Determination of Semantic Types of Tags in Social Tagging Systems†

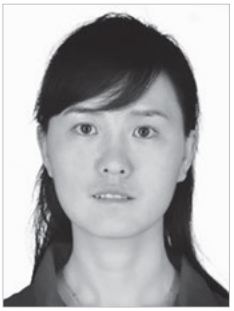
Yongfang Wang*, Yangfang Tai**, Yongfang Yang***

*Business College of Shanxi University, Taiyu Road, Taiyuan, Shanxi Province, China 030031, <wyf-6821006@163.com>

**Management School of Shanxi Medical University, No. 56, Xin-Jian S. Road, Taiyuan, Shanxi Province, China 030001, <yangfangtai@163.com>

*** Business College of Shanxi University, School of Foreign Languages, Taiyu Road, Taiyuan, Shanxi Province, China 030031, <1834930997@qq.com>

Yongfang Wang is a lecturer at the Business College of Shanxi University, China. She has published research papers in Chinese journals such as *Information Science*, *Library and Information Service*, *New Technology of Library and Information Service*, and *Journal of Intelligence*. Her research interests include information organization and information retrieval.



Yangfang Tai is a lecturer at Shanxi Medical University, Taiyuan, China. She has published academic papers in more than forty Chinese and international academic journals, such as *Journal of the China Society for Scientific and Technical Information*, *Information Science*, *Library*, *Journal of Modern Information*, and *International Journal of Education and Management*. Her research interests include knowledge organization and knowledge discovery.

Yongfang Yang is a PhD student at Shanxi University; her current work is focused on conversation analysis and talk-in-interaction. She is an associated professor at the Business College of Shanxi University. She has a master's degree in English literature and language and has written several articles in this area.

Wang, Yongfang, Yanfang Tai and Yongfang Yang. 2018. "Determination of Semantic Types of Tags in Social Tagging Systems." *Knowledge Organization* 45(8): 653-666. 38 references. DOI:10.5771/0943-7444-2018-8-653.

Abstract: The purpose of this paper is to determine semantic types for tags in social tagging systems. In social tagging systems, the determination of the semantic type of tags plays an important role in tag classification, increasing the semantic information of tags and establishing mapping relations between tagged resources and a normed ontology. The research reported in this paper constructs the semantic type library that is needed based on the Unified Medical Language System (UMLS) and FrameNet and determines the semantic type of selected tags that have been pretreated via direct matching using the Semantic Navigator tool, the Semantic Type Word Sense Disambiguation (ST WSD) tools in UMLS, and artificial matching. And finally, we verify the feasibility of the determination of semantic type for tags by empirical analysis.

Received: 31 August 2017; Revised: 24 May 2018; Accepted 23 June 2018

Keywords: semantic relations, social tagging, tags

† This paper is one of the results of the National Social Science Foundation of China's project, "Retrieve on Semantic Analysis of Tag System based on Frame Network Ontology" (project number: 13CTQ030).

1.0 Introduction

Social tagging is also called "cooperative tagging," namely, network users can define a set of tags spontaneously to describe a certain digital object. With the rapid development of information technology, social tagging is becoming more and more popular, and various social tagging platforms have emerged in succession, such as resource sharing websites (e.g., Delicious), video sharing websites (e.g., YouTube), photo sharing websites (e.g., Flickr), network radio platforms (e.g., Last.fm), and blogging and micro-blogging platforms (Xiong and Jiang 2017). In these social tagging systems, users can add tags to resources randomly, and these tags, with large numbers and lacking structure, are not limited by standardized vocabularies. A tag classification can make the decentralized tags highly structured, is conducive to mining the deep

semantic relations between a class's tags, and can mine the deeper semantic relationship among a certain category of tags and the relationships between resources that were tagged as well as the potential relationships between taggers. These classification methods are realized based on statistics and clustering algorithms (Li 2016).

The semantic types are used to describe the inherent and context-free lexical features and semantic features. The semantic type works as a good classifier that can classify different tags in a social tagging system, and each class's tag group shares a specific semantic type. There are specific hierarchical structures and semantic relations between specific semantic types. So, tags have structures and semantic relationships by mapping semantic types (Jia and Tai 2007). In addition, determining the semantic types of tags can enrich semantic information of tags.

Based on this we combine the semantic types of UMLS and the semantic types of the top-level ontology FrameNet to construct the semantic type library. We classify the semantic types of tags collected from BibSonomy by means of the SPECIALIST Lexicon Natural Language Processing (NLP) tools in UMLS and artificial determination. This paper is structured as follows: first, we summarize and analyze the relevant research results; then, we introduce the method of constructing a semantic type library; after that, we review the thinking behind classifying semantic types of tags and verify the method by empirical analysis; and, finally, we provide a conclusion.

2.0 Related work

With the growing popularity of social tagging, research on social tagging has also become popular. Scholars have conducted research on the following aspects: motivations for tagging, types of tags, modes of tagging, recommendations of tags, semantics of tags, visualization of tags, application and related problems of tags (Gupta 2010), power law of tag distribution, tagging communities, strategies of tagging (Munk and Mørk 2007a; 2007b), comparison among tagging practice, authors' keywords and descriptors from professional indexers in journal articles (Kipp 2007), etc.

2.1 Tag classification

The research on tag classification in social tagging systems is mainly based on three aspects: morphology, clustering alogrithms and semantics. Using methods based on morphology, scholars such as Al-Khalifa and Davis (2006) use the root reduction method to standardize the tags and categorize the same tags with the same roots, and Specia and Motta (2007) use string distance measuring to classify the tags with the same morphology into the same category. In work based on clustering algorithms, scholars such as Li et al. (2014) use a co-occurrence spectrum clustering method to classify tags in social tagging systems, and Radelaar et al. (2011) use the spectral bisection method to classify tags. In methods based on semantics, scholars such as Cui et al. (2011) use semantic distance measurement to classify tags, and Wu and Zhou (2012) take a variety of measures of network semantic aggregation to classify tags from semantic information. Also, scholars such as García-Plaza et al. (2012) classify tags manually according to the Open Directory Project (ODP). Relatively speaking, semantic information is better for tag classification, but no scholars have applied semantic types to tag classification at present.

2.2 Mapping of tags and ontology

The ontology mapping system is generally composed of the element layer and the structure layer mapping systems, and the result of the element layer mapping will affect the mapping of the structure layer (Xiong et al. 2013). The element layer mapping mainly calculates the similarity between concepts, and forms a concept-to-concept mapping process. The existing mapping method, in the process of element layer processing, mostly combines the characteristics of English grammar by using prefixes, removing plural forms, cutting suffixes, and other methods (Ghali 2011). Structural layer mapping not only considers a single tag element, but also considers the relationship between the element and other elements, mapping through the relationship between elements in a large structure. Scholars such as Han et al. (2010), based on a probability algorithm, determine the semantics between tags by mapping the tags using a domain ontology and use the co-occurrence tag environment to define the meaning of tags so as to understand users' diverse interests at the semantic level. García Silva et al. (2012) use data structure and algorithms to extract domain terms from public classifications and enrich the semantic information of tags by linking open data clouds. The domain ontology is obtained by mapping tags to an existing formal knowledge ontology. Because the relationship of concept elements contains a lot of latent semantics and has great influence on similarity, mapping based on structure layer has a better effect.

2.3 Tag semantic enrichment

Scholars use a variety of methods to enrich the semantic information of tags. Scholars such as Lux and Dosinger (2007) extract an ontology from the tags of social tagging systems. Djuana et al. (2011) propose linking social tagging with the online dictionary WordNet. Lee and Sohn (2013) propose drawing tag knowledge maps. Kiu and Tsui (2011) propose that social tagging systems can be combined with

controlled vocabularies. Wei Lai (2010) proposes integrating social tagging systems with network semantic resources (such as Swoogle). Yoo et al. (2013) develop a knowledge organization system CTKOS (CT-based Knowledge Organization System) based on classified tags to solve the problem of unclear semantics. Good et al. (2007) develop a semantic annotation system ED (Entity Describer) and use the ED knowledge base within Connotea. This system can enrich the semantic information of tags by creating a mechanism for social taggers to intentionally form connections between their tags and concepts from controlled, structured terminologies. Generally speaking, at present, the semantics of tags in social tagging systems are enriched mainly with existing controlled vocabularies or dictionaries based on algorithms and lexical mappings. If the semantic types and the rich relationships between semantic types are applied to the semantic richness of tags, the effect is better.

2.4 Evaluation of semantic similarity measures

In social tagging systems, a certain degree of similarity exists between tags, resources, and users. Scholars also put forward and evaluate a variety of similarity measures. Cattuto et al. (2008) analyze the characteristics of the three measures of tag relatedness: tag co-occurrence, cosine similarity of co-occurrence distributions, and FolkRank, by mapping the tags between Delicious and the synonyms of WordNet and point out the application situation of each measure. Markines et al. (2009) build an evaluation framework to compare various general folksonomy-based similarity measures derived from established information-theoretic, statistical, and practical measures, including matching, overlap, Jaccard, dice, cosine and mutual information. The evaluation framework first summarizes various aggregation methods, including projection, distribution, macro-aggregation and collaboration. Then, based on WordNet and ODP, they measure similarity between tags and resources in BibSonomy and analyze and compare the advantages and disadvantages of various measures and their feasibility. Lee and Schleyer (2012) compare *Medical Subject Headings* (*MeSH*) and CiteULike tags assigned to 231,388 papers. They measure the Jaccard similarity between *MeSH* and CiteULike tags.

2.5 Determination and application of semantic types

In terms of semantic type research, scholars such as Jia and Tai (2007) analyze the semantic types and characteristics of FrameNet. Fang (1999) reviews the semantic types and semantic relations in UMLS and analyzes the characteristics of their semantic relations. Regarding semantic type determination, Jia and Wang (2010) based on the semantic type of FrameNet, automatically determine the semantic types of the frame elements in the framework network ontology using a variety of methods. Regarding semantic type application Delbecque et al. (2005) use the semantic types of UMLS for medically specific named entity annotations; Wang and Tai (2017) propose applying the semantic types of UMLS to social tagging systems to classify tags of social tagging systems and enrich the semantic information of social tags. Mi and Cao (2012) construct a medical literature ontology based on semantic types and semantic relations of UMLS.

Scholars have initiated useful discussions on the semantic research related to tags and achieved some results. However, their work does not involve research on identifying semantic types of tags in social tagging systems to enrich their semantic information. Therefore, we attempt to classify and enrich the semantic information of tags based on semantic type. This article reports on how to determine the semantic type of social tags.

3.0 Construction of a semantic type library

3.1 Semantic type data sources

3.1.1 UMLS

The Unified Medical Language System (UMLS) is an integrated information retrieval language system for the biomedical and health fields developed by the National Library of Medicine (NLM), mainly used in natural language standardization processing, information indexing and intelligent retrieval (Fang 1999). This system includes UMLS knowledge sources and related tools. The knowledge sources include three parts: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus is a large vocabulary database that contains multi-language versions for multiple uses. It includes concepts and terms, the relationship between these concepts and the semantic types of the concepts in the field of biomedicine and health. The Semantic Network consists of a set of basic semantic types and the semantic relations describing the relationships among these semantic types. Each concept in the Metathesaurus at least can be assigned to a corresponding semantic type. The semantic types comprise the nodes in the Semantic Network, and the semantic relations are the links between them. The SPECIALIST Lexicon is a comprehensive English dictionary, which includes both common English words and biomedical terms. The SPECIALIST Lexicon Language Processing System can process the differences in the syntax and spelling of natural language words and terms, such as the spelling differences between British English and American English and the variation of character sets (UMLS Reference Manual 2016). Although the semantic types in UMLS can be used for tag classification, its

focus is on the classification of tags and terms in the biomedicine field. The effect is unsatisfactory for the classification of tags and terms in general fields.

3.1.2 FrameNet

FrameNet is a frame network ontology. As the top-level ontology, the semantic type of FrameNet classifies concepts from the most general point of view. In a social tagging system, the user adds both professional and non-professional tags to a specific domain of resources. For example, in the Douban reading website (http://book.douban.com/), we can search for the book *Medical Biochemistry*, which is tagged by the members of Douban as follows: "biochemistry," "medicine," "professional," "college textbooks," "learning," "teaching materials" and "my medicine." Although this book is a professional book from the biomedical field, each user's point of view is different, and thus, the tags may also have nonprofessional characteristics. If we combine the semantic types of UMLS and Frame Net, we can enrich the semantic types of UMLS in the public domain and better apply it to the classification of tags in social tagging systems.

3.2 Semantic type mapping between UMLS and FrameNet

The basic information about the name, abbreviation, definition and number of each semantic type in UMLS and FrameNet is sorted out, and the term "definition" explains the concrete meaning and the application scope of the semantic type. Therefore, we analyze the corresponding condition of each semantic type in UMLS and FrameNet based on the meaning and spelling of semantic types, in order to supplement the semantic types of the common domain in UMLS.

According to the correspondence between each semantic type in UMLS and FrameNet, based on the vocabulary mapping method (Chaplan 1995; Zeng and Chan 2004), the semantic type match is divided into six types:

1) Exact match: semantic types that are completely the same in spelling and meaning, such as "Physical_object (68)" in FrameNet and "Physical_Object (T072)" in UMLS. A total of six semantic types meet this type, accounting for 13.33% of total semantic types in FrameNet.
2) Concept match: semantic types that have different spellings but the same meaning, such as "Artifact (61)" in Frame Net and "Manufactured Object (T073)" in UMLS. A total of two semantic types meet this type, accounting for 4.44% of total semantic types in FrameNet.
3) Subordination match: semantic types that have a part-whole relationship, such as "Location (54)" in FrameNet and "Spatial Concept (T082)" in UMLS. A total of eight semantic types meet this type, accounting for 17.78% of total semantic types in FrameNet.
4) Superordination match: semantic types that have a whole-part relationship, such as "Attribute (154)" in FrameNet and "Group Attribute (T102)" in UMLS. A total of three semantic types meet this type, accounting for 6.67% of total semantic types in FrameNet.
5) Near-synonym match: semantic types that have different spellings but similar meanings, such as "Material (63)" in FrameNet and "Substance (T167)" in UMLS. A total of five semantic types meet this type, accounting for 11.11% of total semantic types in FrameNet.
6) No match: the semantic types of FrameNet that are not included in UMLS, such as: "Container (15)," "Point (175)," or "Line (176)." A total of twenty-one semantic types meet this type, accounting for 46.67% of total semantic types in FrameNet.

Table 1 shows the matching conditions of semantic types of UMLS and FrameNet.

By comparison, we find that semantic types in UMLS are more detailed. For example, semantic type "Group (76)" in FrameNet and "Group (T096)" in UMLS completely correspond. In FrameNet, the semantic type "Group" only has a hypogenous semantic type "Organization (58)," while in UMLS, the semantic type "Group" includes the following lower semantic types: "Professional or Occupational Group (T097)," "Population Group (T098)," "Family Group (T099)," "Age Group (T100)," "Patient or Disabled Group (T101)." Therefore, a condition exists where one semantic type in FrameNet corresponds to many semantic types in UMLS. For example, semantic type "Human_act (69)" in FrameNet corresponds to the semantic types "Social Behavior (T054)" and "Individual Behavior (T055)."

In addition to the corresponding twenty-four semantic types, twenty-one semantic types in FrameNet are not included in UMLS, such as "Message (56)," "Speed (234)," or "Relation (174)" that can be used to classify some nonprofessional tags. Therefore, the forty-five semantic types in FrameNet and 133 semantic types in UMLS were combined according to the principle of seeking common ground and a total of 154 semantic types were set as tag classification classifiers in social tagging systems.

Match type	Meaning	Matched number	Matched semantic types	Percent
Exact match	Semantic types that are completely same in spelling and meaning.	6	Physical_object (68)-Physical_Object (T072)	13.33%
			Organization (58)-Organization (T092)	
			Group (76)-Group (T096)	
			Human (80)-Human (T016)	
			Event (75)-Event (T051)	
			Activity (8)-Activity (T052)	
Concept match	Semantic types that have different spelling but same meaning	2	Artifact (61)-Manufactured Object (T073)	4.44%
			Living_thing (66)-Organism (T001)	
Subordination match	Semantic types that have a part-whole relationship.	8	Location (54)-Spatial Concept (T082)	17.78%
			Body_part (10)-Body Part, Organ, or Organ Component (T023)	
			Physical_entity (70)-Entity (T071)	
			Quantity (59)-Quantitative Concept (T081)	
			Time (141)-Temporal Concept (T079)	
			Duration (142)-Temporal Concept (T079)	
			Degree (172)-Laboratory or Test Result (T034)	
			Intentional_act (181)-Behavior (T053)	
Superordination match	Semantic types that have a whole-part relationship.	3	Attribute (154)-Group Attribute (T102), Clinical Attribute (T201), Organism Attribute (T032)	6.67%
			Structure (62)-Anatomical Structure (T017)	
			Animate_being (65)-Animal (T008)	
Near-synonym match	Semantic types that have different spelling but similar meaning.	5	Material (63)-Substance (T167)	11.11%
			Sentient (5)-Vertebrate (T010) Mammal (T015)	
			State (77)-Phenomenon or Process (T067)	
			Human_act (69)-Social Behavior (T054) Individual Behavior (T055)	
			Region (17)-Geographic Area (T083)	
No match	Semantic types of FrameNet that are not involved in UMLS.	21	Container (15), Point (175), Line (176), Body_of_water (2), Running-water (3), Landform (7), Shape (60), Manner (173), Temperature (233), Speed (234), Relation (174), Social relation (57),Locative_relation (182), Source (151), Path (152), Goal (153),State_of_affairs (177), Content (55), Message (56), Achievement (19), Accomplishment (20)	46.67%

Table 1. The matching conditions of semantic types of UMLS and FrameNet. (Note: The numeral in the table is the number of each semantic type with no substantive significance.)

4.0 The determination of semantic types for social tags

4.1 UMLS natural language processing tools

UMLS natural language processing tools are powerful, providing multi-processing tools, such as the SPECIALIST Lexicon, LexAccess, Lexical Tools, Text Tools, Text Categorization, GSpell, DTagger, Visual Tagging Tool, Sub-Term Mapping Tools, MEDLINE N-Gram and Semantic Navigator (UMLS Reference Manual 2016).

Among them, the LexAccess tool can be used to find out the prototype of a word, part of speech, spelling variants and other information. The Norm prototype tool in Lexical Tools can process any input words in turn as follows: removing the possessive case, using spaces instead of punctuation, removing stop words, converting uppercase to lowercase, extracting the prototype of each word, and, fi-

nally, sorting words in alphabetical order. The GSpell tool can be used to check spelling and provide suggestions for the misspelled words. The DTagger tool can tag the part of speech. The Semantic Navigator tool can suggest possible semantic types of a tag based on the knowledge base. The ST WSD tool in the Semantic Navigator tool can be used to filter out the best semantic type of a tag from the selected semantic types by combining the contextual information of the tag (UMLS Reference Manual 2016).

4.2 The steps for determining semantic type for tags

Step 1: The collected tags are pre-processed, such as by filtering and standardizing.

Step 2: For tags that have been preprocessed, we directly match the semantic type from 154 semantic types in the library. If the semantic type can exactly match the tag, then this result is used as the semantic type of the tag. It also holds the semantic types that only partially match as possible candidates.

Step 3: For tags whose semantic type cannot be determined by direct match in step two, we retrieve a semantic type for each word via the Semantic Navigator tool in UMLS. If only one result and no partially matched semantic types are retrieved in step two, then we determine the one result is the semantic type of this tag.

Step 4: If there are multiple results in step three or there are partially matched semantic types in step two, then it is necessary to further use the ST WSD tool, combined with the abstract of the resource that was tagged to filter out the best semantic type of the tag. If there is no result from step three, the tag's spelling variant is used to query again.

Step 5: If the semantic type is still unable to be determined through above steps, we do so manually.

The overall process is shown in Figure 1. Then, we analyze the process with examples.

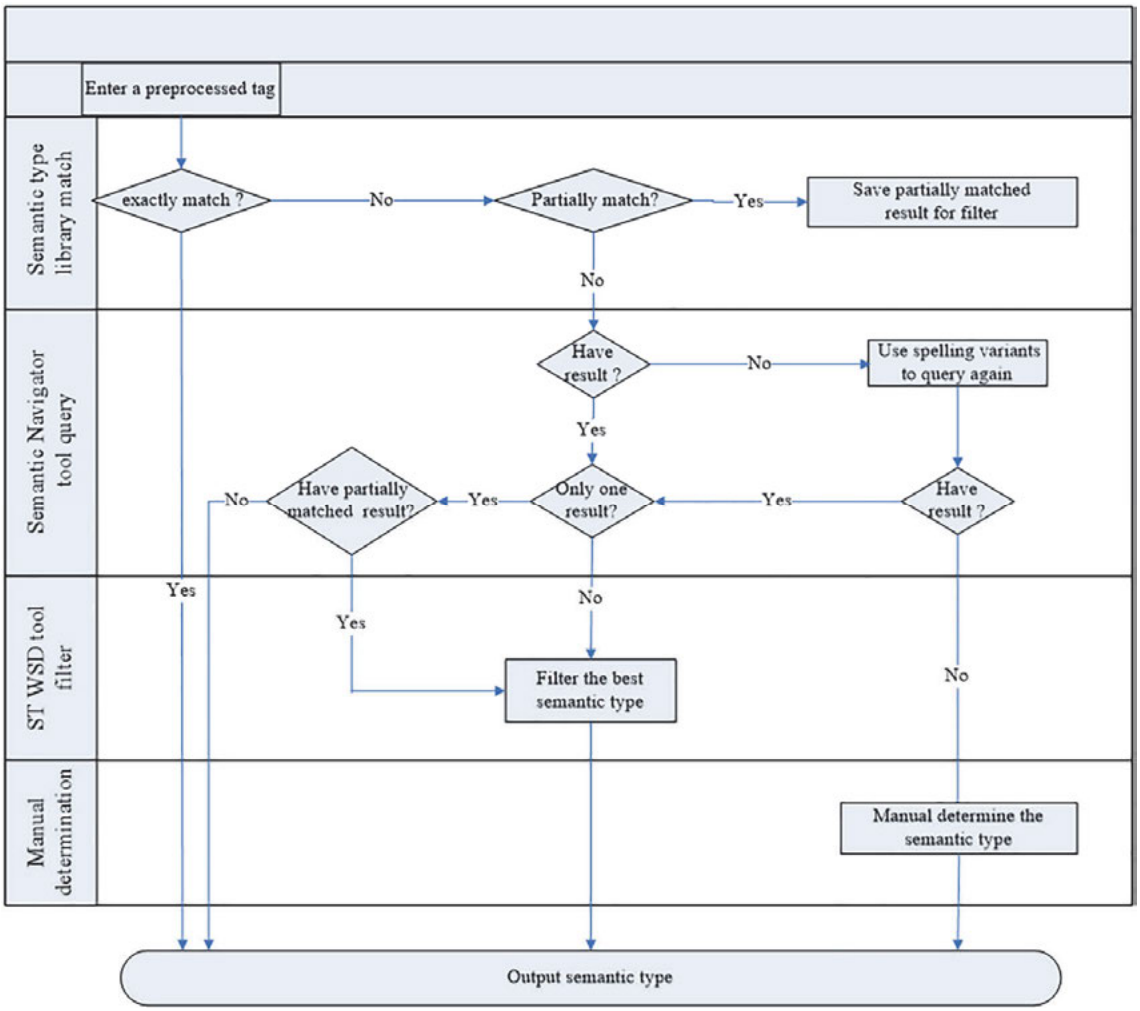


Figure 1. The determination flow of semantic type for tag.

4.2.1 Tag collection and preprocessing

4.2.1.1 Tag collection

At present, some mature social tagging systems exist in the medical field (Tai et al. 2014) such as BibSonomy, DocGuide, PatientsLikeMe, TuDiabetes, Qingko and Docin Medical. In these systems, users can add tags to resources that they released, and they can also collect resources that other people released or add tags for these resources. Through tags, users can find similar resources or find users who have similar interests or a similar community with which to interact. Some systems provide tag retrieval and recommendation functions. Moreover, users can also customize tags that interest them. Among these systems, the BibSonomy system mainly has webpage resources, electronic documents, books, journals and other publications about the biomedicine field. It is the one of the existing three annotation systems (BibSonomy, CiteULike, Connotea) mainly for academic researchers (Borrego et al. 2012). Therefore, we collected the tags from BibSonomy (http://www.bibsonomy.org/). For this research, we chose ten subject heading terms in the level three category "Diseases[C]" of from *MeSH*, as well as the entry terms of the ten subject heading terms as retrieval words retrieved by tag matching. For example, we conducted tag matching by using the subject heading term "Liver Neoplasms" in *MeSH* as the search term. One document (a "publication" type) was retrieved, with a total of sixty-three tags. The subject heading term "Liver Neoplasms" has twenty-one entry words in *MeSH*: "Neoplasms, Hepatic," "Neoplasms, Liver," "Liver Neoplasm," "Neoplasm, Liver," "Hepatic Neoplasms," "Hepatic Neoplasm," "Neoplasm, Hepatic," "Cancer of Liver," "Hepatocellular Cancer," "Cancers, Hepatocellular," "Hepatocellular Cancers," "Hepatic Cancer," "Cancer, Hepatic," "Cancers, Hepatic," "Hepatic Cancers," "Liver Cancer," "Cancer, Liver," "Cancers, Liver," "Liver Cancers," "Cancer of the Liver" and "Cancer, Hepatocellular." Then we used the twenty-one entry words as keywords for retrieval in the BibSonomy system by tag matching. Eight documents were retrieved, of which seven are "publication" type and one is "bookmarks" type. One of the eight documenta is a duplicate. There are 402 tags in total, excluding the repeated documents. Some of them were repeated in the same document or in different documents. There are 197 tags in total except for repeated tags. Table 2 shows all the tags collected in this study.

4.2.1.2 Tag features

After analyzing the tags, we found that various forms occur and mainly have the following features:

1) using numbers as tags, such as "1," "40," "15117829;"
2) using letters as tags, such as: "A;"
3) using numbers and letters as tags, such as: "99m;"
4) using prepositions as tags, such as: "of," "as;"
5) using words with special meanings as tags, such as: "Gov't;"
6) using abbreviations as tags, such as: "CHO;"
7) using compound words as tags, such as: "Non-U.S;"

Subject heading terms	MeSH Unique ID	Number of entry terms	Document number excluding repeated documents	Number of tags	Number of tags excluding repeated tags
Liver Neoplasms	D008113	21	8	402	197
Gastrointestinal Neoplasms	D005770	11	11	393	274
Voice Disorders	D014832	13	4	75	46
Tooth Abnormalities	D014071	9	2	36	27
Fibromyalgia	D005356	25	6	56	41
Osteoarthritis	D010003	8	5	48	31
Bone Neoplasms	D001859	6	7	229	154
Tuberous Sclerosis	D014402	27	18	486	238
Periodontal Diseases	D010510	6	7	90	40
Salivary Gland Diseases	D012466	5	2	74	67
Total			70	1889	1115

Table 2. The tags information of collected tags. (Note: The date range of tag acquisition is 2016.04-2017.11)

8) using a combination of several words as a tag, such as: "diagnosis/genetics/mortality;"
9) using different forms of the same word as two tags, such as: "human, humans;"
10) the same tag appearing repeatedly in the same resource or different resources; and,
11) a symbol before or after the tag, such as: "$-Calmodulin."

4.2.1.3 Tag preprocessing

Among these various forms of tags, some tags are without any real meaning, such as preposition tags. Some tags can only be understood by the taggers themselves, such as number tags, letter tags, and some tags that are not prototypes, such as "humans." Therefore, it is necessary to preprocess these tags for standardization before determining the semantic type.

During preprocessing, we filter the numbers, letters, prepositions, and non-substantive tags first. Regardless of repetition, they make up 4.06% of the total tags. In the remaining tags, 3.55% of the total tags have no results found using the LexAccess tool to retrieve their prototype. Part are misspelled, and the GSpell tool is used to retrieve the correct spelling (for example, "triterpenoids" spelled as "triterpinoids;" "therapy" spelled as "erap") and finding the correct prototype is attempted. The remaining tags are abbreviations or compound words, which need to be combined with other tags to provide the original meaning. The remaining 92.39% of the total tags can be output to the prototype directly. The tag preprocessing flow is shown in Figure 2.

In the process of tag prototype preprocessing with the LexAccess tool, we retain all the parts of speech and spelling variants of tags for use in the next step. As shown in Figure 3, the tag "Dose-Response" is prototyped in the form of "doseresponse," with the spelling variants "dose response" and "dose-response."

In the process of tag prototyping, some tags result in the same prototype; for example, the tags "Cell" and "Cells" have the same prototype, "Cell." Therefore, without repetition, 1,013 tag remains after prototype processing.

4.2.2 Matching semantic type directly

The semantic types of some tags can be determined directly by matching entries from the semantic type library. For example, the tag "animal" and the semantic type "Animal (T008)," the tag "region" and the semantic type "Region (17)" match exactly. In those cases, the semantic types of these tags are the same as the tags themselves. Some tags partly match some semantic types. For example, the tag "acid" partly matches the semantic types "Amino Acid Sequence (T087); Nucleic Acid, Nucleoside, or Nucleotide (T114); Amino Acid, Peptide, or Protein (T116)." We also save the partly matched semantic types as alternative semantic types for the use in the filtering phase.

4.2.3 Determining the semantic type using the Semantic Navigator tool

For tags that do not directly match with a semantic type from the library, we use the Semantic Navigator tool. The Semantic Navigator tool retrieves two categories of semantic type. One is a tag that only has a semantic type result, such as the tag "Alignment," which only has a semantic type "Quantitative Concept (T081)" retrieved with the Semantic Navigator tool. The other type has many semantic type results. For example, the tag "fusion" has the semantic types "Functional Concept (T169)" and "Therapeutic or Preventive Procedure (T061)" retrieved with the Semantic Navigator tool. For these tags, we need to filter to determine their specific semantic types. For tags that only have one result when retrieved with the Semantic Navigator tool and no partly matched result when matched with the semantic type library, we determine the result as the semantic type of this tag. For example, the tag "endothelial" has no result when retrieved in the semantic type library, and there is only one result, "Tissue (T024)" when retrieved with the Semantic Navigator tool. Thus, we determine "Tissue (T024)" is the semantic type of tag "endothelial."

The Semantic Navigator tool can judge the vocabulary that was input according to certain rules; if necessary, some tags will be replaced with their synonyms when retrieving. For example, when we input the tag "multiple" that was prototyped for retrieval with the Semantic Navigator tool, the output is the word "numerous." For this automatic replacement, we still must manually audit. Some of the results of the automatic replacement are not consistent with the original intention of the tagger. For example, when the tag "Cornell" that was prototyped is entered into the Semantic Navigator tool, the output is the words "Cornell Medical Index." Obviously, this is not consistent with the original intention of the tagger. For these tags, we need to manually modify the semantic type.

Some words retrieve no result with the Semantic Navigator tool. In those cases, we retrieve again using the words' variants searched in the LexAccess tool. For example, the prototype of the tag "DNA-Binding" is "DNAbinding," but no results are retrieved when we use the word "DNAbinding" to retrieve using the Semantic Navigator tool. Then, we use the tag's spelling variant "DNA-binding" to retrieve again, and only one semantic type results: "Genetic Function (T045)." Therefore, we determine "Genetic Function (T045)" as the semantic type of this tag.

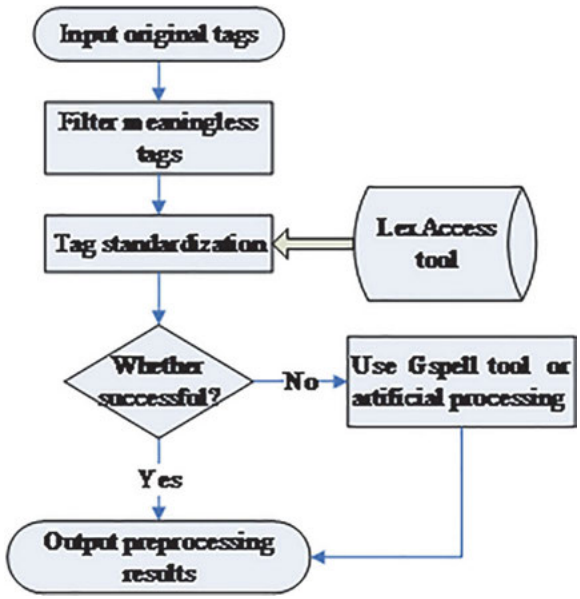


Figure 2. The flow of tag preprocessing.

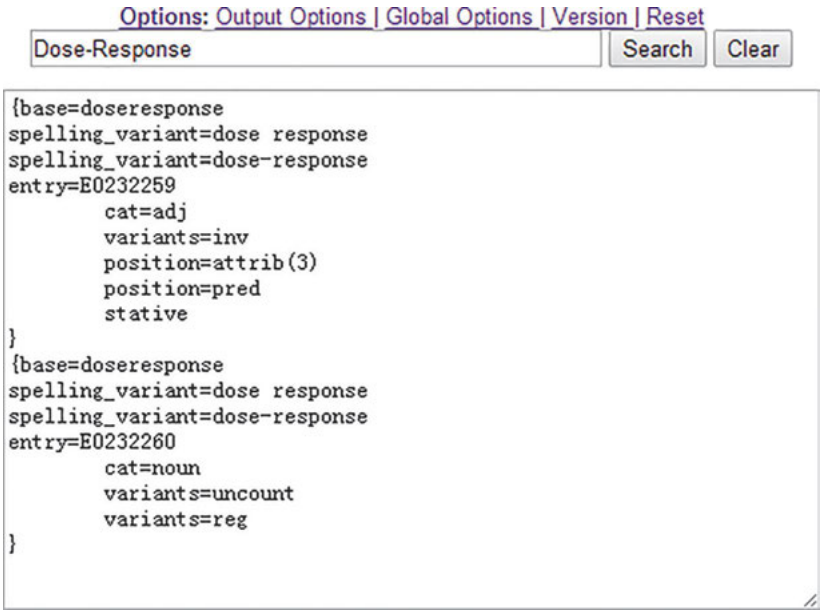


Figure 3. The result of tag prototype processing with the LexAccess tool.

4.2.4 Determination of the semantic type using the ST WSD tool

For tags that result in partial matches with the semantic type library or have more than one result when retrieved with the Semantic Navigator tool, we need to use the ST WSD tool combined with the tag's contextual information to select the semantic type most suitable for the tag among the existing alternative semantic types. As shown in Figure 4, the tag "acid" has two alternative semantic types: "Laboratory Procedure (T059)" and "Chemical (T103)" when retrieved using the Semantic Navigator tool and the partly matched semantic types of "Amino Acid Sequence (T087)," "Nucleic Acid, Nucleoside, or Nucleotide (T114)" and "Amino Acid, Peptide or Protein (T116)." Then, we use the ST WSD tool combined with the abstract information of the resource that was tagged to determine that the best semantic type of the tag is "Amino Acid, Peptide, or Protein (T116)."

4.2.5 Determining the semantic type manually

For tags whose semantic types cannot be determined by the methods described above, or tags that have been automatically changed when the semantic type was retrieved with the Semantic Navigator tool and the result is still in doubt, then, we need to determine their semantic types

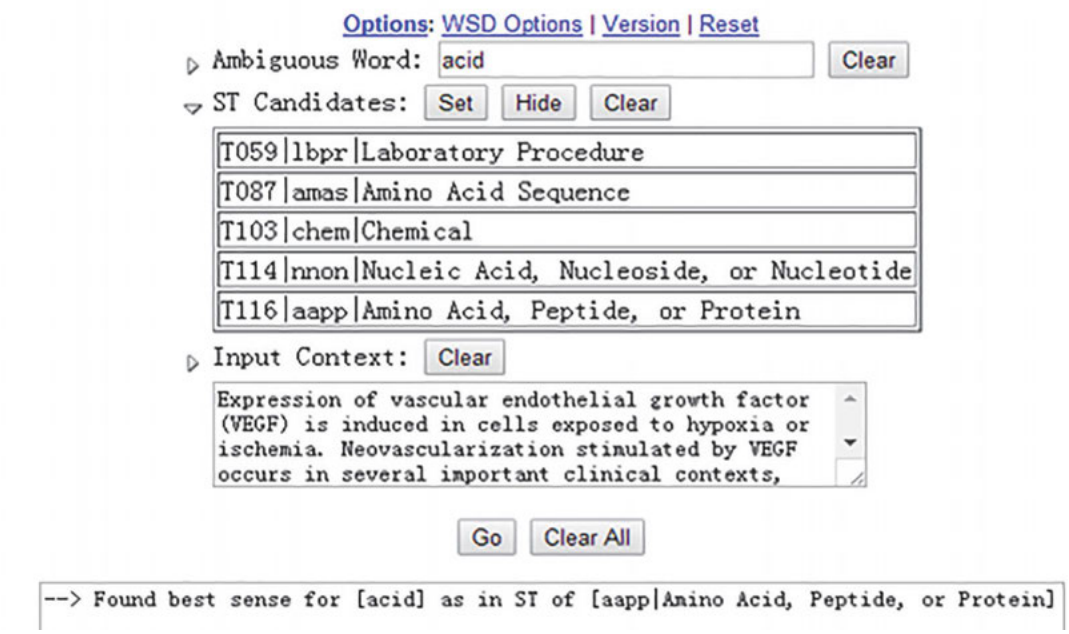


Figure 4. Determining the semantic type using the ST WSD tool.

manually. For example, the semantic type of the tag "Hypoxia-Inducible" cannot be determined by the methods above, so we combine other tags and this annotation resource information and determine the tag's semantic type is "Immunologic Factor (T129)." When we retrieve the semantic type of the tag "Cornell" with the Semantic Navigator tool, the output result is the phrase "Cornell Medical Index," which is not in UMLS's semantic type library. Therefore, we combine the abstract and heading infor mation of this annotated resource to manually modify the semantic type as "Geographic Area (T083)."

4.3 An example

We take the resource titled "Activation of endothelial factor gene transcription by hypoxia-inducible factor growth" in the resource library as an example. This resource has a total of sixty-three tags, some of which have no real meaning, such as "1," "40," "A," some of which have the same prototype with different spelling variants, such as "Cells," "Cell," and some tags that repeatedly appear. After filtering tags that have no real meaning and standardization processing, a total of fifty tags remain, excluding repeated tags. As shown in Table 3, among the fifty-three tags, fifty-two tags' semantic type can be determined through matching with the semantic type library directly, using the Semantic Navigator and ST WSD tools combined with the resource abstract information. Only one tag needs its semantic type to be determined manually. These fifty-three tags can be classified into thirty-four categories, of which there are thirty-three kinds of semantic types in UMLS; the other one is a semantic type in FrameNet. The semantic type of the tag "Regions" is "Region (17)" in FrameNet. Semantic type shows to be a good classifier for tags.

4.4 Classification results of semantic types

In this paper, the semantic type determination methods described above are used to classify the 1,013 tags that were prototyped. Among them, forty-five tags' semantic types were determined though matching with the semantic type library directly, accounting for 4.47%; 272 tags' semantic types were determined directly using the Semantic Navigator tool, accounting for 26.82%; 685 tags' semantic types were determined using the ST WSD tool, accounting for 67.6%; eleven tags' semantic types were determined manually, accounting for 1.12%. On the whole, the proportion of automatic determination is higher, and the results are consistent with manual judgment.

5.0 Conclusion

In this study we found that:

1) During the process of collecting tags, 95.76% of the *MeSH* terms have no corresponding search result when retrieved from BibSonomy; 2.47% of the *MeSH* terms rarely retrieved results; only 1.77% of the *MeSH* terms retrieved more items to meet the demands of the study. Therefore, the choice of the

Resource number	Tags	Semantic types	The source of semantic type	Determination methods
biotags-01-01	genetics	T169|Functional Concept	UMLS	ST WSD tool
	fusion			ST WSD tool
	Hypoxia-Inducible			Manually determined
	acid	T116|Amino Acid, Peptide, or Protein	UMLS	ST WSD tool
	lymphokine			ST WSD tool
	protein			ST WSD tool
	simian			ST WSD tool
	subunit			ST WSD tool
	alignment	T081|Quantitative Concept	UMLS	Semantic Navigator tool
	animal	T008|Animal	UMLS	Matches library directly
	base	T028|Gene or Genome	UMLS	ST WSD tool
	gene			ST WSD tool
	homology			ST WSD tool
	bind	T044|Molecular Function	UMLS	ST WSD tool
	molecular			ST WSD tool
	carcinoma	T191|Neoplastic Process	UMLS	Semantic Navigator tool
	neoplasm			ST WSD tool
	cell	T025|Cell	UMLS	Matches library directly
	chemistry	T059|Laboratory Procedure	UMLS	ST WSD tool
	culture			ST WSD tool
	rat			ST WSD tool
	pathology			ST WSD tool
	DNAbinding	T045|Genetic Function	UMLS	Semantic Navigator tool
	genetic			ST WSD tool
	expression			ST WSD tool
	transcription			Semantic Navigator tool
	data	T074|Medical Device	UMLS	ST WSD tool
	endothelial	T024|Tissue	UMLS	Semantic Navigator tool
	factor	T077|Conceptual Entity	UMLS	ST WSD tool
	reporter			ST WSD tool
	growth	T052|Activity	UMLS	ST WSD tool
	hepatocellular	T080|Qualitative Concept	UMLS	Semantic Navigator tool
	vascular			ST WSD tool
	human	T016|Human	UMLS	Matches library directly
	hypoxia	T046|Pathologic Function	UMLS	ST WSD tool
	liver	T121|Pharmacologic Substance	UMLS	ST WSD tool
	mouse	T015|Mammal	UMLS	ST WSD tool
	nuclear	T082|Spatial Concept	UMLS	Semantic Navigator tool
	nucleic	T026|Cell Component	UMLS	ST WSD tool
	promoter	T123|Biologically Active Substance	UMLS	ST WSD tool
	recombinant	T001|Organism	UMLS	Semantic Navigator tool
	region	17|Region	FrameNet	Matches library directly
	regulation	T038|Biologic Function	UMLS	ST WSD tool
	biosynthesis			ST WSD tool
	research	T062|Research Activity	UMLS	ST WSD tool
	sequence	T087|Amino Acid Sequence	UMLS	ST WSD tool
	transfection	T063| Molecular Biology Retrieve Technique	UMLS	Semantic Navigator tool
	tumor	T033|Finding	UMLS	ST WSD tool
	alfa	T170|Intellectual Product	UMLS	Semantic Navigator tool
	metabolism	T043|Cell Function	UMLS	ST WSD tool
	physiology	T039|Physiologic Function	UMLS	ST WSD tool
	response	T032|Organism Attribute	UMLS	ST WSD tool
	virus	T005|Virus	UMLS	Matches library directly

Table 3. The tags of a resource and the corresponding semantic types.

subject heading term is very important when the tag is collected.

2) The standardization of tags in tagging systems is too weak, and consequently, meaningless tags, number tags, letter tags, compound word tags and symbol tags appear frequently, and only 53.6% of the tags collected can be studied after pre-processing. Therefore, it is also important to select a reasonable tagging system.
3) When the semantic type is determined, most tags need to combine information from the abstract of the tagged resources to determine the final semantic type from the alternative semantic types, while 25.7% of the collected resources do not have abstracts, which then will affect the accuracy of the results. Therefore, it is necessary to further screen the resources collected.
4) Tags that need the semantic type to be determined manually are highly dependent on the relevant background knowledge of the determiner, and therefore the accuracy of the determined semantic type is difficult to guarantee. Therefore, the determination of semantic types requires the participation of experts in the field.

We collected tags from the biomedical field tagging system and constructed a semantic type library for the classification of tags. For tags that were preprocessed, we determined their semantic types by matching with the semantic type library directly, using the Semantic Navigator tool, the ST WSD tool, and manual determination. This method can effectively determine the semantic type of tags and tag classification in a social tagging system. It also lays the foundation for enriching the semantic information of tags, tagging system mapping with ontologies, and other follow-up work.

We also summarized the determination rules of semantic type of tags by only analyzing the sample data. In further phases, we will expand the research sample and continue to supplement the details of tags' semantic type determination, for example, using the WordNet tool to aid determination of the tags whose semantic types need to be determined manually. Finally, the classification of the semantic types of tags can be realized automatically.

References

Al-Khalifa, Hend S. and Hugh C. Davis. 2006. "Folks Annotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies." *Innovations in Information Technology* 1-5. doi: 10.1109/INNOVATIONS.2006.301927

Borrego, Angel and Jenny Fry. 2012. "Measuring Researchers' Use of Scholarly Information Through Social Bookmarking Data: A Case Study of BibSonomy." *Journal of Information Science* 38: 297-308. doi:10.1177/0165551512438353

Cattuto, Ciro, Dominik Benz, Andreas Hotho and Gerd Stumme. 2008. "Semantic Grounding of Tag Relatedness in Social Bookmarking Systems." In *The Semantic Web - ISWC 2008: 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008, Proceedings*, ed. Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin and Krishnaprasad Thirunarayan. Lecture Notes in Computer Science, vol 5318. Berlin: Springer, 615-31. doi:10.1007/978-3-540-88564-1-39

Chaplan, Margarent A. 1995. "Mapping 'Laborline Thesaurus' Terms to Library of Congress Subject Headings: Implications for Vocabulary Switching." *Library Quarterly* 65: 39-61. doi:10.1086/602752

Cui, Jianwei, Hongyan Liu, Jun He, Pei Li, Xiaoyong Du and Puwei Wang. 2011. "Tagclus: A Random Walk-Based Method for Tag Clustering." *Knowledge and Information Systems* 27: 193-225. doi:10.1007/s10115-010-0307-y

Delbecque, Thierry, Pierre Jacquemart and Pierre Zweigenbaum. 2005. "Indexing UMLS Semantic Types for Medical Question-Answering." *Studies in Health Technology and Informatics* 116: 805-10.

Djuana, Endang, Yue Xu and Yuefeng Li. 2011. "Constructing Tag Ontology from Folksonomy based on WordNet." In *Proceedings of the IADIS International Conference on Internet Technologies and Society 2011, International Association for Development of the Information Society (IADIS), The East China Normal University, Shanghai*, ed. Piet Kommers, Jiping Zhang, Tomayess Issaand and Pedro Isaías. https://eprints.qut.edu.au/46776/

Fang, Ping. 1999. "Study on Characteristics and Structure of Semantic Network of UMLS Knowledge Sources." *Journal of the China Society For Scientific and Technical Information* 18:129-34. doi10.3969/j.issn.1000-0135.1999.02.006

García-Plaza, Alberto Pérez, Arkaitz Zubiaga, Victor Fresno and Raqel Martinez. 2012. "Reorganizing Clouds: A Study on Tag Clustering and Evaluation." *Expert Systems with Applications* 39: 9483-9493. doi:10.1016/j.eswa.2012.02.108

García-Silva, Andres, GJael arcía-Castro, Alexander García Alexander, Oscar Corcho and Asuncion Gomez-Perez. 2012. "Building Ontologies from Folksonomies and Linked Data: Data structures and Algorithms." [Technical Report] Facultad de Informática (UPM), Ontology Engineering Group. Facultad de Informática. Universidad Politécnica de Madrid.

Ghali, Fawaz, Mike Sharp and Alexandra I. Cristea. 2011. "Folksonomies and Ontologies in Authoring of Adaptive Hypermedia." *Educational Technology & Society* 8: 6-8.

Good, Benjamin M., Edward A. Kawas and Mark D. Wilkinson. 2007. "Bridging the Gap between Social Tagging and Semantic Annotation: E.D. the Entity Describer." *Nature Precedings* http://precedings.nature.com/documents/945/version/2

Gupta Manish, Rui Li, Zhijun Yin and Jiawei Han. 2010. "Survey on Social Tagging Techniques." *ACM SIGKDD Explorations Newsletter* 12:58-72. doi:10.1145/1882471.1882480

Han, Xiaogang, Zhiqi Shen, Chunyan Miao and Luo, Xudong. 2010. "Folksonomy-Based Ontological User Interest Profile Modeling and Its Application in Personalized Search." In *International Conference on Active Media Technology, 6th International Conference, AMT 2010, Toronto, Canada, August 28-30, 2010, Proceedings*, ed. Aijun An, Pawan Lingras, Sheila Petty and Runhe Huang. Lecture Notes in Computer Science, vol. 6335. Berlin: Springer, 34-46. doi:10.1007/978-3-642-15470-6_6

Jia, Junzhi and Yangfang Tai. 2007. "Research on Semantic Types of FrameNet." *Information Studies Theory & Application* 30: 689-92. doi:10.3969/j.issn.1000-7490.2007.05.032

Jia, Junzhi and Yongfang Wang. 2010. "Automatic Determination of Semantic Types in Frame Elements." *Journal of Intelligence* 29:120-23. doi:10.3969/j.issn.1002-1965.2010.06.027

Kiu, Ching-Chieh and Eric Tsui. 2011. "TaxoFolk: A Hybrid Taxonomy–Folksonomy Structure For Knowledge Classification and Navigation." *Expert Systems with Applications* 38: 6049-58. doi:10.1057/kmrp.2009.33

Kipp, Margaret E. I. 2007. "Tagging for Health Information Organisation and Retrieval." *Proceedings of the North American Symposium on Knowledge Organization* 1:63-74.

Lee, Danielle H. and Titus Schleyer. 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: 1747-57. doi:10.1002/asi.22653

Lee, Hyun Jung and Mye Sohn. 2013. "Tag-Based Integrated Semantic Ontology Construction and Evolution." In *Proceedings: Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 3-5 July, 3013, Asia University, Taichung, Taiwan*, ed. Leonard Barolli, Ilsun You, Fatos Xhafa, Fang-Yie Leu and Hsing-Chung Chen. Piscataway, N.J.: IEEE, 221-27. doi:10.1109/IMIS.2013.45

Li, Huizong. "Tag Clustering Method in Social Annotation Environment." PhD diss. HeFei University of Technology.

Li, Huizong, Xuegang Hu, Wei He and Jianhan Pan. 2014. "Tags Co-occurrence Spectral Clustering Method in Social Tagging Environment." *Library & Information Service* 58: 129-35. doi:10.13266/j.issn.0252-3116.2014.23.020

Lux, Mathias and Gisela Dosinger. 2007. "From Folksonomies to Ontologies: Employing Wisdom of the Crowds to Serve Learning Purposes." *International Journal of Knowledge and Learning* 3: 515-28. doi:10.1504/IJKL.2007.016709

Markines, Benjamin, Ciro Cattuto, Fillippo Menczer, Dominik Benz, Andreas Hotho and Gerd Stumme. 2009. "Evaluating Similarity Measures for Emergent Semantics of Social Tagging." In *Proceeding WWW '09: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, April 20-24, 2009*. New York: ACM, 641-50. doi:10.1145/1526709.1526796

Munk, Timme Bisgaard and Kristian Mørk. 2007a. "Folksonomy, the Power Law & the Significance of the Least Effort." *Knowledge Organization* 34: 16-33.

Munk, Timme Bisgaard and Kristian Mørk. 2007b. "Folksonomies, Tagging Communities, and Tagging Strategies—An Empirical Study. *Knowledge Organization* 34: 115-27.

Mi, Yang and Jindan Cao. 2012. "An Empirical Study on the Construction of Medical Literature Ontology by Using UMLS Semantic Network." *Research on Library Science* 7: 55-60. doi:10.15941/j.cnki.issn1001-0424.2012.07.009

Radelaar, Joni, Aart-Jan Boor, Damir Vandic, Jan-Willem Van Dam, Frederik Hogenboom and Flavius Frasincar. 2011. "Improving the Exploration of Tag Spaces Using Automated Tag Clustering." In *Web Engineering, 11th International Conference, ICWE 2011, Paphos, Cyprus, June 20-24, 2011*, ed. Sören Auer, Oscar Díaz and George A. Papadopoulos. Berlin: Springer, 274-88. doi:10.1007/978-3-642-22233-7_19

Specia, Lucia and Enrico Motta. 2007. "Integrating Folksonomies with the Semantic Web." *The Semantic Web: Research and Applications: 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007, Proceedings*, ed. Enrico Franconi, Michael Kifer and Wolfgang May. Berlin: Springer, 624-39. doi:10.1007/978-3-540-72667-8_44

Tai, Yangfang, Fangfang Li and Peifeng He. 2014. "Study on the Application of Social Tagging in Medical and Health Internet Information Resources." *Digital Library Forum* 123:7-13. doi:10.3775/j.issn.1673-2286.2014.08.002

Unified Medical Language System. 2017. *UMLS Reference Manual.* http://www.ncbi.nlm.nih.gov/books/NBK9676

Wang, Yongfang and Yangfang Tai. 2017. "Study on the Application of UMLS Semantic Network in Social Tag-

ging Systems." *Library and Information Services* 61:89-99. doi:10.13266/j.issn.0252-3116.2017.01.011

Wei, Lai. 2010. "Review of the Research of Semantic Enrichment of Folksonomy Abroad." *Information and Documentation Services* 3: 40-44. doi:10.3969/j.issn.1002-0314.2010.03.009

Wu, Chao and Bo Zhou. 2012. "Tags Are Related: Measurement of Semantic Relatedness Based on Folksonomy Network." *Computing & Informatics* 30:165-85.

Xiong, Huixiang, Min Deng and Siyuan Guo. 2013. "A Research Overview on the Combination of Tag and Ontology in Social Tagging System." *Journal of Intelligence* 8: 136-41. doi:10.3969/j.issn.1002-1965.2013.08.026

Xiong, Huixiang and Wuxuan Jiang. 2017. "Clustering and Recommending Users Based on Tags and Relation Network." *Data Analysis and Knowledge Discovery* 1: 36-46. doi:10.11925/infotech.2096-3467.2017.06.04

Yoo, Donghee, Keunho Choi, Yongmoo Suh, Gunwoo Kim and Hu Changping. 2013. "Building and Evaluating a Collaboratively Built Structured Folksonomy." *Journal of Information Science* 39: 593-607. doi:10.1177/0165551513480309

Zeng, Marcia Lei and Lois Mai Chan. 2004. "Trends and Issues in Establishing Interoperability among Knowledge Organization Systems." *Journal of the American Society for Information Science and Technology* 55: 377-95. doi:10.1002/asi.10387