An Approach to the Construction of Personalized Knowledge Map Based on Collaborative Tagging

Ming Li, Mengyue Yuan, and Haitao Xiong

***Abstract*—Knowledge is the strategic resource for the organization. Knowledge map is an important tool for knowledge sharing. Providing the personalized knowledge map based on the preference of users can ease the burden of learning the knowledge map and facilitate the finding of the required knowledge. The tagging to documents reflects the user's preference of classification. In the paper, the approach to the personalized knowledge map construction based on the collaborative tagging is proposed. Firstly, the weight of the tag in documents is identified. Secondly, the similarity of users on the preference of classification is defined and then users that have the similar classification preference are identified to expand the current user's preference of personalized classification. Then the text vector of document and text similarity is identified. Afterwards, the knowledge is clustered according to both the personalized classification similarity and text similarity. Finally, the topics of each cluster are identified. In the topic identification, both the weight of the term in the text and the weight of the term in the tags are integrated. The experiment shows that proposed method is feasible and performs well.**

***Index Terms*—Knowledge map, knowledge management systems, personalized knowledge map.**

I. Introduction

Knowledge is an important asset for an organization [1]-[3]. It improves the ability of the organization to response to new situations [4], [5]. With the development of globalization and the rapid changing of the competitive environment, finding and mastering the knowledge rapidly can make the decision more rational and improve the core competitive capability [6]. More and more enterprises implement the knowledge management to management the value knowledge both inside and outside organizations.

There are two kinds of knowledge which are implicit knowledge and explicit knowledge [7]-[9]. The explicit knowledge refers to the kind of knowledge that can be expressed and codified such as the rules, theorem and law. It often exists in manual, reports and documents [10]. On the contrary, the knowledge that cannot be codified is called implicit knowledge such as experience. It often exists in the owner's brain [11], [12]. Since explicit knowledge can be transformed into an electronic format more easily than implicit knowledge, it is often a core part in the implementation of knowledge management. With the accumulation of the explicit knowledge, the knowledge explosion occurs and it is more and more difficult to find the appropriate knowledge for the user [13], [14]. Therefore, there need an efficient way to find the required knowledge more quickly and easily. Knowledge map is the commonly used tool to facilitate the retrieving and understanding of the explicit knowledge [15]. It organizes the scattered knowledge and presents them in meaningful categorizations. Users can look for the knowledge by categories through the knowledge map. It facilitates the finding of knowledge and makes the knowledge finding more easily [16].

The construction of the knowledge map attracts many researchers and many important processes have been made [17]-[20]. For example, the genetic algorithm, information retrieval, and multi-dimension scaling method are integrated to construct topic knowledge maps [17]. Both information retrieval technique and data mining technique are used to developed knowledge map [18]. An improved interface combining a 1D alphabetical hierarchical list and a 2D Self-Organizing Map island display are integrated to construct the knowledge map [19]. The classification preferences of each person are not identical. Because of the different background knowledge, they always have their own classification preferences. Providing the personalized knowledge map can ease the burden to learn the knowledge map and facilitate the finding of required knowledge. The personalized knowledge map construction method in knowledge management systems is proposed [20]. However, in the method, the number of documents that are tagged by the single person is limited and the weight of tag is not discriminated in the study. In order to resolve the problems and provide the personalized knowledge map for the user, we proposed the personalized knowledge map construction approach. Firstly, we define the weight of the tags. The classification similarity between documents is given. Then the users that have the similar preferences are found to extend the preferences. Afterwards, the text similarity between documents is defined. After that, the knowledge is clustered based on the text similarity and personalized classification similarity. Finally the topic of each cluster is identified.

The rest of the paper is organized as following. In the following section, the research on the knowledge map and text clustering are discussed. In the third section, the approach to the personalized knowledge map construction is given. In the fourth section, the experiment and the results are provided. Finally, conclusions are provided.

Manuscript received June 2, 2015; revised August 24, 2015. The research is supported by the National Natural Science Foundation of China under Grant No. 71101153, 71201004, and Science Foundation of China University of Petroleum, Beijing (No. 2462015YQ0722), Humanity and Social Science Youth Foundation of Ministry of Education in China (Project No. 13YJC790112).

Ming Li and Mengyue Yuan are with School of Business Administration, China University of Petroleum, China (e-mail: brightliming@outlook.com).

Haitao Xiong is with School of Computer and Information Engineering, Beijing Technology and Business University, China.

II. Literal Review

A. Knowledge Map

Knowledge map refer to the organization of lots of documents based on the classification characteristics of documents. In the knowledge map, the documents or the information of the documents are organized by categories. Users can find the knowledge or the place of the knowledge by browsing categories. It is the popular used tool for the knowledge management and attracts the attention of many researchers.

For example, the knowledge map is used to represent organizational knowledge [21]. In the study, a practical six step method for capturing and representing the knowledge in organizations is given. The case study in a manufacturing company is given to illustrate the proposed method. For the study of knowledge map in tourist destinations, in the work [22], the author compares knowledge maps of four destination types which are city, mountain, historic and island resort tourism. He advises that public sector should be involved in the construction of the tourist destination knowledge depository. In the work proposed in [23], the method to develop the workflow based knowledge map is given. In the constructed process-perspective knowledge map, both the structure of processes and tasks defined in workflow are used. In the study the prototype is developed and applied in the automobile industry. Aiming to organize the knowledge on P2P networks, the visualized cognitive knowledge map method is proposed [24]. In the method, the self-organized map is extended to merge the other peers' documents visually. In order to construct the knowledge map for construction industry, the knowledge map model is constructed to build knowledge map, which includes five steps in the model [25]. The first step is to identify problems. The second step is to discuss with experts and users. After that, classification structure is established. Then the document base is constructed. Finally, the display model is determined. Moreover, in the study, the knowledge map model system is developed to assist the knowledge map reused and shared in the practical processes.

B. Text Clustering

With the exponential growth of knowledge, how to organize text data efficiently and effectively arises as an important problem. Text clustering is an important tool to solve the problem. Text clustering means the clustering of documents based on the predefined similarity metrics without predefined categories [26]. The common characteristic of text classification and text clustering is that the documents are organized into categories. The main different is whether there are the predefined categories [27], [28]. KNN is a simple but effective method for text categorization and clustering. It has been popular used in the text clustering and classification [29]-[31].

In text clustering, it is based on the assumption that the similarity degree of documents in the same cluster is the highest, while in different clusters to the lowest. In the text clustering, the document is selected as the centroid of the cluster. Then the other documents are classified into clusters according to the similarity between the document and the cluster. The document is classified into the cluster that has the highest similarity. Then the representation of the cluster is updated.

The detailed processes of the document clustering based on the KNN include the following steps [32].

1) Create a new empty cluster and read a document as the centroid of the cluster.
2) If there are no documents left in the document collection, go to (4), otherwise read a new document and calculate the similarity between the new document and all the clusters.
3) The new document is classified into the cluster that has the highest similarity and the value of the similarity is larger than the predefined new document. Then the cluster is updated. Otherwise, go to (1).
4) Stop the clustering and the created cluster are the clustering results.

III. The Approach to the Construction of Personalized Knowledge Map

The classification preferences are not identical for users. Each user has his personalized classification preference. The tagging reflects the user's personalized classification preferences. The documents that the user considered belonging to the same categorize will be tagged the same tag. The tagging can be used to get the user's personalized classification preferences. The documents that have the same tag represent the user think they are relevant. In fact, the number of documents that one user tagged is limited because of the limit of knowledge scope. Each user can only classify part of documents. The other users that have the similar classification preference can be used to expand the classified documents of the current user.

Therefore, we need to find similar users. After that, the knowledge can be clustered. In the clustering, not only the textual similarity but also the personalized classification similarities are used. The detail steps of the approach are given as follows.

Step 1. Determination of the weight of tags

The tags on documents have different weigh for the classification of the document. The less tag the document has, the more important of the tag for the document. The tag which less documents uses represents the documents that have the tag have the special similar characteristic of the classification. In the determination of the weight, based on the idea TFIDF [33], the weight w_{ij} of the tag t_i for document $d_j \in D$ can be calculated by

$$w_{ij} = 1/NT_j \times \log(NAD/NTD_i) \quad (1)$$

where, NT_j means the number of tags that the document d_j has, NAD denotes number of documents in the collection D, NTD_i represents the number of documents that have the tag t_i.

Step 2. Calculation of the personalized classification similarity between documents

The personalized classification similarity between the documents $d_i, d_j \in D$ is determined by the same tags along with the weight that the documents have. Since the same tag may mean differently for different users. In the step, same tags refer to the tags given by the same user.

Therefore, the personalized classification similarity between documents $d_i, d_j \in D$ can be derived by

$$sim_c\left(d_i, d_j\right) = \sum_{t_k \in T_i \cap T_j} \left(1 - \left|w_{ki} - w_{kj}\right|\right) \quad (2)$$

$$d_i, d_j \in D$$

where, T_i and T_j represent the tags that the document d_i and d_j have respectively, w_{ki} and w_{kj} represent the weight of tags t_k for document d_i and d_j.

Step 3. Deriving the similarity of users in the personalized classification of documents

We use the term vector to represent the personalized classification characteristics of users. Each element in the vector is the personalized classification information between the two documents. The classification information is represented by the weight of classification similarity between two documents. That is, in the calculation of the similarity between documents, we just take the personalized classification information into consideration but not whether the two tags are the same between users.

Therefore, the similarity between the user u_x and user u_y can be derived by

$$sim_c(u_x, u_y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(sim_c^x\left(d_i, d_j\right) \times sim_c^y\left(d_i, d_j\right)\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} sim_c^x\left(d_i, d_j\right)^2} \times \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} sim_c^y\left(d_i, d_j\right)^2}} \quad (3)$$

where, n is the number of documents in the document collection, d_i, d_j represents the two documents in the document collection.

In the equation, we can see that the both belongings to the same category and not belonging to the same category are considered. That is, besides the two documents that are deemed relevant by both users, the two documents that are deemed irrelevant by both users also reflects the personalized classification similarity.

Step 4. Deriving the weight of terms in documents

We use the term vector to represent the document. In the vector, each element is the term in the document along with the weight. The weight is calculated by the TFIDF method [33].

In the method, the terms that exist more in the document and exits less in the other documents represents the terms are important to the documents and the higher weight needs to be got.

The weight w_{ij} of the term W_i for document d_j can be defined as [33]

$$w_{ij} = \frac{F_{ij}}{\max\limits_{W_i \in d_j}(F_{ij})} \times \log\frac{N_D}{m_i} \quad (4)$$

where, F_{ij} means the occurrence times of the term W_i in the document d_j, m_i represents the number of documents that contains the term W_i, N_D is the number of documents in the collection D.

Step 5. Deriving the text similarity between documents

The documents that terms in which has the similar weights are considered similar.

So, the text similarity between documents d_i, d_j can be derived by

$$sim_t\left(d_i, d_j\right) = \sum_{k=1}^{P}\left(w_{ki} \times w_{kj}\right) \Big/ \sqrt{\sum_{k=1}^{P}\left(w_{ki}\right)^2} \times \sqrt{\sum_{k=1}^{P}\left(w_{kj}\right)^2} \quad (5)$$

where, w_{ki} and w_{kj} represent the weights of term k for the document d_i and d_j, P represents the number of terms.

Step 5. Personalized document clustering

The KNN method [32] is extended to cluster the documents to construct the personalized knowledge map. The key steps in the KNN method are the determination of the similarity between the cluster and the document.

In the calculation of the similarity, both the personalized classification similarity and the text similarity are considered simultaneously.

The similarity between the document d_i and the cluster C_j can be derived by

$$\begin{aligned} sim(d_i, C_j) &= \lambda sim_c\left(d_i, C_j\right) + (1-\lambda) sim_t\left(d_i, C_j\right) \\ &= \lambda \frac{\sum_{d_k \in C_j} sim_c\left(d_i, d_k\right) \times sim_c\left(u_x, u_y\right)}{N_{C_j}} \\ &+ (1-\lambda) \frac{\sum_{k=1}^{P}\left(w_{ki} \times v_{kj}\right)}{\sqrt{\sum_{k=1}^{P}\left(w_{ki}\right)^2} \times \sqrt{\sum_{k=1}^{P}\left(v_{kj}\right)^2}} \end{aligned} \quad (6)$$

where, C_j represents the cluster j, d_k represents the document k in the cluster C_j, N_{C_j} represent the number of documents in the cluster C_j, $\lambda \in [0,1]$ is the adjusting parameter of the degree of personalization. If the user prefer more on the personalization, then the λ can be given a larger value. If the two documents have the same tag that are given by the current user, the similarity between users is set value one. If the two documents that do not have the same tag given by the current user, the similarity between the user and the other user is used.

In the updating of each cluster, only the text vector needs to be updated. In the calculation of the similarity on the personalized classification, the similarity is derived by the

similarity between the new document and each document in the cluster.

Step 6. Identification of topic words in each cluster

In the identification of topic words, not only the weight of the term in the tags but also theweight in the cluster is integrated.

By extending the idea of TFIDF [33], the weight w_{ij}^T of the term i in the cluster j can be derived by the integration of the weight w_{ij}^w in the cluster and the weight w_{ij}^t in the tags, which is shown as

$$\begin{aligned} w_{ij}^T &= \gamma w_{ij}^w + (1-\gamma) w_{ij}^t \\ &= \gamma \frac{F_{ij}}{\max_{W_t \in C_j}(F_{tj})} \times \log \frac{N_C}{m_i} + (1-\gamma) \frac{G_{ij}}{\max_{W_t \in C_j}(G_{tj})} \times \log \frac{N_C}{g_i} \end{aligned} \quad (7)$$

where, F_{ij} is the occurrence times of term i in the text of cluster j, N_C is the number of clusters, m_i is the number of clusters that have the term i in the text, G_{ij} is the occurrence times of term i in tags of the cluster j, N_C is the number of clusters, g_i is the number of clusters that have the term i in the tags, γ is the adjusting parameter of the importance of the personalization.

IV. Experiment

In the experiment, there are about 174 documents are collected. They are classified into four categories manually and parts of document are tagged. Then we compare the accuracy between text classification and the personalized classification. Firstly, we use the KNN method to cluster the documents and construct the knowledge map. Then, we use the proposed method to cluster the documents and construct the personalized knowledge map. The accuracy is defined as the ratio of number of the correct classified documents to the number of documents in the category. In the four categories, the max improvement of accuracy is 14.03%, and the average of the improvement of accuracy is 5.75%. From the experiment results, we see that the proposed method is feasible and performance well. It makes the knowledge map be fitter for the personalized classification preference of users.

V. Conclusions

In the paper, the approach to the construction of personalized knowledge map is proposed. The importance weights of tag for the different classification of documents are not the same. So firstly, the weight of the tag is identified. Then the classification similarity between documents is proposed. Considering the documents that one user tagged is limited, the users that have the similar preference of classification are found. After that the text similarity between documents are identified. Then traditional KNN method is extended to construction the personalized knowledge map by the integration of both the text similarity and personalized classification similarity. Moreover, the topic words of each cluster are identified. The experimental results show the proposed method performs better than the traditional method in the knowledge map construction.

References

[1] T. H. Davenport, D. W. D. Long, and M. C. Beers, "Successful knowledge management projects," *Sloan Management Review*, vol. 39, no. 2, pp. 43-57, 1998.

[2] T. A. Stewart, "The wealth of knowledge: Intellectual capital and the twenty-first century organization," *Crown Business*, 2007.

[3] E. Civi, "Knowledge management as a competitive asset: A review," *Marketing Intelligence & Planning*, vol. 18, no. 4, pp. 166-174, 2000.

[4] A. Carneiro, "How does knowledge management influence innovation and competitiveness?" *Journal of Knowledge Management*, vol. 4, no. 2, pp. 87-98, 2000.

[5] J. P. Liebeskind, "Knowledge, strategy, and the theory of the firm," *Strategic Management Journal*, vol. 17, no. S2, pp. 93-107, 1996.

[6] R. Sharkie, "Knowledge creation and its place in the development of sustainable competitive advantage," *Journal of Knowledge Management*, vol. 7, no. 1, pp. 20-31, 2003.

[7] Q. Huang, R. M. Davison, and J. Gu, "The impact of trust, guanxi orientation and face on the intention of Chinese employees and managers to engage in peer−to−peer tacit and explicit knowledge sharing," *Information Systems Journal*, vol. 21, no. 6, pp. 557-577, 2011.

[8] R. S. Masters, "Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure," *British Journal of Psychology*, vol. 83, no. 3, pp. 343-358, 1992.

[9] E. A. Smith, "The role of tacit and explicit knowledge in the workplace," *Journal of knowledge Management*, vol. 5, no. 4, pp. 311-321, 2001.

[10] Y. S. Hau, B. Kim, H. Lee, and Y. G. Kim, "The effects of individual motivations and social capital on employees' tacit and explicit knowledge sharing intentions," *International Journal of Information Management*, vol. 33, no. 2, pp. 356-366, 2013.

[11] P. M. Leonardi and D. Bailey, "Transformational technologies and the creation of new work practices: Making implicit knowledge explicit in task-based offshoring," *MIS Quarterly*, vol. 32, no. 2, pp. 159-176, 2008.

[12] J. H. Woo, M. J. Clayton, R. E. Johnson, B. E. Flores, and C. Ellis, "Dynamic knowledge map: reusing experts' tacit knowledge in the AEC industry," *Automation in Construction*, vol. 13, no. 2, pp. 203-207, 2004.

[13] R. Dove, "Knowledge management, response ability, and the agile enterprise," *Journal of knowledge management*, vol. 3, no. 1, pp. 18-35, 1999.

[14] G. D. Bhatt, "Knowledge management in organizations: Examining the interaction between technologies, techniques, and people," *Journal of Knowledge Management*, vol. 5, no. 1, pp. 68-75, 2001.

[15] J. Zuhua, S. Hai, and H. Yongwen, "Knowledge map and knowledge management tools to support distributed product design," *Advances in Automation and Robotics*, vol. 1, pp. 647-654, 2012.

[16] M. J. Eppler, "Making knowledge visible through intranet knowledge maps: concepts, elements, cases," in *Proc. the 34th Annual Hawaii International Conference on System Sciences*, 2001, p. 9.

[17] D. Y. Chiu and Y. C. Pan, "Topic knowledge map and knowledge structure constructions with genetic algorithm, information retrieval, and multi-dimension scaling method," *Knowledge-Based Systems*, vol. 67, pp. 412-428, 2014.

[18] F. R. Lin and C. M. Hsueh, "Knowledge map creation and maintenance for virtual communities of practice," *Information Processing & Management*, vol. 42, no. 2, pp. 551-568, 2006.

[19] T. H. Ong, H. Chen, W. K. Sung, and B. Zhu, "Newsmap: A knowledge map for online news," *Decision Support Systems*, vol. 39, no. 4, pp. 583-597, 2005.

[20] S. Pyo, "Knowledge map for tourist destinations—needs and implications," *Tourism Management*, vol. 26, no. 4, pp. 583-594, 2005.

[21] S. Kim, E. Suh, and H. Hwang, "Building the knowledge map: an industrial case study," *Journal of knowledge management*, vol. 7, no. 2, pp. 34-45, 2003.

[22] M. Li, B. W. Sun, and W. Zhang, "An approach to the construction of the personalized knowledge map in knowledge management systems," *Applied Mechanics and Materials*, vol. 598, pp. 736-739, 2014.

[23] I. Kang, Y. Park, and Y. Kim, "A framework for designing a workflow-based knowledge map," *Business Process Management Journal*, vol. 9, no. 3, pp. 281-294, 2003.
[24] F. R. Lin and J. H. Yu, "Visualized cognitive knowledge map integration for P2P networks," *Decision Support Systems*, vol. 46, no. 4, pp. 774-785, 2009.
[25] H. P. Tserng, S. Y. L. Yin, and M. H. Lee, "The use of knowledge map model in construction industry," *Journal of Civil Engineering and Management*, vol. 16, no. 3, pp. 332-344, 2010.
[26] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," *Mining Text Data*, 2012, ch. 4, pp. 77-128.
[27] M. W. Berry, "Survey of text mining," *Computing Reviews*, vol. 45, no. 9, pp. 548, 2004.
[28] S. Bloehdorn, P. Cimiano, and A. Hotho, "Learning ontologies to improve text clustering and classification," *From Data and Information Analysis to Knowledge Engineering*, Springer Berlin Heidelberg, 2006, pp. 334-341.
[29] Y. Liao and V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," *Computers & Security*, vol. 21, no. 5, pp. 439-448, 2002.
[30] M. R. Brito, E. L. Chavez, A. J. Quiroz, and J. E. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Statistics & Probability Letters*, vol. 35, no. 1, pp. 33-42, 1997.
[31] Y. Zhou, Y. Li, and S. Xia, "An improved KNN text classification algorithm based on clustering," *Journal of computers*, vol. 4, no. 3, pp. 230-237, 2009.
[32] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503-1509, 2012.
[33] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," in *Proc. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation*, vol. 69, 2013, pp. 1356-1364.

Ming Li is an associate professor at School of Business Administration, China University of Petroleum. His research interests include information management, knowledge management and multiple criteria decision making methods.

Mengyue Yuan is the graduate student at School of Business Administration, China University of Petroleum. Her research interests include information management and knowledge management.

Haitao Xiong is an associate professor at School of Computer and Information Engineering, Beijing Technology and Business University. His research interests include the information management, knowledge management and big data.