrete to the most abstract concepts, a top-down approach identifies from the most abstract to the most concrete concepts. A *middle-out* strategy (Uschold and Gruninger, 1996) combines the top-down and bottom-up approaches.
There are several factors to be considered when constructing an ontology, such as level of detail, commonality, stability and consistency which are associated with efforts or rework (Figure 2). Since a top-down approach requires an expert-based approach, it is costly and there is 'a risk of less stability in the model which in turn leads to rework and greater efforts' (Uschold and Gruninger, 1996, p.20). On the other hand, a bottom-up approach resulting in high level of detail can allow for detection of inconsistencies and expand concepts by incorporating new emerging concepts, but a bottom-up approach '1) increases overall effort, 2) makes it difficult to spot commonality between related concepts, and 3) increases risk of inconsistencies which leads in turn to 4) rework and yet more effort' (Uschold and Gruninger, 1996, p.20). A middle-out approach identifies the most relevant to the most abstract and most concrete concepts. In a middle-out approach, since detail arises only as necessary by specialising or generalising the basic concepts, it does not require as much effort. To put it another way, a middle-out approach starts with the most important concepts first, and defines higher level categories, which does not require so much effort or reworking.

This study uses the middle-out strategy, with which core key terms are selected and then are specialised or generalised. In this approach, main concepts or core concepts are identified. That is, core concepts are listed in the high level of hierarchy, and then the concepts are
specialised or generalised in the lower level of hierarchy. For example, terms body, activity, contextual factors are identified as main concepts based on reference tools such as the *International classification of functioning, disability, and health* and the *International classification of diseases*. Next, concepts are specialised and generalized. For example, body structure is specialised into more specific concepts such as skeleton and joint.

**Data collection**

Social data were collected from Delicious, which is one of the most popular social bookmarking services. For a preliminary analysis, 1,326 tags from 153 Web documents were collected. For professionally-generated keywords, terms provided by Intute subject specialists were collected. Intute is a subject directory which includes the collections of quality assessed Web resources organized by subject specialists. Intute offers a searchable and browsable database of Web resources that subject specialists select, evaluate and describe. Among nineteen subject categories organized by Intute, subject categories such as Medicine including dentistry and nursing, midwifery and allied health are related to health and medical areas and Web documents were randomly selected from those categories. After that, Delicious tags assigned to the Web documents were collected and compared with professionally-generated keywords which are provided by Intute subject specialists. Among professionals' keywords, terms associated with the type of documents or publications, that is, image or any names of journals or conferences were not applied for the analysis of latent semantics.

**Results**

**Semantic analysis of social tags**

The study collected concepts by examining the semantics of social tags in order to build a class hierarchy of an ontology. As discussed in the Methods' section of, in terms of professionals' keywords, terms associated with the types of publication or documents were excluded for the analysis of latent semantics. The examples of these terms include patient education, NIH publication, and teaching materials, etc. Table 2 presents the examples of the
collected professionals' keywords and users' tags regarding Web documents in medicine. Table 2 illustrates that while Delicious and Intute include some common terms between them, Delicious tags also include users' preferred terms which are not found in professionals' keywords. Table 2 also shows the latent semantic analysis values which ranged from zero (or N/A) to 1.00. Where the values were greater than 0.10, the terms were used for building the class hierarchy.

<table>
<thead>
<tr>
<th>Web document</th>
<th>Professionals' keywords (Intute)</th>
<th>Users' tags (Delicious)</th>
<th>Latent semantic analysis values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporo-mandibular joint and muscle disorders (<a href="#">National Institute... 2014</a>)</td>
<td>T disorders; parent education</td>
<td>jaw dental-problems odontologia dentistry temporo-mandibular-joint-disorders</td>
<td>0.12 0.27 N/A 0.33 1.00</td>
</tr>
<tr>
<td>OPETA: abdomen exam. (<a href="#">Cavanagh, et al., 2004</a>)</td>
<td>abdomen physical examination</td>
<td>clinical abdomen gastroenterology</td>
<td>0.29 0.15</td>
</tr>
<tr>
<td>NIH consensus statement on acupuncture (<a href="#">US. National Institutes of Health, 1997</a>)</td>
<td>acupuncture</td>
<td>oriental or Chinese-medicine acupuncture</td>
<td>0.55 1.00</td>
</tr>
<tr>
<td>EKG arrhythmia review (<a href="#">Crimando, 1999</a>)</td>
<td>arrhythmia electrocardiography</td>
<td>cardiovascular useful physiology physical therapy medical school</td>
<td>0.21 0.07 0.14 0.05 0.02</td>
</tr>
</tbody>
</table>

**Table 2: Professionals' keywords vs. users' tags and latent semantic analysis values**

The examples of the latent semantic values of tags are graphically illustrated (Figure 3-4). The term *odontologia* is not an English word and does not exist in the latent semantic analysis corpus. In Figure 3-4, tags representing lower values (i.e., less than 1.00) include *odontologia*.
(Figure 3), and useful, physical therapy, medical school (Figure 4). Also, it indicates that these tags are not related to subject or topics of documents. Since this study aims to focus on building an ontology which is conceptualisation in the domain, those tags not related to subject or topics of documents were excluded for building the class hierarchy of the ontology.

![Figure 3: Latent semantic analysis value of tags regarding the document, temporo-mandibular joint and muscle disorders](image1)

![Figure 4: Latent semantic analysis values of tags regarding the document, EKG arrhythmia review](image2)

Representing a consumer health ontology

In this section, we show how a concept list is developed based on the existing categories by utilising social tags,
and how relations among concepts are identified for ontological reasoning. With the middle-out strategy, core key terms are selected and then are specialised or generalised. For example, terms body, activity, contextual factors are identified as main concepts based on reference tools such as the International classification of functioning, disability, and health, and the International classification of diseases. Next, concepts are specialised and generalised. For example, body structure is specialised into more specific concepts such as skeleton and joint. Social tags were used for developing specialised concepts in the hierarchy. For instance, regarding the Web document Temporo-mandibular joint and muscle disorders (Table 2), concepts collected from social tags were jaw, dental-problems, and temporo-mandibular-joint-disorders. Table 3 shows that collected concepts from social tags are extended with the existing categories. Like concepts, properties are also specialised or generalised. The right column of the table lists identifies relations, for example, has_subclass, affects, is_affected_by, is_located in, is_connected_to, and is_concerned_with. Additionally, the following properties were created for relations:

- Transitive: the property relates class A to class B, and also class B to class C, then we can infer that class A is related to class C via the property. For example, if a class body has subclass body structure, and class body structure has subclass skeleton, then we infer skeleton is subclass of body.
- Symmetric: the property relates class A to class B, then class B is also related to class A through the property. For example, temporo-mandibular joint is connected to ear, and then ear is connected to temporo-mandibular joint.
- Inverse: if there is a property linking class A and B, then its inverse property will link Class B to A. e.g., body structure affects body, and also body is affected by body structure.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation and its definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thing (root)</td>
<td>A has subclass B=def.</td>
</tr>
<tr>
<td>Body</td>
<td>A has a subdivision of</td>
</tr>
<tr>
<td>Body structure</td>
<td></td>
</tr>
<tr>
<td>Skeleton</td>
<td></td>
</tr>
</tbody>
</table>
- Joint
  - Temporomandibular joint

- Body function
  - Ear
  - Mouth
    - Jaw
    - Teeth
    - Gums
    - Lips
    - Tongue
    - Dental problem

- Eye
  - Activity
  - Contextual factors

- Class B which is related to A in a taxonomic category. E.g., skeleton has subclass joint. Transitive, e.g., If a class body has subclass body structure, and class body structure has subclass skeleton, then we infer skeleton is subclass of body.

- A affects B=def. (B is affected by A)
  A causes or produce an effect or change in B. E.g., body structure affects body.
  Inverse. E.g., body structure is affected by body.
  Domain: body
  Range: body structure

- A is located in B=def.
  A is placed at a certain location in B. E.g., teeth is located in mouth.

- A is connected to B=def.
  A is linked with one or more other physical units.
  Symmetric. E.g., temporo-
Table 3: Concepts and relations of consumer health ontology (an example)

In order to implement the ontology, the study uses Protégé-OWL, which supports Web Ontology Language. The Protégé-OWL ontology modeller was used to present the diagrammatic notation (Figure 5) which is based on concepts and relations from Table 3. Since the graph in Figure 5 is mainly diagrammed for addressing the document *Temporomandibular joint and muscle disorders*, only applied relations or properties are indicated in the graph.
Discussion

Since health consumers and healthcare professionals tend to use different terms to describe health-related concepts (Zeng et al., 2001; Zielstorff, 2003; Vydiswaran et al., 2014), it has given rise to a need for bridging the terminology gap between health consumers and healthcare professionals. In an early stage of the project, this paper shows how social tags are used for the design and development of an ontology which would assist health consumers in finding relevant documents to their needs. Social tags, or user-generated terms, provide additional access points (Trant, 2009; Choi, 2014), which improves information access and promote effective reasoning for retrieval. The results of our study indicated how social tags can be successfully utilised for developing class hierarchies in the ontology. It also identified unambiguously implicit relations among social tags.

The following example indicates the importance of our study. It demonstrates that social tags reflect terms and concepts that are more familiar to users and plays a role of communicating difficult concepts. These terms also
provide additional access points which are not found in controlled vocabulary. In Table 2, regarding a Web document Temporomandibular joint and muscle disorders, terms assigned by professionals are temporomandibular joint disorders and patient education. The assigned social tags to the same document included several terms such as jaw, dental-problems, odontologia, and temporomandibular-joint-disorders (Table 2). The temporo-mandibular joint is the joint of the jaw and is frequently referred to as 'TMJ'. The temporo-mandibular joint is connected from jaw to ear.

The consumer health ontology was partially implemented by using Protégé ontology modular (Figure 5) to show the framework of the ontology. Figure 5 shows how terms are related with similarity and relationships in hierarchies, and helps understand how the consumer health ontology would improve user access and retrieval.

There are inverse relations linking two classes, body structure and body function, i.e., body structure 'affects' body function and body function 'is_affected_by' body structure. The temporo-mandibular joint has a symmetric relation of 'is_connected_to' with two classes, ear and jaw. That is, the relation 'is_connected_to' relates class temporomandibular joint to class ear and also relates class ear to class temporomandibular joint. There are also super- and sub-hierarchical relations among classes, for example, between class body and class body structure and between class body structure and class skeleton, etc. The 'has subclass' relation is transitive, so when joint has subclass temporomandibular joint, temporomandibular joint is also subclass of body structure. Class mouth has several subclasses such as gums, teeth, lips, and tongue, and then these subclasses are also linked to class mouth through 'is_located_in' relation. In addition, the relation 'is_concerned_with' is symmetric, that is, the relation 'is_concerned_with' relates class mouth to class dental problem and also relates class dental problem to class mouth. As discussed, it is illustrated that in the ontology, semantics and relations between concepts are explicitly represented, which allows for analysing the context and semantics of the query. Furthermore, since social tags provide
additional access points as user-generated terms, it would improve information access and promote effective reasoning for retrieval. The scope of the consumer health ontology represented in this paper is limited to a specific category of medical condition, for example, oral health.

For further development of domain-specific ontology in the health domain, other specific categories of medical conditions, such as pregnancy and childbirth, can be represented by expanding concepts in the ontology with domain-specific properties. There have been very few studies conducted on building health vocabularies which features concepts and vocabularies familiar to health consumers (Seedorff et al., 2013). Previous work in consumer health vocabulary such as the Open-access and collaborative consumer health vocabulary (U.S. National Library of Medicine, 2012) was not implemented using a knowledge representation language, but our proposed consumer health ontology using Protégé-OWL supporting the Web Ontology Language improves accessibility to related documents, because it allows for semantic search by exploiting semantic characteristics of consumers' search queries and documents. Therefore, our preliminary results indicate the feasibility of developing health consumer-preferred information systems using ontology.

**Conclusions and future research**

Due to the unfamiliarity of some health consumers to current terminology used for organizing health or medical information, medical information systems need to include user-friendly vocabulary. A powerful semantic-based ontology is required in order to support the search for health-related resources and to enhance the communication between health consumers and health professionals. This paper presents a discussion of the process for developing an ontology for consumer health information for health consumers to assist them to access health-related documents which are relevant to their needs. In the middle-out approach, core key terms were identified and then specialised or generalised. In this approach, main concepts or core concepts are identified. The results of our study are summarised as follows:

- The results from the study showed that the proposed
consumer health ontology could improve user access and retrieval, since it allows for semantic search by exploiting semantic characteristics of health consumers' search queries and documents.

- The proposed consumer health ontology implemented using Web Ontology Language explicitly represented semantics and relations between terms extracted from social tags by defining ontological relations. Thus, it demonstrated convincingly how terms extracted from tags are related to each other with similarity and relationships within hierarchies in the ontology.

Health communities need to establish a closer link between health consumers' information needs and health science librarians' or information professionals' responses. It is of interest to health communities to learn and understand the significant impact of ontologies on health information organization for health consumers. Given the number of online health resources, the growing interest in assessing quality health information will have the brunt of the work to provide health consumers with effective access to relevant resources. Nevertheless, little study exists regarding how ontology can best support health consumers' needs with regard to searching relevant resources to manage their health conditions. This paper shows how social tags can be used for the design and development of consumer health ontology. This study will have implications for better design of ontology applications that support the search for health-related resources and enhance the communication between health consumers and health professionals.

Once the concept list is completed and all ontological relations are identified, the consumer ontology will be fully implemented by identifying the domain and ranging constraints for properties and cardinality. In order to validate the content of the ontology, the study will perform the ontology evaluation and conduct semi-structured interviews with both health consumers and domain experts to assess the usefulness and effectiveness of the ontology for representing terms in the domains.

**Acknowledgements**

This research was partially supported by the Connecticut
State University - American Association of University Professors (CSU-AAUP) Research Grants.

About the author

Dr. Yunseon Choi is a Visiting Scholar at the School of Library and Information Studies, University of Wisconsin-Madison. She teaches a course on social media for information agency. She received her Ph.D. in Library and Information Science from the University of Illinois at Urbana-Champaign. She can be contacted at ychoi249@wisc.edu or dr.yunseon.choi@gmail.com

References


Sen, S., Lam, S., Rashid, A.M., Cosley, D., Frankowski,


How to cite this paper


Find other papers on this subject

Check for citations, using Google Scholar

© the author, 2016.

Last updated: 15 November, 2016