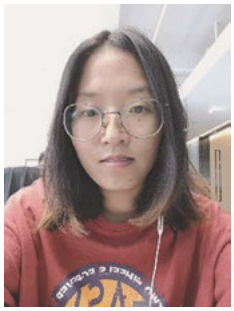
Information Organization Patterns from Online Users in a Social Network†

Chengzhi Zhang*, Hua Zhao**, Xuehua Chi*** and Shuitian Ma****

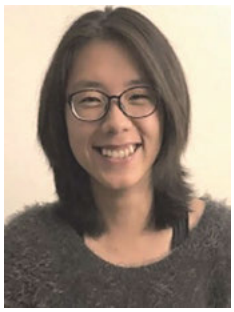
Nanjing University of Science and Technology, Department of Information Management, Nanjing, 210094, China,
*<zhangcz@njust.edu.cn>, **<joyce.chi@qq.com>, ***<1249662620@qq.com>,
****<mashutian0608@hotmail.com>

Chengzhi Zhang is Professor of information science in the Department of Information Management, Nanjing University of Science and Technology. He is a also visiting scholar at the University of Pittsburgh, Pittsburgh (2013) and Senior Research Associate at the City University of Hong Kong, Hong Kong SAR (2010~2011). His research interests are information organization, digital libraries, information retrieval, text mining, natural language processing, etc. He has published approximately 100 articles in refereed journals and conference proceedings including *JASIST*, *ASLIB JIM*, *OIR*, *Scientometrics*, etc. He is an editorial advisory board member of *The Electronic Library*, *Data Intelligence* and *Technology Intelligence Engineering*.



Hua Zhao received her master's degree in information science as well as bachelor's degree in information management and system program from the Nanjing University of Science and Technology, Nanjing, China, in 2014 and 2017, respectively. From 2018, she joined as an algorithm engineer with the Ctrip.com International Ltd, Shanghai, China. Her research interests include text mining and natural language processing.

Xuehua Chi is a graduate student at Nanjing University of Science and Technology. She received her bachelor's degree in information management and information system from Nanjing University of Science and Technology as well. Her research interests include user modeling, text mining, information systems, and knowledge organization.



Shutian Ma is a doctoral student at Nanjing University of Science and Technology. She is in the last year of a master-doctor combined program now and received her bachelor's degree in information management and information system from Nanjing University of Science and Technology as well. Her research interests include citation recommendation, embedding algorithms, clustering and classification, and knowledge organization.

Zhang, Chengzhi, Hua Zhao, Xuehua Chi and Shuitian Ma. 2019. "Information Organization Patterns from Online Users in a Social Network." *Knowledge Organization* 46(2): 90-103. 37 references. DOI:10.5771/0943-7444-2019-2-90.

Abstract: Recent years have seen the rise of user-generated contents (UGCs) in online social media. Diverse UGC sources and information overload are making it increasingly difficult to satisfy personalized information needs. To organize UGCs in a user-centered way, we should not only map them based on textual topics but also link them with users and even user communities. We propose a multi-dimensional framework to organize information by connecting UGCs, users, and user communities. First, we use a topic model to generate a topic hierarchy from UGCs. Second, an author-topic model is applied to learn user interests. Third, user communities are detected through a label propagation algorithm. Finally, a multi-dimensional information organization pattern is formulated based on similarities among the topic hierarchies of UGCs, user interests, and user communities. The results reveal that: 1) our proposed framework can organize information from multiple sources in a user-centered way; 2) hierarchical topic structures can provide comprehensive and in-depth topics for users; and, 3) user communities are efficient in helping people to connect with others who have similar interests.

Received: 3 June 2018; Sixth revision: 17 January 2019; Accepted: 18 January 2019

Keywords: topic model, user, information, social networks

† This work is supported by the National Social Science Fund (No. 14BTQ033), Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Institute of Scientific and Technical Information of China (No. ZD2018-07/01) and Qing Lan Project.

1.0 Introduction

Various online social networks have been created for people to communicate and share information with each other around the world. There is also increasing growth in the numbers of social media users. For example, at the end of September 2017, Twitter had 330 million active monthly users (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/), while Sina Weibo (https://weibo.com), one of the biggest Chinese microblog platforms, had 376 million active monthly users (http://www.useit.com.cn/thread-17562-1-1.html). Simultaneously, vast quantities of user-generated contents (UGCs), referring to images, videos, text, audio, and any other form of content, are being posted by users in online platforms (https://en.wikipedia.org/wiki/User-generated_content). Due to their huge volume, uneven quality, and dynamic changes, UGCs pose new challenges for mining and organizing content (Zhu et al. 2013, 233).

To provide organized information for users, most of the current social networks methodically present UGCs. For instance, Sina Weibo displays posts in chronological order, whereas users may prefer contents to be sorted by topics of their interests. There are a large number of redundant communities and inactive groups in social networks, since most cannot organize information efficiently to meet users' needs (Treem and Leonardi 2012, 143). Researchers are seeking ways to group and streamline large amounts of UGCs (Kietzmann et al. 2011; Van Damme et al. 2007). Traditional knowledge organization tools employ conventional relations between concepts, subjects, and information units, whereas current studies focus more on users, aiming to organize and integrate information in a user-centered way (Hjørland 2003; 2014).

There are three indispensable elements of a social network platform: UGCs, users, and user communities. Since most related research is focused on one element in attempting to organize information (Ming et al. 2014; Zhu et al. 2014), we aim to develop a multi-dimensional information organization system by linking the three elements. To represent UGCs, we employ a topic model to construct a topic hierarchy, which has been widely used in many studies (Zhang 2017; Zhu et al. 2013; Zhu et al. 2014). We then generate profiles of users' interests through user modeling and detect community structure to reveal network organization of users. Finally, to support user-centered information organization, we perform similarity calculations to associate UGCs, users, and user communities with one another.

2.0 Related works

Of the rich set of studies on the organization of online information organization, in this section we discuss mainly research into UGCs and user information behavior.

2.1 Information organization based on UGCs

With the popularization of web 2.0, UGCs have received much attention from researchers. Social tags and folksonomies are becoming popular among different types of UGCs (Kim 2008). Researchers have made particular progress on information organization (Munk and Mørk 2007). Noruzi and Alireza (2006) explored the folksonomy tagging phenomenon and discussed relevant problems. Mathes (2004) provided the first review and survey of social tagging systems. Potnis (2011, 32-35) allowed users to participate in information organization by using folksonomy. Finally, Van Damme et al. (2007) proposed an integrated approach for turning folksonomies into ontologies and using networked knowledge organization systems for information organization.

While mass UGCs provide rich information resources, they also present new challenges, like the integration of heterogeneous contents. Although classifying UGCs in pre-customized categories is simple, classification accuracy depends on the manual maintenance of a category system (Gao 2012, 761). Chen et al. (2011) organized UGCs through a topic model, which does not require maintenance of a classification system. However, the topics obtained from their model were somewhat difficult to understand, and its classification accuracy needs to be improved. Gupta et al. (2010) used social labels for information organization, which greatly reduces manual costs, but their method has problems of data sparseness due to label scarcity. Zhu et al. (2013) proposed topic hierarchy construction for UGCs, while Li (2013) constructed the hierarchical architecture in an entity-based formalism. These studies are very relevant but do not distinguish between different types of UGCs, and few studies have attempted multi-dimensional information organization.

2.2 Information organization based on user information behavior

The research related to user information behavior can be divided into theoretical and empirical works. Theoretical research has explored, for instance, the behavior of information query (Bawden 2006), information interaction (Chen et al. 1998; Keenan et al. 2013), information creation (Maria et al. 2008), information utilization (Kaplan et al. 2010), and information sharing (Stutzman 2006). Empirical research has mainly focused on information behav-

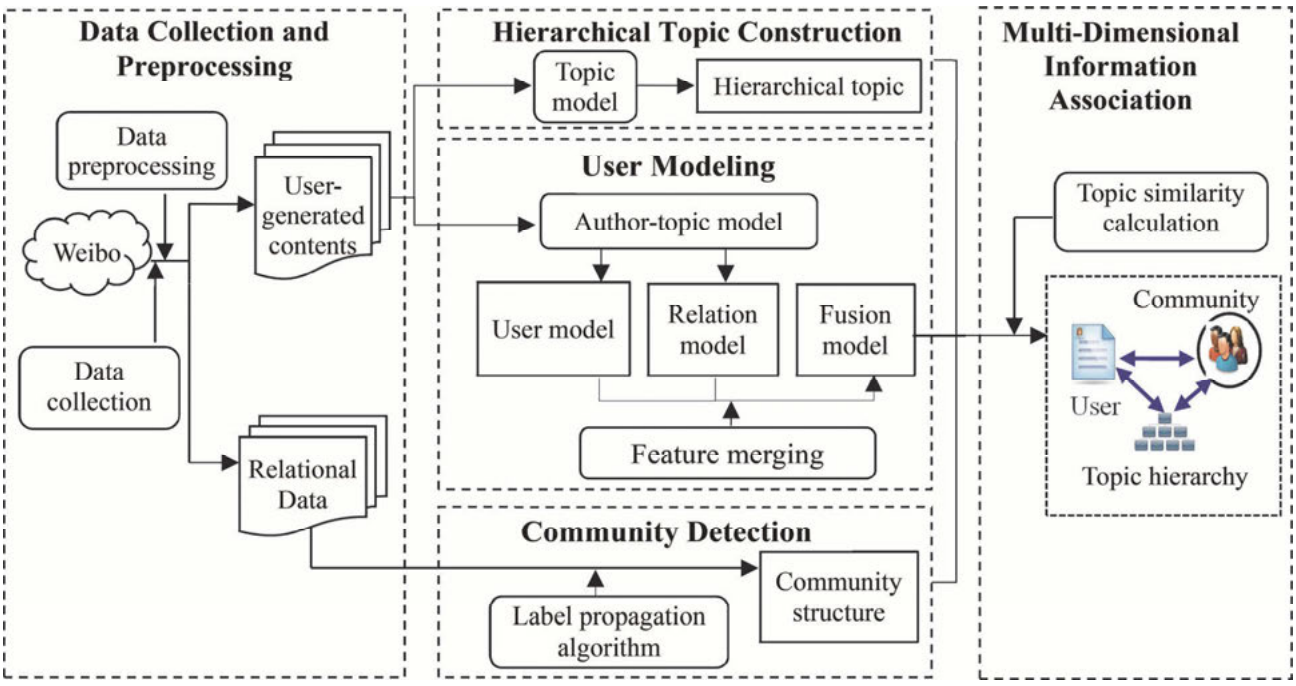


Figure 1. Framework of the study's multi-dimensional information organization.

ior (Benevenuto et al. 2009), user influence (Agichtein et al. 2008), and user social relationships (Kwak et al. 2010). Overall, few studies have explored user behavior with respect to organizing information. Bar-Ilan and Belous (2007) studied the theory and practice of information organization and its relationship with human perception, proposing an intuitive information organization system. Bu et al. (2010) emphasized the importance of considering user behavior, participation, and experience when designing an information organization interface. There is clearly a need for more in-depth research of user behavior, especially with respect to organizing information on social networks.

Traditional methods of information organization rely on semantic or topical relations between textual contents to organize UGCs, without considering the users who generate contents and the user communities in which similar users gather. Despite the increasing shift toward user-centered information integration (Zhou 2010, 36-40), focusing especially on relationships between information and users, only a few studies have considered a multi-dimensional approach to organizing information, combining UGCs with users and user communities. By fully considering the features of UGCs, this study endeavors to associate these three elements to generate a new pattern of knowledge organization in social networks.

3.0 Methodology

3.1 Framework

To connect UGCs, users, and user communities, we attempt to map them in a topic space. Our framework comprises four main steps: 1) hierarchical topic construction for UGCs; 2) user modeling to represent user interests; 3) community detection; and, 4) multi-dimensional information association to link UGCs, users, and user communities. Our experimental process is depicted in Figure 1.

We begin by collecting UGCs from verified users' microblogs and the relationships between users in five domains (football, internet, literature, medicine, law) on Sina Weibo. We then use these UGCs (taken from posts that users create, like, and repost) to generate the topic hierarchy. To obtain a fusion model of users, we merge features learned from users' interests and relations (follower, following) using an author-topic model. To detect user communities, we employ a label propagation algorithm. Finally, a multi-dimensional information organization system is constructed by calculating topic similarities among the UGC topic hierarchy, the user fusion model, and detected user communities.

In the next section, we will discuss the key techniques employed in UGC topic hierarchy construction, user modeling, community detection, and multi-dimensional information association.

3.2 Key techniques

3.2.1 UGC topic hierarchy construction

3.2.1.1 Latent Dirichlet Allocation (LDA)

LDA is a three-level hierarchical Bayesian model (Blei et al. 2003). It assumes that each item in a collection is generated by the finite mixture of topics, each of which is modeled as a multinomial mixture of vocabulary. It also assumes that each document is modeled as a finite mixture over an underlying set of topics. The topic mixture is then drawn from a conjugate Dirichlet prior that remains the same for all documents. The LDA model contains the following parameters: *a*, the Dirichlet priori parameter of document-topic distribution; *β*, the Dirichlet prior parameter of topic-word distribution; *K,* the topic number; *d*, the document; and *z*, the topic. This paper uses the LDA model to derive topical representations for verified users in the five domains of law, medicine, literature, football, and internet.

3.2.1.2 Document topic extraction

After topic modeling, each document is projected into the topic space with different probability distributions, as shown in equation 1 below. We then set a threshold to assign a specific topic to each document; if the topic probability of a document in the topic space exceeds the predefined threshold, the document is assigned to this topic.

$$doc_i = \{topic_1 : w_{i,1}, topic_2 : w_{i,2}, , \ldots, topic_j : w_{i,j}, \ldots, topic_n : w_{i,n}\} \quad \text{(equation 1)}$$

Where $topic_j$ represents topic *j*, doc_i represents document i, $w_{i,j}$ represents the weight of topic *j* in document i. $w_{i,1} + w_{i,2} + \cdots + w_{i,j} + \cdots + w_{i,n} = 1$, $p=1/n$; if $w_{i,j} > p$, we assign this document to $topic_j$, thereby collecting the documents for each topic.

3.2.2 User modeling

3.2.2.1 Author-topic model

The author-topic model uses a topic-based representation to model both document contents and author interests. This model supposes that each author has a topic probability distribution θ, and each topic has a term probability distribution φ, as shown in Figure 2. The model generation process (Steyyers et al. 2004, 307-310) is as follows:

1) Extract the polynomial probability distribution θ for each author;
2) Extract the polynomial probability distribution φ for each topic;
3) For each item in document *d*:
   a) extract an author *x*;
   b) extract a subject *z*;
   c) extract a term *w*;
4) Repeat extraction N_d times to generate document d.

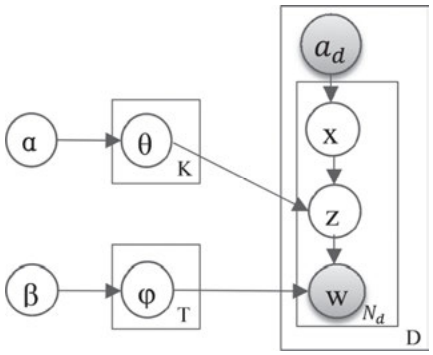


Figure 2. Author-topic model.

In the author-topic model: *θ* represents the author-topic probability distribution; *φ* represents the topic-term probability distribution; *a* is the Dirichlet prior parameter of document-topic probability distribution; *β* is the Dirichlet prior parameter of the topic-term probability distribution; $\underline{a}_d$ represents the uniform distribution of the author's set; *x* represents the author; *z* represents the topic; *w* is the term; *D* represents the document set; N_d represents the number of repeated samples; *K* represents the number of authors; and *T* represents the number of topics.

3.2.2.2 User fusion model

After first using author-topic data to model user interests using posts in five domains, we then used relational data collected on followers and followings to generate follower collection model and following collection model by the author-topic model. Finally, the fusion model is obtained by fusing the three models (user model, follower collection model, and following collection model).

User model of each user is denoted as equation 2:

$$u = \begin{Bmatrix} topic_1 : w_1, \ldots, topic_i : w_i, \ldots, \\ topic_n : w_n \end{Bmatrix} \quad \text{(equation 2)}$$

Where, $w_1 + \cdots + w_i + \cdots + w_n = 1$, *u* represents the user model, $topic_i$ represents topic *i*, and w_i represents the weight of topic *i*.

With reference to Hannon (2010, 201), the follower collection and following collection model are obtained by merging the follower model and the following model separately, which are complemented to the user model by the following equation:

$$U_{all} = \begin{Bmatrix} u_{self}: w_{self}, \mathrm{U}_{followers}: w_{followers}, \\ \mathrm{U}_{followings}: w_{followings} \end{Bmatrix}$$ (equation 3)

Where, $w_{self} + w_{\mathrm{followers}} + w_{followings} = 1$; u_{self} represents the user model; $\mathrm{U}_{followers}$ represents the followers collection model; $\mathrm{U}_{followings}$ represents the followings collection model; w_{self} represents the weight of the user model; $w_{followers}$ represents the weight of the followers collection model; $w_{followings}$ represents the weight of the following collection model; and U_{all} represents the final user model, also called the user fusion model. $U_{followers}$ and $U_{followings}$ is computed by the following formula:

$$\mathrm{U} = \sum_{i=1}^{N} u_i$$ (formula 1)

Where u_i represents user *i*'s own model, and U represents the model results of user collection $\{u_1, \ldots, u_i, \ldots, u_N\}$.

3.2.3 User community detection

3.2.3.1 Label propagation algorithm

The label propagation algorithm (Raghavan et al. 2007) is the classical algorithm for finding communities and is widely used in large-scale networks. The algorithm assumes that a node's label is the one carried by the largest number of its neighbors. Nodes with the same label are grouped into the same community. The steps of the label propagation algorithm are as follows:

1) Initialize the labels of all nodes in the network, and give each node a unique label;
2) Set t=1, with t representing the number of iterations;
3) Randomly arrange nodes in the network, and generate sequence X;
4) According to the order in sequence X, let each node's label be $\arg\max_{\mathrm{N}^1(\mathrm{v})}$, where $\mathrm{N}^1(\mathrm{v})$ represents the set of neighbors with label v;
5) Update labels until each node changes its label to the one carried by the largest number of its neighbors; set t=t+1, and return to the third step.

Through this repeated process, nodes with the same label are grouped into the same community.

3.2.3.2 Community detection

Modularity is a benefit function that measures the quality of a network's division into groups or communities (Newman et al. 2004). The formula is:

$$\mathrm{Q} = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j)$$ (formula 2)

Where A_{ij} represents the adjacency matrix of the network graph; m represents the number of edges of the network graph; and P_{ij} represents expectations of the edge between node *i* and node *j* in an empty model. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. If nodes *i* and *j* are in the same community, $\delta(C_i, C_j) = 1$; otherwise, $\delta(C_i, C_j) = 0$.

3.2.4 Multi-dimensional information association

We use the Jensen–Shannon (JS) divergence distance to calculate the similarity of topics, based on the "topic-term" matrix obtained from the user model and topic model. The formula of the JS divergence distance is as follows:

$$D_{JS}(P||Q) = 1/2(\sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i) + \sum_i \ln\left(\frac{Q(i)}{P(i)}\right)Q(i))$$ (formula 3)

Where $P(i)$ represents the probability of word i in topic P, and $\mathrm{Q}(i)$ represents the probability of word i in topic Q.

We can then calculate the similarity among hierarchy topics, users, and user communities to build a multi-dimensional information organization system.

4.0 Experiment

4.1 Dataset

For this study, we collected 4,966 verified users from Sina Weibo. Microblogs were sourced from Tu et al. (2015). The maximum number of microblogs was set to 500 for each specific user. In total, there are 1,887,633 user post profiles across the five domains, as elaborated in Table 1.

In addition, 152,976 follower and following relationships were crawled. As Table 2 shows, we obtain 106,388 pairs of followers and 110,367 pairs of following, and there are fewer follower–following pairs than the sum of follower and following pairs, which indicates that existing user pairs follow each other.

4.2 Experimental results analysis

4.2.1 UGC topic hierarchy construction

LDA is used to derive the first topic layer using the open source Gibbs sampling tool. The parameters are set as follows (see Section 3.2.1 for parameter definitions): $K = 5$, $a = 50 / K$, and $\beta = 0.01$. Document collections for each

domain	football	internet	law	literature	medicine	Total
posts	409,389	393,187	201,477	655,816	227,764	***1,887,633***

Table 1. Microblogs distribution across different domains.

Relationship type	Follower	Following	Follower + following
pairs	106,388	110,367	152,976

Table 2. Relational data distribution.

User model		Following model		Follower model		Fusion model	
Topic0	0.3527	Topic0	0.3391	Topic0	0.2526	Topic0	0.3300
Topic19	0.1875	Topic19	0.1339	Topic19	0.1378	Topic19	0.1669
Topic47	0.1070	Topic3	0.0560	Topic36	0.0752	Topic47	0.0864
Topic26	0.0970	Topic47	0.0551	Topic47	0.0558	Topic26	0.0772
Topic3	0.0664	Topic10	0.0481	Topic3	0.0527	Topic3	0.0616
Topic18	0.0382	Topic26	0.0480	Topic26	0.0471	Topic18	0.0318
Topic13	0.0259	Topic36	0.0463	Topic10	0.0440	Topic36	0.0261

Table 3. Model results based on user and relationship data

topic are extracted by document extraction method. We set the topic probability threshold to 1 / K (0.2) and get five new document collections. Subsequently, we use LDA to derive the second topic layer based on these five new collections of documents. By repeating the last steps, we derive the third topic layer, thus completing the UGC topic hierarchy structure, in which the first, second, and third hierarchical layers comprise five, twenty-five, and 125 topics, respectively. The Appendix presents the ten most-common terms for each topic, together with their probabilities.

As shown in the Appendix, the first-layer topics are "literature," "law," "medicine," "football," and "internet"; the second-layer topic terms for the first-layer topic "medicine" are "experts," "daily recuperation," "surgical," "female medical treatment," and "diet"; the third-layer topic terms for the second-layer topic of "female medical treatment" are "pregnancy," "diseases of affluence," "treatment," "nursery," and "pregnancy test." The third topic layer is the fine-grained description of the first topic layer, showing the hierarchical relationship and distribution of topics in each level.

4.2.2 User modeling

Each user is modeled based on microblogs and relationships. We applied author-topic modeling (ATM) to formulate the user model using microblogs. The ATM parameters are set as follows (see Section 3.2.2 for parameter definitions): $K = 50$, $\alpha = 50 / K$, and $\beta = 0.01$. We also generate the follower collection model and following collection model for each user by merging user's each follower's posts and following's posts respectively. Note that the follower model and following model are normalized in this study. Finally, the user fusion model is formed by integrating the user, follower, and following models, respectively weighted 0.6, 0.2, and 0.2.

To demonstrate the process, we choose user "178****763" on Sina Weibo, who has a large number of followers. As Table 3 shows, the high-frequency topics of the user model, follower collection model, and following collection model for this individual are "Topic0," "Topic19," "Topic47," "Topic26," and "Topic3." The only difference between the follower model and following model is found in the topics' weights. Compared to the

训练 (Training) Topic0: 加油 (Fighting) 0.0257 比赛 (competition) 0.0241 足球 (Football) 0.0142 兄弟 (brother) 0.0126 训练(training) 0.0112 球迷 (soccer fans) 0.0108	足球赛事 (Football game) Topic19: 足球 (football) 0.0368 比赛 (competition) 0.0205 俱乐部 (club) 0.0089 联赛 (league) 0.0059 冠军 (champion) 0.0053 国安 (Guoan) 0.0049 进球 (goal) 0.0048	人生感悟 (Life thoughts) Topic47: 人生 (life) 0.0219 生活 (living) 0.0207 世界 (world) 0.0133 生命 (live) 0.0085 梦想 (dream) 0.0067 内心 (heart) 0.0058	家庭 (Family) Topic26: 孩子 (child) 0.0090 妈妈 (mother) 0.0061 回家 (go home) 0.0047 儿子 (son) 0.0046 爸爸 (father) 0.0044 女儿 (daughter) 0.0034
工作 (work) Topic3: 同学 (classmate) 0.0064 回家 (go home) 0.0049 心情 (mood) 0.0040 上班 (go to work) 0.0037 吃饭 (eat) 0.0036 公司 (company) 0.0029	公益 (charity) Topic18: 孩子 (child) 0.0351 爱心 (love) 0.0120 父母 (parent) 0.0096 祝福 (blessing) 0.0089 生命 (live) 0.0074 传递 (delivery) 0.0074	生活记录 (life record) Topic36: 馋嘴 (glutton) 0.0140 星座 (constellation) 0.0121 休息 (rest) 0.0113 熊猫 (panda) 0.0061 加油 (fighting) 0.0095 睡觉 (sleep) 0.0074	宗教 (religion) Topic13: 佛教 (Buddhism) 0.0108 菩萨 (buddha) 0.0092 修行 (discipline) 0.0078 众生 (beings) 0.0070 法师 (master) 0.0070 智慧 (wisdom) 0.0065

Table 4. The terms and probabilities of topics presented in Table 3.

user model, the fusion model substitutes "Topic36" for "Topic13." Table 4 details the probability of the terms for each topic in Table 3.

High-frequency topics of the three models are "training," "football game," "life thought," "family," and "work"; in the fusion model, "religion" was replaced by "life record." The high weight of the topics "training" and "football game" indicates that this user likes football or does work related to football; they are also interested in other topics, like "life thoughts," "family," "work," "charity," and "life record,"

4.2.3 Community detection

We detected four communities based on the relationship dataset. Figure 3 shows the population distribution of the communities' members.

As Figure 3 shows, Community 3 has the largest number of members (almost 40% of the total); the other three communities have similarly sized populations, each being around 20%.

As Figure 4 shows, users in the Community 4 are mainly related with the football domain and a few users are related with the other four domains. It indicates that users from different domains are linked. There are significantly more men than women in the "football" community, indicating that male users are more interested in football. Finally, the registration date findings indicate that Weibo's popularity increased significantly in 2010 and 2011, with a high proportion of registrations in both years.

As Figure 5 shows, members of the "football" community come from different provinces. Besides Beijing, Shanghai and Guangzhou, there are many users from other regions like Liaoning, Guangdong, Shandong, Tianjin, Hubei, etc.

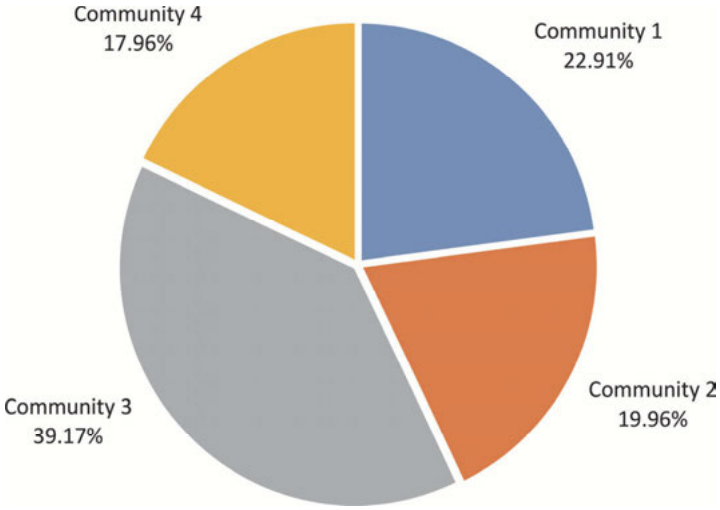


Figure 3. Community population distribution.

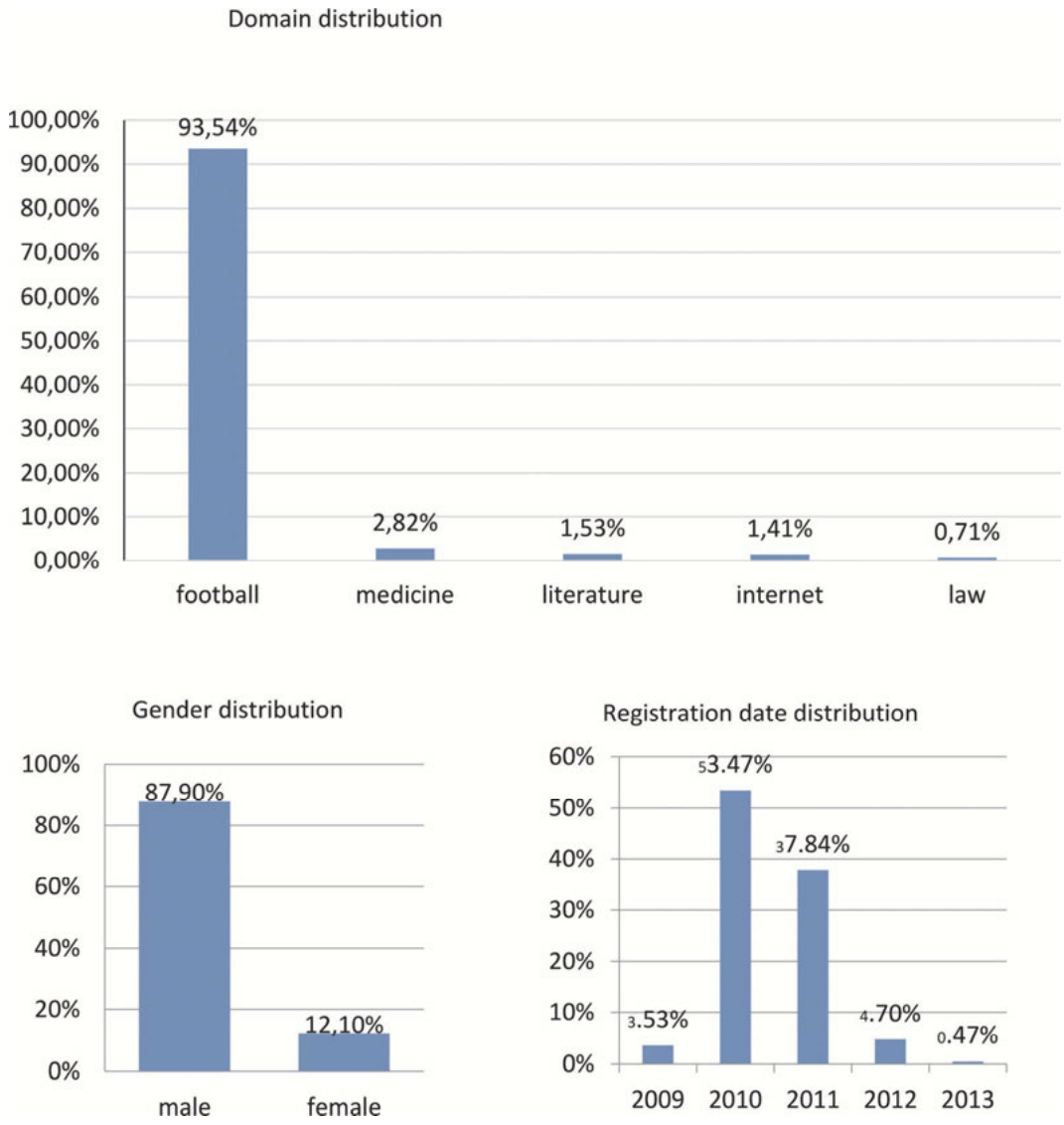


Figure 4. Domain, Gender, and Registration Date Distribution of Community 4.

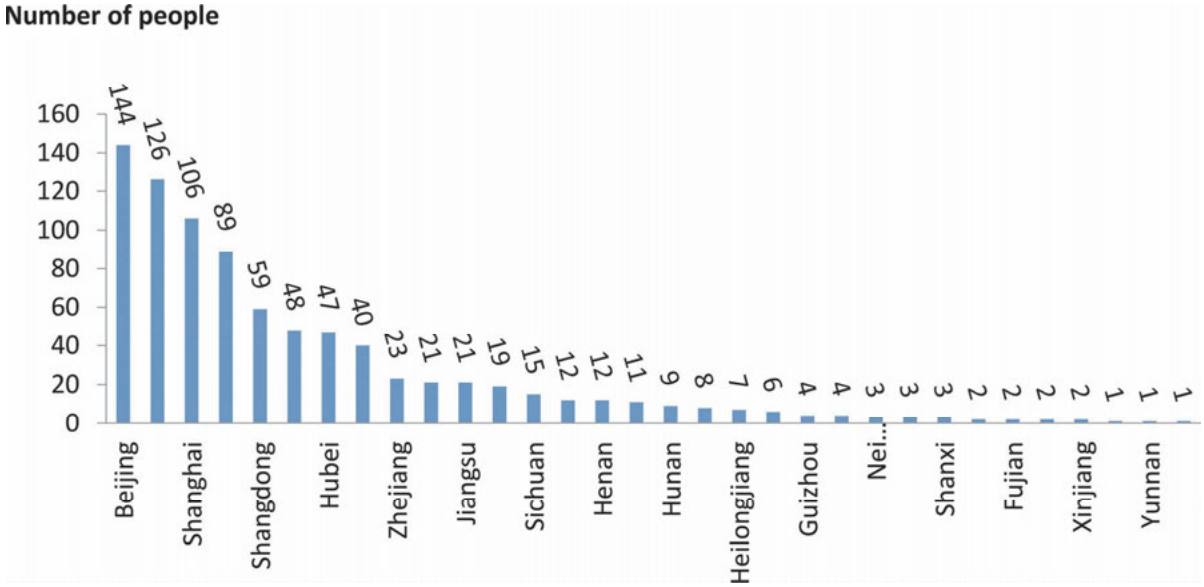


Figure 5. Province Map of Community 4.

4.2.4 Multi-dimensional association

In this section, we will show the relationship between community topics and hierarchical topic for the "football" community. Figure 6 shows the associations among user, community, and topic hierarchy.

User "178****763" in Sina Weibo is assigned to the "football" community from the experiment results, so we can recommend the "football" community and other members to them. Community topic0 is closely related to first-layer topic3, which is itself closely related to second-layer topic4, which is, in turn, closely related to third-layer topic1.

Table 5 shows the term distribution of topic0 ("football"), topic4 ("football game"), and topic1 ("domestic match"). These topics are closely related to topic0 ("football training").

For fine-grained detail on the community's interest topics, we present the third-layer topic distribution of second-layer topic 4 in Table 6.

As Table 6 shows, the third-layer topics of second-layer topic4 are "international match," "domestic match," "player training," "football show," and "person." We can thus recommend topics in a more refined and accurate way based on the associated community and topic hierarchy. Here, we can recommend these third-layer topics for the "football"

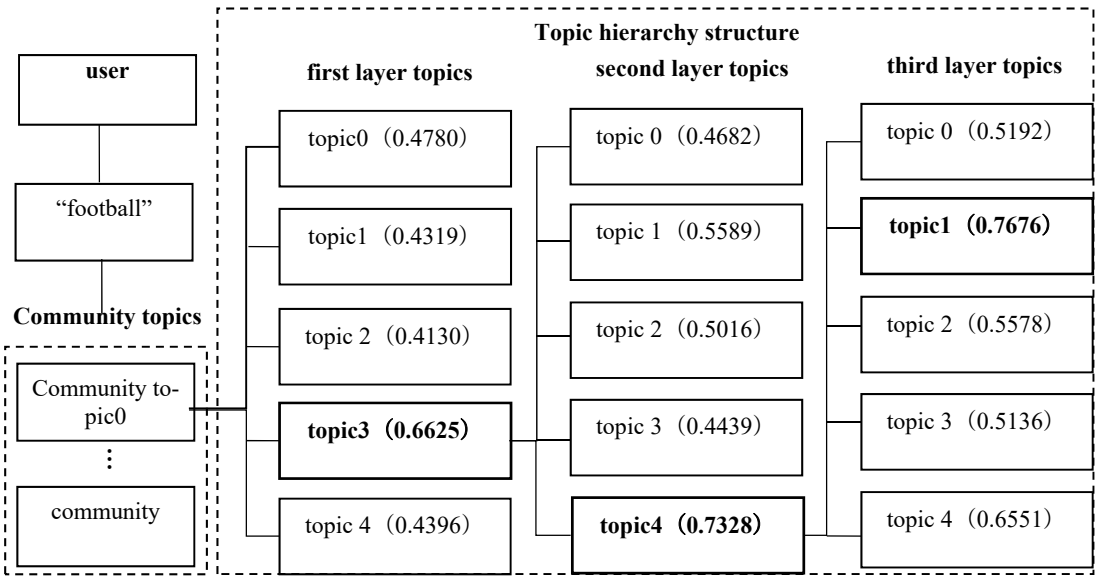


Figure 6. Associations among user, community, and topic hierarchy in the "football" community.

First-layer topic-Topic3 “足球” (football)	Second-layer topic-Topic4 “足球赛事” (football game)	Third-layer topic-Topic1 “国内赛事” (domestic match)
加油 (fighting) 0.0053	足球 (football) 0.0196	比赛 (competition) 0.0369
足球 (football) 0.0051	比赛 (competition) 0.0191	球迷 (football fan) 0.0211
比赛 (competition) 0.0050	加油 (fighting) 0.0161	足球 (football) 0.0191
回家 (go home) 0.0031	球迷 (football fan) 0.0088	球队 (football team) 0.0102
妈妈 (mother) 0.0027	俱乐部 (football club) 0.0054	广州 (Guangzhou) 0.0073
睡觉 (sleep) 0.0025	球队 (football team) 0.0054	联赛 (league) 0.0069
孩子 (child) 0.0025	训练 (train) 0.0053	北京 (Beijing) 0.0060
球迷 (football fan) 0.0023	球员 (footballer) 0.0048	青岛 (Qingdao) 0.0054
爱心 (love) 0.0019	兄弟 (brother) 0.0047	主场 (home court) 0.0049
兄弟 (brother) 0.0019	体育 (sports) 0.0042	上海 (Shanghai) 0.0048

Table 5. Relevant topics of community topic0.

Topic0 **“世界赛事”** **(international match)**	**Topic1** **“国内赛事”** **(domestic match)**	**Topic2** **“球员训练”** **(player training)**	**Topic3** **“足球节目”** **(football show)**	**Topic4** **“风云人物”** **(person)**
比赛 (competition) 0.0167	比赛 (competition) 0.0369	足球 (football) 0.0288	体育 (sports) 0.0108	加油 (fighting) 0.0122
球员 (footballer) 0.0085	球迷 (football fan) 0.0211	教练 (coach) 0.0060	足球 (football) 0.0080	国安 (Guoan) 0.0059
进球 (goal) 0.0076	足球 (football) 0.0191	孩子 (child) 0.0058	比赛 (competition) 0.0059	足球 (football) 0.0032
球队 (football team) 0.0075	球队 (football team) 0.0102	训练 (train) 0.0048	上海 (Shanghai) 0.0046	女足 (women’s soccer) 0.0025
西班牙 (span) 0.0061	广州 (Guangzhou) 0.0073	运动 (sport) 0.0039	球迷 (football fan) 0.0045	青岛 (Qingdao) 0.0023
冠军 (champion) 0.0055	联赛 (league) 0.0069	球员 (footballer) 0.0038	直播 (broadcasting) 0.0041	王永珀(Wang Yongbo) 0.0022
皇马 (Real Madrid) 0.0054	北京 (Beijing) 0.0060	俱乐部 (football club) 0.0036	参加 (participate) 0.0038	李 帅(Li Shuai) 818 0.0017
巴萨 (Bass) 0.0049	青岛 (Qingdao) 0.0054	比赛 (competition) 0.0035	俱乐部 (football club) 0.0037	王晓龙 (Wang Xiaolong) 0.0016
意大利 (Italy) 0.0048	主场 (home court) 0.0049	运动员 (athlete) 0.0032	现场 (scene) 0.0031	邵佳一(Shao Jiayi) 0.0016
决赛 (final) 0.0047	上海 (Shanghai) 0.0048	体育 (sports) 0.0032	节目 (show) 0.0028	徐云龙(Xu Yunlong) 0.0015

Table 6. The third-layer topic distribution of second-layer Topic4.

community. In addition, other second-layer topics can be recommended to the community to help users access more relevant information. In short, users can easily obtain a comprehensive, in-depth picture of their topics of interest.

4.2.5 Comparisons between different organization patterns

There are three basic elements for information organization in this paper: topic hierarchy, user, and community. As Table 7 shows, we analyze different combinations of the main elements with respect to their information organization methods and associated advantages and disadvantages. The table also provides examples of relevant social media platforms.

As described in Table 7, for those organization models using only one element (all elements are user, hierarchical topic, and community), it is more costly for users to access information from another two sources. For those models using two elements, shortages might exist when information organization is trying to build on the missing element. For instance, user interaction is not sufficiently

Basic Elements	Organization Model	Advantages	Disadvantages	Social Media Cases
User	Users in a friend relationship on social media can send all types of messages and be fully connected.	Information is usually shared through a one-to-one connection and with known friends, which guarantees privacy.	Obtained information resources are limited to the range of friends.	Communication software e.g., Tencent QQ, Messenger
Hierarchical Topic	Information content is based on entries and is organized in a hierarchical manner according to a certain classification system or topic.	Information organized in a hierarchical topic structure can help users to quickly search for and find needed knowledge.	Classification system is not clear enough; the data need to be maintained and updated in real time.	Internet encyclopedia projects e.g., Wikipedia, Baidu Encyclopedia
Community	Users with the same information needs are integrated in the same virtual space and communicate with one another within this community.	Users within the same community can quickly share information.	Accessible information resources are limited by community themes.	Online communities, e.g., Google Groups
User + Hierarchical Topic	Users who seek specific information will consult their friends or look for relevant information based on the topic.	Interaction between users and information acquisition can be efficient.	User interaction is not sufficiently strong.	Socialized question and answer platforms, e.g., Zhihu, Stack Overflow
User + Community	Users make friends gradually and communities are basically groups of users with a common interest or goal, who are also highly likely to be friends.	Convenient for users to effectively achieve their social goals and maintain personal connections.	Community construction is affected by users' social activities over social media.	Business social platforms, e.g., LinkedIn, Dajie Network
Community + Hierarchical Topic	Users within each community have highly personalized interactions; different communities have different topics that can be organized in a hierarchical structure.	Users in the same community can share information in real time and find corresponding sub-communities according to their own information needs.	Information sharing within social media communities is mainly in the form of text and pictures.	Web forums e.g., Tianya
User + Hierarchical Topic + Community	Users can obtain information through friendships, communities to which they belong, and hierarchical topics on social media.	Users can engage in social behaviors, access rich information sources, and obtain information effectively from different sources.	There might be information security risks.	Social networking platforms, e.g., Facebook, Douban

Table 7. Comparisons between different information organization patterns.

strong when community information is not considered. Therefore, the most efficient solution is to incorporate all three elements in the information organization model, enabling users to obtain information via friends, communities to which they belong, and hierarchical topics on social media.

5.0 Conclusion

We have described our investigation into how to organize information on social networks in a user-centered way to meet personalized needs. Our proposed method linked UGCs, users, and user communities in multi-dimensional framework.

First, we constructed a three-layer topic hierarchy based on UGCs. Second, we developed a user interests model using UGCs and relationship data, fusing the user, follower, and following models. Third, inspired by previous work on community detection, we proposed a new approach for detecting the topics of each community. Finally, we derived a multi-dimensional information organization pattern through similarities among three dimensions: the UGC topic hierarchy, user interest model, and user communities.

Our results show that the topic hierarchy is effective in providing supplementary and recommended information. Problems of spare data in user modeling can be partly solved by integrating it with the relational model. We can also help users find communities of interest using community detection. As the user modeling in this paper is not evaluated, we propose to conduct a follow-up study, in which the evaluation will involve both expert scoring and user assessment, such as user satisfaction of information recommended through platforms.

References

Agichtein, Eugene, Carlos Castillo, Debora Donato, Aristides Gionis and Gilad Mishne. 2008. "Finding High-quality Content in Social Media." In *WSDM'08: Proceedings of the 2008 international conference on web search and data mining. February 11-12, 2008, California, USA*. New York, NY: Association for Computing Machinery, 183-194. Doi:10.1145/1341531.1341557

Bar-Ilan, Judit and Yifat Belous. 2007. "Children as Architects of Web Directories: an Exploratory Study." *Journal of the American Society for Information Science & Technology 58*: 895–907.

Bawden, David. 2006. "Users, User Studies and Human Information Behaviour." *Journal of Documentation* 62: 671-679.

Benevenuto, Fabrício, Tiago Rodrigues, Meeyoung Cha and Virgílio Almeida. 2009. "Characterizing User Behavior in Online Social Networks." In IMC'09: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. November 4-6, 2009, Chicago, Illinois*. New York, NY: Association for Computing Machinery, 49-62. doi:10.1145/1644893.1644900

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *J Machine Learning Research Archive 3*: 993-1022

Bu, Shuqing, Huamei Liu and Guangping Wang. 2010. "国外近几年网络环境下知识组织理论、方法的深化与拓展." [A Summary of Recent Research on Knowledge Organization.] 中国索引 [*Journal of the China Society of Indexers]* 1:2-12.

Chen, Sherry Y. and Nigel J. Ford. 1998. "Modelling User Navigation Nehaviours in a Hypermedia-based Learning System: An Individual Differences Approach." *Knowledge organization* 25: 67-78.

Chen, Enhong, Yanggang Lin, Hui Xiong, Qiming Luo and Haiping Ma. 2011. "Exploiting Probabilistic Topic Models to Improve Text Categorization under Class Imbalance." *Information Processing & Management* 47: 202-214.

Gao, Xia and Jiancheng Guan. 2012. "Network Model of Knowledge Diffusion." *Scientometrics* 90: 749-62.

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

Hannon, John, Mike Bennett and Barry Smyth. 2010. "Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches." In RecSys'10: *Proceedings of the fourth ACM conference on Recommender systems. September 26-30, 2010, Barcelona, Spain.* New York, NY: Association for Computing Machinery, 199-206. doi:10.1145/1864708.1864746

Hjørland, Birger. 2003. "Fundamentals of Knowledge Organization." *Knowledge organization* 30: 87-111.

Hjørland, Birger. 2014. "User-based and Cognitive Approaches to Knowledge Organization: A Theoretical Analysis of the Research Literature." *Knowledge organization* 40: 11-27.

Kaplan, Andreas M. and Michael Haenlein. 2010. "Users of the World, Unite! The Challenges and Opportunities of Social Media." *Business Horizons* 53: 59-68.

Keenan, Andrew and Ali Shiri. 2013 "Sociability and Social Interaction on Social Networking Websites." *Library Review* 58: 438-450.

Kietzmann, Jan H., Kristopher Hermkens, Ian P. McCarthy, Bruno S. Silvestre. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons* 54: 241-251.

Kim, Hak Lae, Simon Scerri, John G. Breslin, Stefan Decker and Hong Gee Kim. 2008. "The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies." In DC2008*: Proceedings of the*

International Conference on Dublin Core and Metadata Applications 2008. September 22-26, 2008, Berlin, Germany. Singapore: Dublin Core Metadata Initiative, 128-137.

Kwak, Haewoon, Changhyun Lee, Hosung Park and Sue Moon. 2010. "What is Twitter, a Social Network or a News Media?" In WWW2010: *Proceedings of the 19th international conference on World Wide Web. April 26-30, 2010, Raleigh, USA.* New York, NY: Association for Computing Machinery, 591-600. doi: 10.1145/1772690.1772751

Li, Jinhai, Changlin Mei and Yuejin Lv. 2013. "Incomplete Decision Contexts: Approximate Concept Construction, Rule Acquisition and Knowledge Reduction." *International Journal of Approximate Reasoning* 54: 149-165.

Maia, Marcelo, Jussara Almeida and Virgílio Almeida. 2008. "Identifying User Behavior in Online Social Networks." In SocialNets'08: *Proceedings of the 1st Workshop on Social Network Systems. April 1, 2008, Glasgow, Scotland, UK.* New York, NY: Association for Computing Machinery, *1-6*.

Mathes, Adam. 2004. "Folksonomies: Cooperative Classification and Communication through Shared Metadata." http://www.adammathes.com/academic/computer -mediated-communication/folksonomies.html.

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

Newman, Mark EJ and Michelle Girvan. 2004. "Finding and Evaluating Community Structure in Networks." *Physical review E 69*: 026113.

Noruzi, Alireza. 2006. "Folksonomies: (Un) controlled Vocabulary?" *Knowledge organization* 33: 199-203.

Potnis, Devendra. 2011. "Folksonomy-based User-centric Information Organization Systems." *International Journal of Information Studies* 3: 31-43.

Qiang, Bi, and Wang Yu. 2013. "Fronts and Hotspots of the Application Research on Folksonomy Abroad." (In Chinese) *Data Analysis and Knowledge Discovery* 29: 36-42.

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. 2007. "Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks." *Physical Review E Statistical Nonlinear & Soft Matter Physics 76*: 036106.

Steyvers, Mark, Padhraic Smyth, Michal Rosen-Zvi and Thomas Griffiths. 2004. "Probabilistic Author-topic Models for Information Discovery." In KDD'04: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* August 22-25, 2004, Seattle, Washington, USA. New York, NY: Association for Computing Machinery, 306-315. doi:10.1145/1014052.1014087

Stutzman, Frederic. 2006. "An Evaluation of Identity-sharing Behavior in Social Network Communities." *International Journal of Performance Arts & Digital Media* 3: 10-18.

Treem, Jeffrey W. and Paul M. Leonardi. 2012. "Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association." *Social Science Electronic Publishing* 36: 143-189.

Tu, Cunchao, Zhiyuan Liu, Huanbo Luan and Maosong Sun. 2015. "PRISM: Profession Identification in Social Media." ACM Transactions on Intelligent Systems and Technology 8, no. 6: 1-16.

Van Damme, Céline, Martin Hepp, and Katharina Siorpaes. 2007. "Folksontology: An Integrated Approach for Turning Folksonomies into Ontologies." Paper presented at Bridging the Gap between Semantic Web and Web 2 .0 (SemNet 2007), at the 4th European Semantic Web Conference. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.379.5516

Ming, Zhao Yan, Jintao Ye, and Tat Seng Chua. 2014. "A Dynamic Reconstruction Approach to Topic Summarization of User-generated-content." In CIKM '14: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. November 03-07, 2014, Shanghai, China.* New York, NY: Association for Computing Machinery, 311-320. doi:10.1631/FITEE.1500402

Zhang, Wei, Jia-yu Zhuang, Xi Yong, Jian-kou Li, Wei Chen and Zhe-min Li. 2017. "Personalized Topic Modeling for Recommending User-generated Content." *Frontiers of information technology & electronic engineering* 18: 708-718.

Zhou, Xiaoying. 2010. "知识链接的发展阶段、发展动因和类型特征分析." [Studies on Development Phases, Development Motivation, Type and Characteristic of knowledge Linkage.] 图书情报工作 [*Library and Information Service*] 54: 36-40.

Zhu, Xingwei, Zhao-Yan Ming, Zhao-Yan Ming, Zhao-Yan Ming. 2013. "Topic Hierarchy Construction for the Organization of Multi-source User Generated Contents." In SIGIR'13: P*roceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. July 28 - August 01, 2013, Dublin, Ireland.* New York, NY: Association for Computing Machinery, 233-242. doi: 10.1145/2484028.2484032

Zhu, Xingwei, Zhao-Yan Ming, Yu Hao, Xiaoyan Zhu, Tat-Seng Chua. 2014. "Customized Organization of Social Media Contents Using Focused Topic Hierarchy." In CIKM '14: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. November 03-07, 2014, Shanghai, China.* New York, NY: Association for Computing Machinery, 1509-1518. doi:10.1145/2661829.2661896

Appendix: The sample of UGC topic hierarchy construction