SCIENCE CHINA
Information Sciences

• LETTER •

July 2021, Vol. 64 179103:1–179103:3
https://doi.org/10.1007/s11432-018-9940-0

Knowledge forest: a novel model to organize knowledge fragments

Qinghua ZHENG[1,2], Jun LIU[1,2*], Hongwei ZENG[1,2], Zhaotong GUO[1,2], Bei WU[1,2] & Bifan WEI[2,3]

[1]*School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;*
[2]*National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an 710049, China;*
[3]*School of Continuing Education, Xi'an Jiaotong University, Xi'an 710049, China*

Received 10 August 2018/Revised 1 March 2019/Accepted 10 June 2019/Published online 16 March 2021

Citation Zheng Q H, Liu J, Zeng H W, et al. Knowledge forest: a novel model to organize knowledge fragments. Sci China Inf Sci, 2021, 64(7): 179103, https://doi.org/10.1007/s11432-018-9940-0

Dear editor,

With the rapid growth of knowledge, there is a steady trend of knowledge fragmentization. Knowledge fragmentization manifests as that the knowledge related to a specific topic in a course is scattered in isolated and autonomous knowledge sources [1]. We term the knowledge of a facet in a specific topic as a knowledge fragment. For example, "A push operation adds an item to the top-most location on the stack." is a knowledge fragment about facet operation of topic Stack. The problem of knowledge fragmentization brings two challenges. First, for knowledge is scattered in various knowledge sources, users need to make great efforts to search for the knowledge they are interested in, thereby leading to information overload [2]. Second, learning dependencies which refer to the precedence relationships among topics in the learning process is concealed by the isolation and autonomy of knowledge sources, thus causing learning disorientation [3]. However, three mainstream knowledge organization models [4–7], including term list, categorization, and relation list, which organize knowledge fragments without facet hyponymy and ignore learning dependencies among topics, can hardly be applied to address these two challenges.

To solve the knowledge fragmentization problem, we propose a novel knowledge organization model, knowledge forest, which consists of facet trees and learning dependencies. Facet trees can organize knowledge fragments with facet hyponymy to alleviate information overload. Learning dependencies can organize disordered topics to cope with learning disorientation. The knowledge forest uses a resource description framework (RDF) for knowledge representation and storage. Compared with RDF, knowledge forest organizes knowledge fragments in a way that is more consistent with human cognition and learning. Furthermore, we propose an effective construction method of knowledge forest. The construction process of knowledge forest contains facet tree construction, learning dependency extraction, and knowledge fragment assembly.

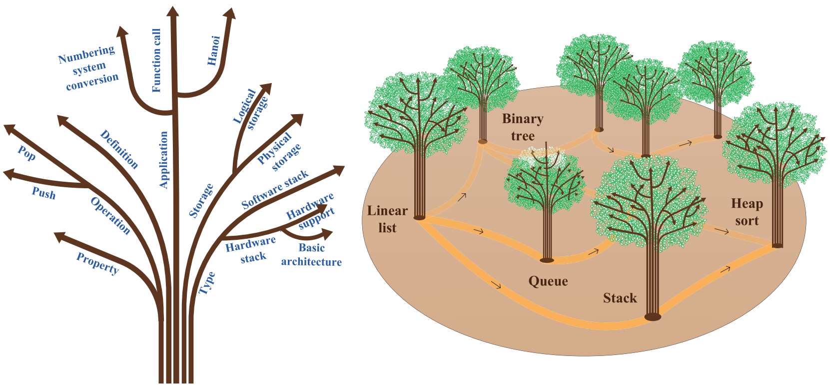
Definitions and notations. The related definitions and formalized notations of knowledge forest are introduced as follows.

Definition 1 (Facet tree)**.** A facet tree is a set of facets with facet hyponymy. Supposing $T = \{t_1, \ldots, t_n\}$ is the topic set of a course, the facet tree of topic $t_i \in T$ can be expressed as a tuple $FT_i = (F_i, RF_i)$. F_i refers to the facet set corresponding to t_i. $RF_i \subseteq (\{t_i\} \cup F_i) \times F_i$ represents topic-facet and facet-facet relationships. For example, Figure 1(a) shows a facet tree of a topic Stack, $F_{\text{Stack}} = \{\text{storage}, \text{operation}, \text{pop}, \text{push}, \ldots\}$ and $RF_{\text{Stack}} = \{(\text{Stack}, \text{operation}), (\text{operation}, \text{pop}), \ldots\}$.

Definition 2 (Materialized facet tree)**.** A materialized facet tree is a facet tree which is assembled with knowledge fragments. The materialized facet tree of topic $t_i \in T$ can be expressed as a triple $\text{MFT}_i = (FT_i, K_i, FK_i)$. K_i refers to the set of knowledge fragments corresponding to topic t_i. $FK_i \subseteq F_i \times K_i$ is the mapping relationship set between facet set F_i and knowledge fragment set K_i. Knowledge fragment $k \in K_i$ will be assembled to facet tree FT_i according to mapping relationship FK_i. For example, knowledge fragment "A push operation adds an item to the top-most location on the stack." will be assembled to facet operation of topic Stack.

Definition 3 (Knowledge forest)**.** A knowledge forest is the combination of materialized facet trees of topics and learning dependencies among topics. The knowledge forest can be expressed as a tuple $KF = (\text{MFT}, LD)$. $\text{MFT} = \{\text{MFT}_i \,|\, t_i \in T\}$ refers to the set of materialized facet trees corresponding to all topics in T. $LD \subseteq T \times T$ represents learning dependencies among topics in T. Figure 1(b) is a partial view of knowledge forest of the data structure course. The relationship with an arrow represents the learning dependency. For example, the learning dependency from topic linear list to Stack indicates that we should learn linear list first.

* Corresponding author (email: liukeen@xjtu.edu.cn)



(a) (b)

Figure 1 (Color online) Visualization of the knowledge organization model. (a) Facet tree of topic Stack; (b) the partial view of the knowledge forest of the data structure course.

Knowledge forest construction. The construction process of the knowledge forest, which can be regarded as the organization process of knowledge fragments, contains three steps as follows.

Step 1. We propose a facet propagation algorithm to construct facet trees. The method aims to give a facet set F_i for each topic $t_i \in T$. Intuitively, a pair of parent-child topics and a pair of brother topics both have similar facet sets. The pair of parent-child topics includes two topics with hypernymy relationship. The pair of brother topics includes two topics whose hypernym topic is the same one. For each topic, we parse and preprocess the contents of the corresponding Wikipedia webpage[1]) to obtain the initial facet set. Then, we use the facet propagation algorithm to complete the facet set of each topic on the basis of the initial facet set. During each facet propagation, the probability of facet $f \in F_i$ will be updated by the facet set similarity between both parent-child topic pairs and brother topic pairs. Until the algorithm converges, the probability of $f \in F_i$ bigger than 0.5 indicates that topic t_i includes facet f, and vice versa.

Step 2. We utilize our early work [8] to extract learning dependencies among topics. This method proposes two useful hypotheses, the distribution asymmetry of core terms and the locality of learning dependencies, which are essential for building the classification model to identify learning dependencies.

Step 3. We propose a mapping method based on convolutional neural network (CNN) to assemble knowledge fragments to corresponding facet tree [9]. This method aims to give one or more facet labels $F_i' \subseteq F_i$ for each $k \in K_i$, which consists of three steps. (i) We employ word embeddings to represent the words of knowledge fragments. Then, we use three convolution layers and three pool layers to represent each knowledge fragment as three matrices indicating the phrases information, corresponding to unigram, bigram and trigram, respectively. (ii) To reduce facet heterogeneity, we propose a text matching strategy to establish the relationship between each knowledge fragment and a facet label text (FaLT). First, we introduce FaLTs from Wikipedia webpages. For example, the FaLT corresponding to facet definition of topic Stack is "In computer science, a stack is an abstract data type that serves as a collection of elements, with two principal operations.[2])" Then, each FaLT is represented as three matrices by the knowledge fragment representation method mentioned above. Finally, three-dimensional similarity matrices are generated by cosine similarity measures between a knowledge fragment and a FaLT. (iii) We utilize the three-dimensional similarity matrices as the input of a three-channel CNN as multiple binary classifications for facet label assignment.

Datasets and basic statistic. We recruit ten participants who major in computer science with enough knowledge to annotate the knowledge forest. They independently annotate three courses, including data structure, data mining, and computer network. The course of data structure contains 193 topics, 35076 knowledge fragments, and 247 learning dependencies. The course of data mining contains 93 topics, 12723 knowledge fragments, and 128 learning dependencies. The course of computer network contains 84 topics, 13081 knowledge fragments, and 113 learning dependencies.

Experiments. To validate the effectiveness of automatic construction method of knowledge forest, we conduct experiments on these three courses. The nDCG score of facet tree construction can achieve more than 82%, and the Macro_F value of knowledge fragment assembly method can reach more than 83% on all three courses. The results indicate that our method implements a good generalization capability and can effectively organize knowledge fragments in different courses.

To evaluate the effectiveness of knowledge forest in alleviating information overload and learning disorientation, we conduct learning performance test. We recruit sixty participants for these three courses, and each course has twenty participants, ten of which in the control group and the other ten in the experimental group. We develop a prototype knowledge forest system named Yotta. The baseline is the control group which do not use Yotta to learn corresponding courses. The comparison metric is the mean and standard deviation of participant' scores in pre-test and post-test. The Student's t-test is used for statistical analysis which can be summarized as follows. First, the scores of pre-test, which are concerned with the participants' prior knowledge, have no significant differences between the control group and the experimental group in the three courses ($p > 0.05$). Second, participants' scores in post-test have significant improvements over pre-test both in the control group and the experimental group ($p < 0.05$). Third, the gain scores of

1) https://en.wikipedia.org/.
2) https://en.wikipedia.org/wiki/Stack(abstract_data_type).

the experimental group are much higher than those in the control group, which indicates that the participants in the experimental group can achieve significantly better learning performance than those in the control group ($p < 0.05$). Thereby, we can conclude that the knowledge forest is useful to alleviate the participants' information overload and learning disorientation.

Conclusion. We propose a novel knowledge organization model, knowledge forest, which consists of facet trees and learning dependencies. We propose an automatic construction method of knowledge forest. The results of extensive experiments conducted on three courses show that the knowledge forest can effectively organize knowledge fragments and alleviate information overload and learning disorientation.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2018YFB1004500), National Natural Science Foundation of China (Grant Nos. 61532015, 61532004, 61672419, 61672418), Innovative Research Group of National Natural Science Foundation of China (Grant No. 61721002), Innovation Research Team of Ministry of Education (Grant No. IRT_17R86), and Project of China Knowledge Centre for Engineering Science and Technology.

References

1 Liu W Q, Liu J, Wang M, et al. Faceted fusion of RDF data. Inf Fusion, 2015, 23: 16–24

2 Huang W D, Khoury R, Dawborn T, et al. WeBeVis: analyzing user web behavior through visual metaphors. Sci China Inf Sci, 2013, 56: 052106

3 Bajaj R, Sharma V. Smart education with artificial intelligence based determination of learning styles. Procedia Comput Sci, 2018, 132: 834–842

4 Hodge G. Systems of knowledge organization for digital libraries: beyond traditional authority files. Council on Library and Information Resources Pub91, 2000. http://www.clir.org/pubs/reports/pub91/pub91.pdf

5 Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of ACM SIGMOD International Conference on Management of Data, 2008. 1247–1250

6 Wu W T, Li H S, Wang H X, et al. Probase: a probabilistic taxonomy for text understanding. In: Proceedings of ACM SIGMOD International Conference on Management of Data, 2012. 481–492

7 Dong X L, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. 601–610

8 Liu J, Jiang L, Wu Z H, et al. Mining learning-dependency between knowledge units from text. VLDB J, 2011, 20: 335–345

9 Wu B, Wei B F, Liu J, et al. Facet annotation by extending CNN with a matching strategy. Neural Comput, 2018, 30: 1647–1672