Comparative Analysis of Index Terms and Social Tags: Medical Subject Headings vs. BibSonomy and Delicious

Danielle H. Lee*

Contents

ABSTRACT

This paper demonstrates the comparative analysis of the similarity and difference between Medical Subject Headings (MeSH) and social tags. Both types of metadata have the same purpose — that is, succinctly abstracting content of a given document — but are created from heterogeneous viewpoints. The former MeSH terms show the aspects of publication related professionals, whereas the latter social tags are from the perspectives of general readers. When both types of metadata are assigned to the same publications, do they consist of different nomenclatures reflecting the heterogeneous viewpoints or are they similar, since both metadata types describe the same publications? Social tags are also compared with family terms of MeSH terms in the given MeSH hierarchy, so as to understand the specificity of social tags, related to MeSH terms. Lastly, given the fact that readers assign social tags in casual ways without any restricted vocabulary, we tested how many social tags contain consumer health terms, which are familiar to laypeople. Through these comparisons, we ultimately aim to examine how much the highly controlled publication index reflects general readers' cognitive understandings and stress the necessity of general readers' involvement in the publication indexing process.

Keywords: Medical Subject Headings, Social Tags, Bibsonomy, Delicious, Metadata

* University of Washington, Bothell, Computing and Software Systems(suleehs@gmail.com)
논문접수일자: 2015년 4월 23일 최초심사일자: 2015년 4월 23일 게재확정일자: 2015년 5월 8일
한국문헌정보학회지, 49(2): 291-311, 2015. [http://dx.doi.org/10.4275/KSLIS.2015.49.2.291]

www.kci.go.kr

1. Introduction

Indexing is a process to succinctly describe the aboutness of documents by tagging with *index terms*. Primarily, index terms target to “give users systematic and effective shortcuts to the information they need (Cleveland and Cleveland 2001, 97).” As one of the most successful indexing systems, the Medical Subject Headings (MeSH, hereafter) of PUBMED database indexes and abstracts biomedical articles (Burnett, Ng and Park 1999) to deliver effective and efficient access to information.

Besides the effective information access, another important purpose of indexing systems is to describe “why it is likely to be of interests to a particular group of users” (Lancaster 2003, 9). In MeSH indexing, because of the sophisticated indexing process involving domain knowledge, expertise, and highly formalized semantic hierarchy, a limited number of professionals, such as publishers, librarians, and domain experts, are in charge of indexing processes. Hence, it was criticized on the grounds that index term definitions and the term assignments by professionals usually do not reflect users’ general viewpoints (Knecht and Nelson 2002). Indexing by limited population of professionals makes it difficult to scale for ever-increasing and rapidly growing biomedical articles as well as online materials (Mathes 2004). Moreover, the complicated semantic structure, which is MeSH hierarchy, prevents frequently updating the vocabulary with more up-to-date technical jargon and terminologies.

To support more user-oriented and diverse perspectives, cope with scalability problem, and ensure up-to-date terms in semantic structure, several researchers have suggested including general users in the indexing process (Brown et al. 1996; Hidderley and Rafferty 1997; Lancaster 2003). With the advent of Web 2.0 technology, specifically thanks to social bookmarking systems such as CiteULike, Delicious, Mendeley, BibSonomy, etc., involving users in indexing processes has become much more feasible. Social bookmarking systems enable general readers of bibliographic references to annotate information resources with free-text words as *tags*.

Social tags represent a tagger’s conceptual understanding or categorization of information from a personal viewpoint. As with index terms, tags abstract and retrieve information (Golder and Huberman 2006). Therefore, social tagging is defined as “a neologism for a practice of collaborative categorization (Mika 2005; Chevalier et al. 2011; Ferrara and Tasso 2011).” However, unlike index terms, which are solely focused on succinct descriptions of information content, people assign tags for a variety of reasons: their daily practices, needs, experiences, concerns, expressions, and even social factors, besides abstracting the content (Golder and Huberman 2006; Marlow et al. 2006; Kim, Decker and Breslin 2010; Ferrara and Tasso 2011). Because one’s personal

www.kci.go.kr

experience is closely intertwined with that of other people's, tags even can be interpreted to contain cultural and communal meanings.

These various intentions of users on social tagging lead us to the question as to when index terms and social tags are assigned to the same publications, whether both kinds of metadata are alike or not. Both share the same purpose — retrieving information — but are generated by disparate groups of people. To answer these questions, we tested the degree of overlap between social tags and MeSH terms assigned to the same publications.

In order to comprehend the unique traits of social tags in contrast to MeSH terms, social tags are also compared with other kinds of metadata, which are highly related to MeSH; thus, family terms of original MeSH terms and consumer health vocabulary.

First, MeSH terms are rooted in a hierarchical structure, so as to represent the relationships among indexed information items. That means that each MeSH term can have its family terms: for instance, parents, children, or siblings. In order to assess the relative specificity, social tags are compared to all parents and children terms of original MeSH terms. Because sibling terms tend to have the same level of specificity, we omit the comparison of social tags with the siblings of original MeSH terms.

Second, even though many readers of biomedical articles are not necessarily general patients or laypeople, given the fact that they add social tags in more casual and uncontrolled ways without any predetermined vocabulary or indexing rules, we reckon that social taggers assign consumer health vocabulary (CHV hereafter) as social tags. In order to test this assumption, social tags were also compared with CHV. The CHV initiation was proposed to link the gab between daily words familiar to patients and highly technical terms or jargon, including MeSH terms used by medical professionals (Zeng-Treitler et al. 2008). For instance, instead of "myocardial infarction," "craniocerebral trauma," "pes," or "patella," the general patients or laypeople use "heart attack," "head injury," "foot," or "kneecap." Since it is likely that social taggers are end users of biomedical literature search, it is critical to understand how much of the social tags vocabulary contains CHV terms.

Most existing studies focusing on the lexical differences between index terms and social tags are executed regardless of the annotated publications or in a too-small scale to generalize the results. In addition, there is no study to consider more than one data source of social tags, as done in this paper. Reflecting upon these problems, one contribution of this paper is to investigate the differences of subject headings with social tags for the same articles in a relative large scale (n = 4300 articles). The data sources of social tags in this paper are from two typical bookmarking systems: BibSonomy and Delicious. Hence, we can alleviate the possible bias when using only

www.kci.go.kr

one data source of social tags. More importantly, this study is designed to explore the difference and similarity between social tags and MeSH in multiple perspectives in order to substantiate whether a highly controlled publication index reflects general readers' cognitive understandings and stresses the necessity of general readers' involvement in the publication indexing process.

2. Related Work

Related to this paper's topic (i.e. a comparative analysis between publication index and social tags), there are two trends of research: one is to explore how the overall characteristics or vocabularies of these metadata kinds are different in general, regardless of the annotated literature; another is to compare two kinds of metadata for the same literature. First, Heckner, Mühlbacher and Wolff (2008), Mathes (2004), Noll and Meinel (2007), and Yi and Chan (2009), Good, Tennis and Wilkinson (2009) are a few examples of the first trend. On the other hand, Bruce (2008), Kipp (2006; 2008; 2011), Lee and Schleyer (2010; 2012) and Lin et al. (2006) belong to the second research trend.

For instance, Mathes (2004) has identified three kinds of metadata for books, reflecting three different view points: (a) indexer's perspectives expressed by publication indices such as the Library of Congress Subject Headings (LCSH); (b) authors' perspectives provided by author-generated vocabularies; and (c) readers' perspectives as expressed through social tags. However, the author did not perform any empirical comparisons among these types of metadata.

Noll and Meinel (2007) compared authors' and readers' viewpoints by examining the annotations of 100,000 Web pages randomly selected from the Open Directory. The study inferred authors' viewpoints from two kinds of metadata: (a) the (X)HTML metadata described using the TITLE element, META keywords, and META description, and (b) content ratings provided by a standard rating system for Web content (e.g. nudity, sexual material, violence, language). Social tags of the 100,000 Web pages from Delicious provided the readers' viewpoints. Of the 100,000 URLs, 5.1% (4,992) had both author metadata and social tags. In total, 58% of the social tags overlapped with the authors' metadata. However, again, page-by-page comparisons were not made.

Yi and Chan (2009) investigated the overlap between a folksonomy (social tags on Delicious) and a publication index (terms on LCSH). Of 409 tags for 4500 Web pages on the Delicious system, approximately 60 percent were contained in the LCSH vocabulary (Yi and Chan 2009), demonstrating a significant degree of overlap. However, the study did not determine to what degree tags and LCSH terms overlapped for the same information items.

www.kci.go.kr

As the similar studies to the one presented in this paper, Lin, et al. (Lin et al. 2006) performed the empirical comparison of a publication index with social tags for the same literature. The authors compared social tags, MeSH terms, and titles for a set of biomedical papers contained in Connotea (http://www.connotea.org), a social reference management system. Forty-five biomedical papers were indexed in the PUBMED database. The literature was tagged with 540 tags by 264 Connotea users. Eleven percent of social tags used by at least one user (n = 59) overlapped with MeSH terms and 19% (n = 102) with titles. Twenty-four percent of social tags used by at least two users (n = 123) overlapped with MeSH terms and 42% (n = 52) with titles. Unfortunately, these results are difficult to be generalized because of the small sample size.

The underlying assumption of Yi's study (Yi 2010) is that frequently used social tags are valuable to replace Subject Heading terms. For 114 books, the author tested the assumption through the LCSH terms and social tags of LibraryThing (http://www.librarything. com). This study concluded that the five most popular (top-five) social tags are the most valuable. The study demonstrated that the social tags voted by the majority of users have the quality equivalent to the controlled vocabulary. However, including more than five social tags largely reduced the chance that social tags are matched to the publication index. Moreover, the author performed this comparison in a very small scale.

In a series of studies, Kipp (2006; 2008; 2011) compared social tags assigned by CiteULike users, keywords assigned by authors, and library descriptors assigned by professional indexers for 165 journal articles. The similarity among the three sets of tags was compared using a structured thesaurus with seven levels of matching (ranging from "same" to "not related"). The study showed that, while users often use terminology that is somewhat like that used in a thesaurus, they tend not to use the exact terminology of the thesaurus to describe their work. The study also highlighted clear differences in the use of tags, keywords, and descriptors with respect to aspects such as time and task management tags, geographic descriptors, specific details and qualifiers, generalities, and emergent vocabulary.

Bruce (2008) compared CiteULike tags and descriptors assigned by the Educational Resources Information Center for 2786 articles on education-based research. Among the 3176 unique tags and 2899 unique descriptors assigned to articles, the author found 240 (7.6%) exact matches. He concluded that users of CiteULike and indexers use different vocabularies.

We used our own previous study (Lee and Schleyer 2012) to empirically compare the two sets of metadata for the largest body of literature. In the study, which compared MeSH terms with CiteULike social tags for 231,388 papers, the results demonstrated that two kinds of metadata are quite distinct lexically; 3.3% overlap at maximum. However, one flaw of this study was we did not consider any

www.kci.go.kr

other folksonomy but the CiteULike's. That means that the comparison of MeSH terms with other folksonomies can yield different results. Therefore, the main motivations of this study are to compare two sorts of metadata for the same articles and to consider more than one source of social tags.

3. Data Sets and Processing

The analysis described in this paper is targeted in comparing two sets of metadata, when these sets were assigned to the same literature. In the data collection for this analysis, specifically, we collected social tag data first, and then, among the articles annotated with social tags, we chose all biomedical articles indexed with MeSH terms. Once the collection of both metadata types had been completed, for effective comparison, various text-processing techniques were applied. This section introduces the ways to collect data, to apply text-process techniques, and to match terms between both types of metadata.

3.1 Data Sources of Social Tags and MeSH Terms

In order to alleviate the possible bias of one folksonomy, as the data sources of social tags, this study chose two typical social bookmarking systems: BibSonomy and Delicious. First, BibSonomy[1] is one of the successful bookmarking systems for managing and sharing bibliographic references. This study used the dataset officially distributed by the administrators and; specifically, the data version of this study was made on February 2, 2014 (Benz et al. 2010). The dataset contains the list of users, users' bookmark histories, time-stamps of bookmarks, and social tags. It also has bibliographic details of bookmarked references;, titles, abstracts, author names, publication years, published journal/conference names, and the Internet URLs of references. The second data source is Delicious[2]; which is a popular social bookmarking system for Web resources. This study collected the data on Delicious using the snowball sampling. To pick the initial set of seed users, I visited Delicious for 15 times randomly on February 2012. All users who posted new bookmarks at the time of visit were chosen. For a group of seed users, the breadth-first search (Becchetti et al. 2006) was performed to collect their online social connections (i.e. so called 'network') on the system. This breadth-first search was executed multiple times for all

1) http://www.bibsonomy.org
2) http://www.delicious.com

www.kci.go.kr

collected users. After collecting users, I collected each user's bookmark information, including his bookmarked Internet URLs, time-stamps of bookmarks, and social tags.

Both BibSonomy and Delicious systems assign pseudo-tags automatically, such as "no-tag," when users do not assign any tags. BibSonomy assigns "*file-import-xx-xx-xx" for a citation imported from a citation file (e.g. endnote, RIS, BibTex, etc.). While pseudo-tags appeared in many bookmarks of multiple users, they were only used for the systems' sake and do not carry any significant meanings of the assigned papers. Prior to moving to the next data collection, all of the pseudo-tags were removed.

The next step was to choose biomedical articles in the given BibSonomy and Delicious data sets. We can derive the biomedical articles from the Internet URLs of bookmarks. Each biomedical article stored in the PUBMED database refers to the address of the National Library of Medicine (i.e. http://www.ncbi.nlm.nih.gov). Moreover, biomedical articles can be identified via the PUBMED IDs on the URLs (e.g. http://www.ncbi.nlm.nih.gov/pubmed/*PUBMED_ID*). Through this process, I found 2548 and 1758 articles in BibSonomy and Delicious, respectively. Between BibSonomy and Delicious datasets, very few articles (six articles) are common. Accordingly, the total number of articles considered in this study is 4300. For the 4300 biomedical articles in both datasets, MeSH terms were crawled from the Web-based MEDLINE/PUBMED database provided by the National Library of Medicine. For the descriptive statistics of social tag datasets and MeSH dataset, refer to Table 2 in the Results section.

3.2 Text Processing and Matching Social Tags and MeSH terms

As explained, compared with the concrete purpose of the MeSH index, which is to abstract documents and help the search of documents, users of social bookmarking systems utilize social tags with various intentions. Therefore, one of the major characteristics of social tags is the idiosyncrasy. Since social tags consist of uncontrolled free-text words, they also have problems such as imprecision, redundancy, and ambiguity (Suchanek, Vojnovic and Gunawardena 2008; Yi and Chan 2009). These characteristics of social tags can introduce noise. In order to alleviate these problems and make the comparison more effective, various text-processing techniques were applied to both social tags and MeSH terms, prior to the comparative analysis. The applied techniques are 1) case normalization; 2) deleting delimiters; 3) tokenization; and 4) stemming.

As the first stage, all social tags and MeSH terms were converted into lower case. The second technique was to replace delimiters (such as comma, slash, backslash, colon, semicolon and asterisk)

www.kci.go.kr

with spaces. Because social bookmarking systems usually do not allow users to enter spaces in social tags, as an alternative, users added various delimiters to make multiple words as a single social tag: for instance, "Neoplasms/Hormone/Dependent," "Models;Adult;Auditory," or "Molecular_Sequence_Data." After converting delimiters into spaces, social tags were tokenized into multiple subcomponents on the center of the spaces. Through this process, compound social tags such as "Neoplasms/Hormone/Dependent" and "Models;Adult;Auditory" can be dissected into one set of "neoplasms," "hormone," and "dependent" and another set of "models," "adult," and "auditory." The same methods – case normalization, deletion of delimiters and tokenization – were applied to MeSH terms. In case of MeSH terms, only commas were replaced with spaces, and, on the center of the spaces, each MeSH term was divided into multiple subcomponents. Through these techniques, MeSH terms and social tags consisting of multiple words can be compared directly. Especially, when all subcomponent words of a MeSH term should appear in social tags, the pair should be counted as a match. Table 1 shows some examples of the matching processes.

The last technique was stemming word variations. Specifically, Porter stemmer was applied (Liu 2007). Stemming is widely used in text retrieval and search engines to reduce word variations by truncating words to their stems or roots. For example, "computer," "computers," "computing" and "compute" are truncated to "comput." Thus, stemming could increase the number of matches, as shown in the Table 2 (i.e. article #3).

〈Table 1〉 An Example Showing How to Count Matches between MeSH terms and Social Tags (Matched terms were bold and underlined)

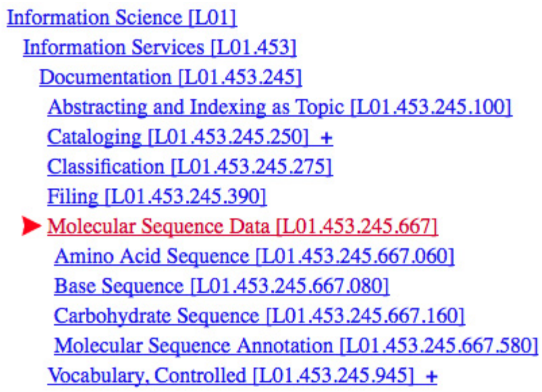
Article	MeSH	Social Tags	No. of Matches
#1	"molecular sequence data", "**<u>dna</u>**"	"data", "**<u>dna</u>**", "sequence"	1
#2	"**<u>breast cancer</u>**"	"**<u>breast</u>**", "**<u>cancer</u>**"	1
#3	"**<u>comput</u>**ing", "device"	"**<u>comput</u>**er"	1
#4	"**<u>anterior cruciate ligament</u>**", "**<u>pain</u>**"	"**<u>anterior</u>**", "**<u>cruciate</u>**", "**<u>ligament</u>**", "**<u>pain</u>**"	2

3.3 Data Collection and Text Processing of the Other Kinds of Metadata

Along with the MeSH terms, which are originally indexed to the publications, social tags were compared with the other two kinds of MeSH-related metadata: family terms of original MeSH and CHV. First, on the center of the original terms, the family terms were limited to the parents of two steps up and the children of two steps down in the given MeSH hierarchy. For instance,

www.kci.go.kr

Figure 1 depicts one part of the MeSH hierarchy. When “molecular sequence data” was the originally indexed MeSH term, “documentation” and “information services” will be considered as the parents (i.e. ancestors), along with “amino acid sequence,” “base sequence,” “carbohydrate sequence,” and “molecular sequence annotation” as the children (i.e. descendants).


Information Science [L01]
Information Services [L01.453]
Documentation [L01.453.245]
Abstracting and Indexing as Topic [L01.453.245.100]
Cataloging [L01.453.245.250] +
Classification [L01.453.245.275]
Filing [L01.453.245.390]
Molecular Sequence Data [L01.453.245.667]
Amino Acid Sequence [L01.453.245.667.060]
Base Sequence [L01.453.245.667.080]
Carbohydrate Sequence [L01.453.245.667.160]
Molecular Sequence Annotation [L01.453.245.667.580]
Vocabulary, Controlled [L01.453.245.945] +

<Figure 1> An Example of MeSH Hierarchy

Second, Zeng-Treitler, et al. developed a corpus of medical lay terms (i.e. CHV) that are familiar to non-experts. Based on search logs of online service provided by the US National Library of Medicine and the corpus of UMLS (the Unified Medical Language System) technical terms and concepts, the authors quantify lay health terms by computing several graph-based algorithms (term co-occurrence network and contextual network) and data mining techniques (familiarity value) (Zeng-Treitler et al. 2008). Especially, the CHV corpus was founded upon UMLS, and MeSH is a part of UMLS meta-thesaurus. The process to create the CHV provides mapping of CHV terms to MeSH terms, thereof. For this comparison, CHV terms mapped to each original MeSH term were chosen.

For these two kinds of metadata (i.e. the family terms of the original MeSH terms and consumer health vocabulary), the same set of text-processing techniques (case normalization, deleting delimiters, tokenization and stemming) was applied.

3.4 Measure of Similarity

To compute similarity between MeSH and social tags corpora, this study uses similarity measures, thus making term-to-term comparisons possible. The first measure is the number of common

www.kci.go.kr

terms between MeSH and social tags for the same articles. Because the corpora of social tags and MeSH terms in this study have different sizes, as with the other similarity measures, two types of coverage ratio are computed as eq.1 and eq.2 depict.

$$Coverage\ Ratio\ of\ Tags = (A \cap B)/A \qquad \text{eq. 1}$$

$$Coverage\ Ratio\ of\ MeSH\ terms = (A \cap B)/B \qquad \text{eq. 2}$$

Where A is the social tag data set, comprised of all distinct tags for a paper, and B is the MeSH term data set, comprised of distinct MeSH terms assigned to the same article. The coverage ratios define the fraction of the common annotations for an article covered by its social tags (eq. 1) and MeSH terms (eq. 2), respectively. The coverage ratio is useful to determine how much of the corresponding metadata is identical and, furthermore, replaceable to another metadata. For instance, if an article has 20 MeSH terms, 100 social tags, and five common terms, the coverage ratios of tags and MeSH are 0.05 (5/100) and 0.25 (5/20), respectively. In this example, social tags cover a relatively large portion of MeSH terms (i.e. one-fourth of the terms), whilst, while little percent of social tags (< 10%) is covered by MeSH. Hence, social tags could be a good substitute for MeSH terms. Otherwise, in the other way around, when the coverage of MeSH is little, it substantiates that MeSH terms do not reflect users' cognitive understandings about the article. To summarize, we measured the similarity between MeSH and social tags for the same articles via 1) the number of common terms; 2) coverage ratio of tags; and 3) coverage ratio of MeSH.

When comparing social tags with the family terms of the original MeSH terms and CHV terms, the number of common terms was also used. In addition, only coverage ratio of tags was computed, so as to calculate how many social tags were overlapped with these other types of metadata.

4. Results

4.1 Description of Social Tags and MeSH Terms Datasets

Table 2 represents a summary of BibSonomy and Delicious social tag-related datasets and MeSH data set used in this paper. The description shows that the average bookmarks possessed by BibSonomy users (i.e. 19.59) are about six times higher than those of Delicious users (i.e. 3.18). The average number of social tags annotated on each BibSonomy article (i.e. 6.59) is

www.kci.go.kr

significantly higher than Delicious (i.e. 2.70), as well. On BibSonomy, 30.70% of articles have five tags or more and 12.03% have in excess of 20 tags. However, 15.63% and 0.4% of Delicious articles have at least five or 20 tags, respectively. We can reckon that it is because BibSonomy is a specialized system to manage and share bibliographic information, whereas Delicious is a more general-purpose system to manage and share Web bookmarks. For users' research activities, the collections of bibliographic information may be more useful than Web bookmark collections. Hence, it seems that BibSonomy users put more effort (collecting more bookmarking and adding more social tags) into collecting and organizing their collections than do Delicious users. We can postulate that the difference in social tag sizes between two systems could reflect the results of our comparisons, and this speculation will be checked in the later part of this section.

Regardless of whether it is from BibSonomy or Delicious, for the 4300 target articles of this study, the average number of social tags per article is 5.18 generated by 1.0 user. The average number of MeSH terms per article is 11.67.

〈Table 2〉 Descriptive Statistics of Social Tags and MeSH datasets

Social Tags	Bibsonomy	Total No. of Articles	2548
		Average Social Tags Per Article	6.59 (σ = 9.5)
		Total No. of Distinct Social Tags	4853
		Total No. of Annotators	140
		Average Bookmarks per User	19.59 (σ = 73.5)
	Delicious	Total No. of Articles	1758
		Average Social Tags Per Article	2.70 (σ = 2.77)
		Total No. of Distinct Social Tags	1872
		Total No. of Users	585
		Average Bookmarks per User	3.18 (σ = 7.4)
MeSH Terms		Total No. of Articles	4300
		Average MeSH Terms Per Article	11.67 (σ = 4.77)
		Total No. of Distinct MeSH Terms	7159

Another interesting point of these data sets is the number of annotators per each article. In the comparison between social tags and MeSH terms, we can investigate social tags in two levels: personomy and folksonomy. The personomy level comparison is based on the tri-partite relation of social tagging (user-article-tag). For each paper, the personomy level test compares an individual set of social tags annotated by *each user* to the assigned MeSH terms. In particular, in the comparison of compound MeSH terms, all subcomponents of the compound terms should appear in the set

www.kci.go.kr

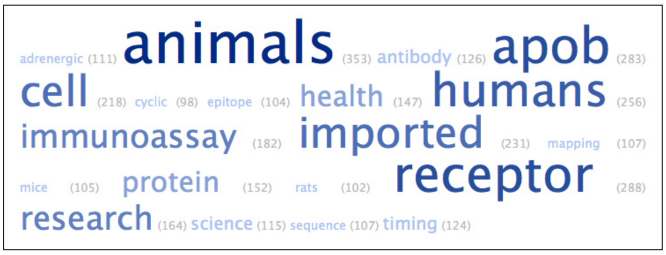
of *one user's social tags*. On the other hand, the folksonomy level comparison is based on the bi-partite relation of social tagging (article-tag), regardless of the annotators. Hence, for each paper, an aggregated set of social tags, which are added by all of the users who bookmarked the corresponding paper, was compared with the assigned MeSH terms. The folksonomy shows much more general viewpoints for a corresponding paper than personomy made by one person. In this folksonomy-level comparison, a match indicates that multiple subcomponents of a compound MeSH term are in the aggregated social tag set. As the example, for article #4 in Table 1, let's assume that user *A* added "anterior" and "ligament" and user *B* added "cruciate" and "pain" as their social tags. On the personomy level, there is only one match: "pain." On the folksonomy level, by aggregating user *A* and *B*'s social tags together, there are two matches: "pain" and "anterior cruciate ligament." However, in both BibSonomy and Delicious data sets, the average number of annotators per article was too few to consider folksonomy. On average, the average number of users who annotated social tags on each paper is 1.00 on BibSonomy ($\sigma = 0.0$) and 1.02 on Delicious ($\sigma = 0.30$), respectively. That is to say, actual results of personomy and folksonomy level comparisons were not much different. Hence, this study executed the folksonomy level comparison.

The next step is to examine the most popular social tags and MeSH terms. Figure 2 shows the top-20 social tags plus the frequencies (the numbers in parenthesis) in either a BibSonomy or Delicious dataset. Figure 3 shows the top-20 MeSH terms and the frequencies. All of the top social tags are single words, while some of the top MeSH terms are compound words. There are some overlaps between two sets. Out of the top-five MeSH terms (i.e. "Humans," "Male," "Aminals," "Female" and "Adult"), two terms also belong to the top social tags. The morphologic difference between two sets such as the MeSH term "antibodies" and the social tag "antibody" indicates the necessity of text-processing techniques prior to the comparative analysis. Compound-based word formation also requires matching based on the subcomponents. The social tags "sequence" and "timing" partially comprise of the MeSH terms "amino acid sequence," "molecular sequence data," and "time factors." Another interesting pattern is that one top MeSH term (i.e. Apolipoprotein B) is abbreviated to one of the top social tags (i.e. apob).

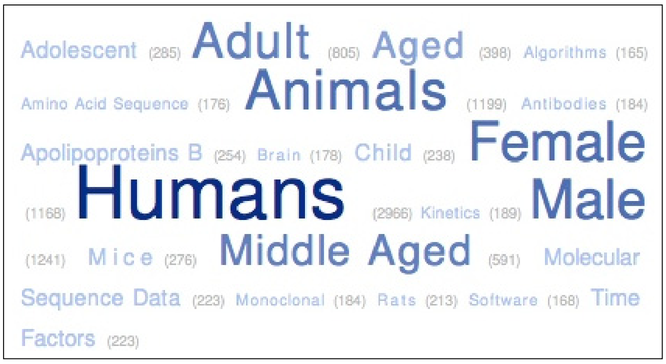
Moreover, Figure 3 shows that general terms such as "humans," "male," "animals," and "female" are used frequently as MeSH terms. These are used too often to make general distinctions among papers. For instance, the MeSH terms "humans" and "animals" were assigned to about 69% and 28% of the papers, respectively, while the social tag "humans" and "animals" were used for about 6% and 8.2% of the papers. However, the descriptive statistics Table 2 shows that MeSH

www.kci.go.kr

term corpus consists of more diverse terms (the number of distinct MeSH terms is 7159) than social tags (the distinct social tags are 6357 in both social tag datasets). Despite the diversity of the vocabulary, the frequency distribution of MeSH terms is skewed to a few popular words.



〈Figure 2〉 Top 20 Most Popular Social Tags



〈Figure 3〉 Top 20 Most Popular MeSH Terms

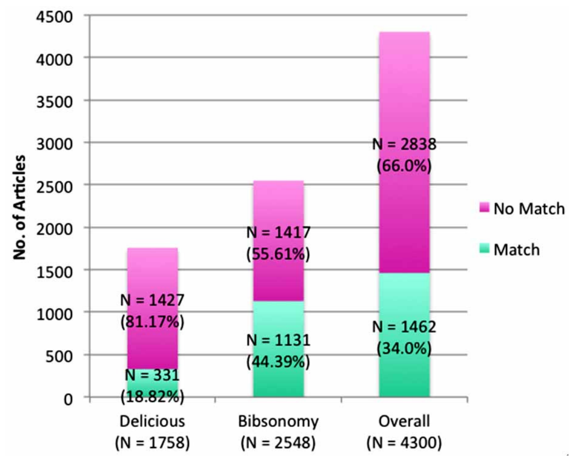
4.2 Comparison Between Social Tags and MeSH Terms

This section demonstrates the results of comparison between social tags and MeSH terms. Of the 4300 articles, 34% (n = 1462) shared at least one identical social tag/MeSH term. That is, two thirds of literature considered in this study did not share any common term between social tags and MeSH terms. On average, the 34% of our articles shared 1.66 common tags/MeSH terms per paper (σ = 4.25). 16.25% of social tags were matched with MeSH terms assigned to the same articles, and 12.80% of the MeSH terms were covered by social tags. Put differently, even though the two types of metadata were assigned to the same publications, almost 90% of MeSH terms were not overlapped with social tags.

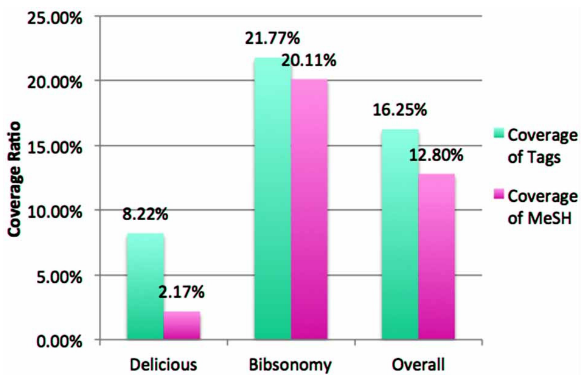
Specifically, 44.39% of BibSonomy articles (n = 1131) and 18.82% of Delicious articles

www.kci.go.kr

(n = 331) shared one or more similar terms with MeSH, respectively. The average number of matched terms between BibSonomy social tags and MeSH terms is 2.64 (σ = 5.28); 21.77% of BibSonomy social tags were identical to the MeSH terms, and 20.11% of MeSH terms were matched with social tags of the same publications. On the other hand, the average number of overlapped terms between Delicious social tags and MeSH terms is 0.23 (σ = 0.54); 8.22% of Delicious social tags were the same as MeSH terms of the same articles, when 2.17% of MeSH terms were overlapped with tags. Figure 4 shows the summary of overlaps.



(a) Number of Articles Matched between Social Tags and MeSH terms



(b) Comparison of Two Coverage Ratios

<Figure 4> Patterns of Overlaps between Social Tags and MeSH terms

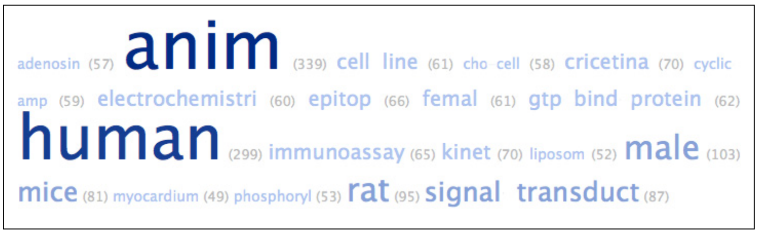
The Figure 4 summarizes that, depending on the dataset, social tags were matched with MeSH terms in quite different degrees. For instance, almost half of the BibSonomy articles have one shared term between social tags and MeSH terms, whereas less than one fifth of Delicious articles share

www.kci.go.kr

one common term. While 20% of MeSH terms were the same with BibSonomy tags, only 2% of MeSH terms were covered by Delicious tags. As we speculated in the first subsection of the Results, this difference might be caused by the tag sizes of two datasets. The average number of BibSonomy social tags per article (n = 6.59) is much higher than Delicious (n = 2.70). In order to prove this speculation is correct, this study computed how much the number of social tags per article is correlated to the number of overlaps with MeSH terms and to the coverage ratio of MeSH.

The number of social tags per each paper was significantly and strongly correlated to how many social tags are matched with MeSH terms (r = 0.88; p < .01). The size of social tags also has a significant correlation with the coverage ratio of MeSH (r = 0.83; p < 0.1). The results of both correlations concluded that the more social tags are annotated to a paper, the more likely that the tags are the same to the MeSH terms of the same paper.

Lastly, the Figure 5 depicts the top-20 most matched social tags/MeSH terms with the frequency of matches. The terms in the figure are in the fully text-processed format. Social taggers and publication professionals still often use popular terms such as “anim (i.e. animal(s)),” “human,” “male,” “femal (a.k.a female(s)),” and “mice” with one accord. Unlike the top-20 social tags displayed in Figure 2, however, this top-20 list includes compound words such as “signal transduct,” “gtp bind protein,” “cell line,” and “cho cell”. Moreover, although they are not used as often as the other top words, the list of words (e.g. “phosphoryl (i.e. Phosphorylation),” “myocardium,” “liposom (i.e. Liposome(s)),” “epitop (i.e. Epitope(s)),” “electrochemistri (i.e. Electrochemistry),” “adenosin,” “cricetina (i.e. Cricetinae)” are repeatedly indexed to the same publications by social taggers and publication professionals.



<Figure 5> Top 20 Most Matched Social Tags/MeSH terms

4.3 Comparison Between Social Tags and Family Terms of Original MeSH Terms

Since MeSH corpus is based on the hierarchical structure, each MeSH term can have the relative terms. For instance, parent terms can be considered to convey broader and extended references

www.kci.go.kr

to the original MeSH term. Sibling terms will associate additional contexts, and children terms can bring more precise sense to original terms. With the purpose to assess the relative specificity, social tags were compared with parent and children terms of MeSH headings that are originally indexed to the 4300 articles. Again, we limited the scope of family terms to two steps up and down. Since this comparison focused on the specificity, sibling terms in the same level were not considered.

Out of 4300 articles, 24.67% (n = 1061) have at least one shared term between social tags and parent terms of original MeSH terms. The average number of common tags/parent terms per paper is 0.61 (σ = 1.64), and 7% of social tags were matched with parent terms. Specifically, 871 BibSonomy articles and 190 Delicious articles shared at least one common item. On average, 0.95 BibSonomy social tags were identical to parent terms for every article. That means 9.73% of BibSonomy social tags were covered by parent terms. On the other hand, the average number of common terms between Delicious social tags and parent terms is 0.12, which indicates 3.38% coverage ratio of social tags.

Compared with the results of the parent terms, a smaller portion of articles (12% of overall articles) has one or more common terms between social tags and children terms of the original MeSH. The average number of social tags matched with children terms is 0.23 (σ = 0.85), which is also fewer than parent terms. The portion of children MeSH terms to cover social tags was 2.41%. The separate results between BibSonomy and Delicious social tags also indicate smaller shares with children terms. 428 BibSonomy articles and 90 Delicious articles shared common words between social tags and children terms, respectively; 0.36 BibSonomy social tags were the same with children terms, which were 2.41% of social tags; 0.05 Delicious social tags shared the common words with child terms; and it was 2.40% of all Delicious social tags.

In order to test the statistical difference between the social tag matches with parent terms and matches with children terms, I ran the t-test. Since the average number of children terms per article (n = 94.48, σ = 123.11) is almost five times larger than the parent terms (n = 19.49, σ = 5.94), we can naturally presume that social tags would have more shares with children terms. However, the statistical test shows the opposite result. Social tags are significantly more overlapped with parent terms than children terms (t = 42.14; p < .01). Conclusively, for more than 20% of our target articles (i.e. 24.67%), users of social bookmarking systems added broader and a less specific sense of indexing words than MeSH terms, so as to abstract and categorize their items of interests.

www.kci.go.kr

4.4 Comparison Between Social Tags and the Consumer Health Vocabulary

This last analysis evaluates how often social taggers use health terminologies familiar to laypeople, instead of highly technical and expertise MeSH terms. The original MeSH terms of all publications in our consideration have been substituted by the corresponding CHV words, as provided by the Consumer Health Vocabulary Initiatives (Keselman et al. 2008).[3]

22% of our publications (n = 947) have the same terms between social tags and CHV. There are 0.51 common terms per article (σ = 1.46), and 6.29% of social tags are the same with CHV words. I found at least one match for 765 BibSonomy articles and 182 Delicious articles; 0.77 BibSonomy social tags (8.07% coverage of social tags) and 0.12 Delicious social tags (3.67% coverage of social tags) are matched with CHV, respectively.

In a quite similar degree of overlap between social tags and parent terms of original MeSH terms, for about 20% of our publications, social taggers use consumer health terms instead of technical MeSH terms.

〈Table 3〉 Summary of Comparative Analyses
(Bib. and Del. stand for Bibsonomy and Delicious dataset respectively.)

		No. of Articles Shared at least One Term (Ratio)	No. of Common Terms	Coverage Ratio of Social Tags	Coverage Ratio of MeSH Terms
Original MeSH vs. Social Tags		1462 (34%) (Bib. = 1131, 44.4%; Del. = 331, 18.8%)	1.66 (Bib. = 2.64; Del. = 0.23)	16.3% (Bib. = 21.8%; Del. = 8.2%)	12.80% (Bib. = 20.1%; Del. = 2.2%)
Family MeSH vs. Social Tags	Parent MeSH	1061 (24.7%) (Bib. = 871, 20.3%; Del. = 190, 4.4%)	0.61 (Bib. = 0.95; Del. = 0.12)	7% (Bib. = 9.7%; Del. = 3.4%)	N/A
	Children MeSH	516 (12%) (Bib. = 428, 10.0%; Del. = 90, 2.1%)	0.23 (Bib. = 0.36; Del. = 0.05)	2.4% (Bib. = 2.4%; Del. = 2.4%)	N/A
Consumer Health Vocabulary vs. Social Tags		947 (22%) (Bib. = 765, 17.8%; Del. = 182, 4.2%)	0.51 (Bib. = 0.77; Del. = 0.12)	6.3% (Bib. = 8.1%; Del. = 3.7%)	N/A

3) http://consumerhealthvocab.org (accessed in February, 2014)

www.kci.go.kr

5. Conclusion and Discussion

This article investigates the question of whether social tags and controlled medical subject headings are similar, when both types of metadata are assigned to the same articles. Our study compares 22,924 social tags (6,357 distinct tags) to 50,202 MeSH terms (7,159 distinct terms), for a set of 4300 biomedical papers on an article-by-article basis. Especially with the purpose to minimize the possible bias of one data set, this study examines two social tagging datasets. In order to comprehend how users assigned their social tags in detail, the tags are also compared with two other types of MeSH related metadata: family MeSH terms and consumer health vocabulary.

As Table 3 succinctly depicts, only a third of the target citations share one or more terms between social tags and MeSH terms. Individual papers have 1.66 tags/MeSH terms in common, on average. In particular, despite the purpose of subject headings to describe users' interests on indexed articles, only 12.80% of the MeSH was matched with social tags, which represent users' cognitive understanding or interest in the corresponding articles. The comparative analysis of social tags with the other kinds of metadata also produced interesting findings. For about 25% of the target articles (1061 articles), the general users added terms carrying broader and less specific senses than the original MeSH terms. Social taggers assigned more specific terms than the original MeSH terms only for about 12% of the target articles. In addition, for one fifth of the target articles (22% or 947 articles), users added consumer health words as social tags, which are familiar to laypeople and general patients.

However, this small matching ratio between CHV and social tags is contradictory to our expectations, since we considered that both types of metadata are from general readers or consumers of biomedical articles. However, at this point, we need to consider the fact that the starting point of CHV is users' query history on UMLS (the Unified Medical Language System), which is developed by PUBMED and the umbrella system of MeSH. That means the creators of CHV only focused on users of the UMLS system and did not consider any vocabulary of other users who haven't ever used the UMLS system. Moreover, when there were no matching UMLS/MeSH terms, the creators of CHV did not consider the consumers' query terms. As of the time when this study was conducted, the author of this paper failed to find any other consumer/general patient-centered health vocabulary. Hence, the future direction of this study is to find another possible patient-centered health vocabulary and compare the quality with social tags.

The three steps of analysis (comparisons of social tags with original MeSH terms, family terms of the original MeSH, and CHV) found matches between social tags and MeSH-related metadata

www.kci.go.kr

for almost half of our target papers (49.77%; n = 2140). That is, even after applying various text-processing techniques and including various MeSH-related terms, another half of the target papers do not share any common term with social tags. Moreover, 55.1% of distinct original MeSH (n = 3942) is still not matched with any social tag. Importantly, these results demonstrate that social tags and MeSH terms are quite discrete lexically, reflecting different viewpoints between the general users and publication professionals. Social tags and MeSH terms are assigned through different processes and differ in many perspectives such as specificity and familiarity. Hence, this article contributes rich evidence for the difference of social tags to MeSH terms and the necessity to let general users participate in indexing processes.

For the future direction of this study, given the challenges in matching tags and index terms computationally, a small-scale study that uses human reviewers to assess matches may be a promising option. Another planned study is to conduct additional, sophisticated statistical tests about term distributions such as topic modeling. Lastly, an empirical study comparing user search performance using either social tags or MeSH terms may be a meaningful way to explore how MeSH terms and tags contribute in the retrieval of the biomedical literature.

Acknowledgments

The author of this paper deeply appreciates Bibsonomy administrators for generously sharing their whole data set.

References

[1] Becchetti, L. et al. 2006. A comparison of sampling techniques for web graph characterization.

[2] Benz, D. et al. 2010. "The Social Bookmark and Publication Management System BibSonomy." *The VLDB Journal*, 19(6): 849-875.

[3] Brown, P. et al. 1996. "The democratic indexing of images." *New Review of Hypermedia and Multimedia*, 2(1): 107-120.

[4] Bruce, R. 2008. "Descriptor and Folksonomy Concurrence in Education Related Scholarly Research." *Webology*, 5(3).

[5] Burnett, K., Ng., K. B. and Park, S. 1999. "A comparison of the two traditions of metadata

www.kci.go.kr

development." *J. Am. Soc. Inf. Sci*, 50(13): 1209-1217.
[6] Chevalier, M. et al. 2011. "Information Retrieval and Folksonomies together for Recommender Systems." *E-Commerce and Web Technologies*, 85: 172-183.
[7] Cleveland, D. B. and Cleveland, A. D. 2001. *Introduction to Indexing and Abstracting*. 3rd ed. USA: Libraries Unlimited, Inc.
[8] Ferrara, F. and Tasso, C. 2011. *Improving collaborative filtering in social tagging systems*. in Proceedings of the 14th international conference on Advances in artificial intelligence: spanish association for artificial intelligence. Spain: Springer-Verlag.
[9] Golder, S. and Huberman, B. 2006. "Usage patterns of collaborative tagging systems." *Journal of Information Science*, 32(2): 198-208.
[10] Good, B., Tennis, J. and Wilkinson, M. 2009. "Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics." *BMC Bioinformatics*, 10(1): 1-17.
[11] Heckner, M., Mühlbacher, S. and Wolff, C. 2008. "Tagging tagging. Analysing user keywords in scientific bibliography management systems." *Journal of Digital Information*, 9(2).
[12] Hidderley, R. and Rafferty, P. 1997. *Democratic Indexing: An approach to the retrieval of Fiction*. PAYS-BAS: IOS Press.
[13] Keselman, A. et al. 2008. "Developing Informatics Tools and Strategies for Consumer-centered Health Communication." *Journal of the American Medical Informatics Association*, 15(4): 473-483.
[14] Kim, H. L., Decker, S. and Breslin, J. 2010. "Representing and sharing folksonomies with semantics." *Journal of Information Science*, 36(1): 57-72.
[15] Kipp, M. 2006. "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords." *Canadian Journal of Information and Library Science*, 30(3).
[16] Kipp, M. 2008. "Tagging for Health Information Organization and Retrieval." *TCDL Bulletin*, 4(1).
[17] Kipp, M. 2011. "User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike." *Canadian Journal of Information and Library Science*, 35(1): 17-48.
[18] Knecht, L. W. S. and Nelson, S. J. 2002. "Mapping in PubMed." *Journal of the Medical Library Association*, 90(4): 475.
[19] Lancaster, E. W. 2003. *Indexing and Abstracting in Theory and Practice*. Champaign, Illinois: Intellitext Corporation.
[20] Lee, D. H. and Schleyer, T. 2010. *A Comparison of MeSH terms and CiteULike Social Tags*

www.kci.go.kr

as Metadata for the Same Items. in the 1st ACM International Conference on Health Informatics (IHI 2010). USA: Washington DC.

[21] Lee, D. H. and Schleyer, T. 2012. "Social tagging is no substitute for controlled indexing: A comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers." *Journal of the American Society for Information Science and Technology*, 63(9): 1747-1757.

[22] Lin, X. et al. 2006. *Exploring Characteristics of Social Classification*. in 17th ASIS&T SIG/CR Classification Research Workshop.

[23] Liu, B. 2007. *Informational Retrieval and Web Search, in Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Edited by B. Liu. New York: Springer.

[24] Marlow, C. et al. 2006. *HT06, tagging paper, taxonomy, Flickr, academic article, to read, in Proceedings of the seventeenth conference on Hypertext and hypermedia*. Denmark: ACM.

[25] Mathes, A. 2004. *Folksonomies - cooperative classification and communication through shared metadata*. Computer Mediated Communication.

[26] Mika, P. 2005. "Flink: Semantic Web technology for the extraction and analysis of social networks." *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3): 211-223.

[27] Noll, M. G. and Meinel, C. 2007. *Authors vs. readers: a comparative study of document metadata and content in the www, in Proceedings of the 2007 ACM symposium on Document engineering*. Canada: ACM.

[28] Suchanek, F. M., Vojnovic, M. and Gunawardena, D. 2008. *Social tags: meaning and suggestions*. in Proceeding of the 17th ACM conference on Information and knowledge management. USA: ACM.

[29] Yi, K. 2010. "A semantic similarity approach to predicting Library of Congress subject headings for social tags." *Journal of the American Society for Information Science and Technology*, 61(8): 1658-1672.

[30] Yi, K. and Chan, L. 2009. "Linking folksonomy to Library of Congress subject headings: an exploratory study." *Journal of Documentation*, 65(6): 872-900.

[31] Zeng-Treitler, Q. et al. 2008. "Estimating Consumer Familiarity with Health Terminology: A Context-based Approach." *Journal of the American Medical Informatics Association*, 15(3): 349-356.

www.kci.go.kr

www.kci.go.kr