Improvement on the Representation and Fusion Method of Fragmented Knowledge Structure

Lili Jia\textsuperscript{1} and Jinhua Yang\textsuperscript{1,*}

\textsuperscript{1}Dianchi College of Yunnan University, Kunming, China

\textsuperscript{*}Corresponding author e-mail: jialilidanchi1982@163.com

Abstract. With the rapid development of Internet, cloud computing, Internet of things and other new technologies, information is growing exponentially, and human society has entered a new stage of fragmented knowledge. Fragmentation is flexible, random, and contains rich information, but fragmented information is fragmented, disordered, incomplete, large in size and fast in updating, so it is difficult to form effective knowledge. To resolve the problem of extracting and mining fragmented knowledge, firstly, this paper studies the fragmented knowledge map and the integration method of fragmented knowledge. secondly, we study the integration method of fragmented knowledge, from the aspects of incomplete and redundant fragmented information; Finally, we put forward several points of fragmented knowledge processing methods Key scientific issues and challenges.

Keywords: Fragmented knowledge, Knowledge map, Structural representation, Knowledge fusion, Deep mining

1. Introduction

With the rapid development of the Internet, a large number of knowledge fragmentation has become the most prominent feature of the new era\textsuperscript{[1]}. Due to fragmentary resources and lack of authority, fragmentary knowledge is often one-sided and incomplete. Therefore, in order to form effective knowledge from fragmented knowledge, it is necessary to extract and mine fragmented information. However, the fragmented information has the characteristics of scattered distribution, diverse sources, large volume, complex composition, weak relevance, and fast update iteration. To fully extract and mine the knowledge contained in the fragmented information, it is necessary to express the fragmented knowledge and establish the correlation between the fragmented knowledge. In order to solve the problem of incomplete and redundant fragmented knowledge, it is also necessary to describe the fragmented knowledge Integrate knowledge.

Knowledge map is a new type of knowledge management technology\textsuperscript{[2-3]}. By using this technology, we can overcome the disadvantages of “Knowledge Island”. The most prominent feature of knowledge map is to realize the best integration of all information\textsuperscript{[4]}, so as to form understandable, descriptive and available knowledge. By using this technology, we can transform all kinds of features such as structured, unstructured and unstructured information into effective knowledge, which is embodied in the form of atlas. At present, the construction method of knowledge structure based on
knowledge map is still in the preliminary research stage, and the specific research results are few. Scholars at home and abroad have carried out some explorations on the non-integrity and redundancy of fragmented knowledge, such as the establishment of core matrix based on the incompleteness of data [5]. Although clear answers can be obtained through this method, this method is limited to core multi-modal learning. Some scholars also proposed the method of incomplete multimodal data fusion [6], such as Li et al. [7] proposed that subspace learning method can integrate norm sparseness, while literature [8] proposed norm regularization and literature [9-10] proposed compatible global structure and other factors to learn the shared features of incomplete multi-modal data. Although these methods can further realize the deep combination of incomplete multi-modal data, only linear or nonlinear transformation can make the multi-modal data tend to be the same, and the processing effect of fragmentation knowledge is not good. Therefore, the method of dealing with incomplete and redundant fragmented information needs further study and improvement.

From the perspective of knowledge map construction, this paper analyzed the representation of fragmented knowledge structure. On this basis, the paper further carried out research on incomplete and redundant fragmented information, proposed the integration method of fragmented knowledge. At last, the paper put forward the key scientific problems and challenges of the method of fragmented knowledge processing.

2. Research on the Representation of Fragmented Knowledge Structure

2.1. Network Topology Representation of Group Intelligent Knowledge Cluster

In view of the fragmented information obtained in the user generated content (UGC) platform, these fragmented information lacks a unified logical structure, such as micro blog, SMS, audio and video and other multi-modal data, which has nonlinear characteristics. Therefore, based on the group intelligence principle, the rapid nonlinear discriminant analysis (ENDA) is used to extract it from the nonlinear high-dimensional feature space. Taking out the most discriminative low dimensional features, these features can help to gather all the fragmented information of the same category together to form a knowledge cluster, and the fragmented information of different clusters should be separated as far as possible, that is to say, the feature with the largest ratio of the dispersion between and within the fragmented knowledge clusters should be selected to form a group intelligent knowledge cluster.

The dimension reduction and classification are carried out by using the extreme speed nonlinear discriminant analysis ((ENDA).First, realize the original characteristics of fragmented knowledge $X = \{x_1^T, x_2^T, \cdots, x_N^T\} \in R^{N \times m}$ to $H = \{h_1^T, h_2^T, \cdots, h_N^T\} \in R^{N \times L}$Fast random mapping.Second, using linear discriminant analysis (LDA) to dimension reduction for $H$, from L dimension to $P( P \leq m-1 )$ dimension, $\hat{H} = \{\hat{h}_1^T, \hat{h}_2^T, \cdots, \hat{h}_N^T\} \in R^{N \times \tilde{L}}$can be obtained,and $\hat{H} = W^T H , W = [w_1, w_2, \cdots, w_p]$, $w_i$ is weight of linear discriminant analysis .The Fisher criterion function is used for discriminant analysis, which is defined as follows:

$$J(w) = \arg \max \frac{w^T S_b w}{w^T S_w w} = [w_1, w_2, \cdots, w_p]$$ (1)

$S_b$, $S_w$ is the inter cluster dispersion matrix and the intra cluster dispersion matrix of fragmentation knowledge, which are defined as,

$$S_b = \sum_{k=1}^{m} \frac{N_k}{N} (\overline{x_k} - \overline{x})(\overline{x_k} - \overline{x})^T \quad S_w = \sum_{k=1}^{m} \sum_{i=1}^{N_k} \frac{1}{N} (x_{ki} - \overline{x})(x_{ki} - \overline{x})^T$$ (2)

$N$ is all fragmentation information, $m$ is number of knowledge clusters, $N_k$ is the number of fragmented information in the $k$-th knowledge cluster, $x_{ki}(k=1,2,\cdots,m, i=1,2,\cdots,N_k)$ is the $i$-th
fragmentation information of the $k$-th knowledge cluster, $\overline{x}$ is the mean value of all fragmentation information, $\overline{x}_i (k = 1, 2, \cdots, m)$ is the mean value of $k$-th knowledge cluster.

Finally, the optimal discriminant projection matrix and the projection weight of the entire extreme speed nonlinear discriminant analysis (ENDA) can be obtained from the generalized feature problem.

![Fig. 1 flow chart of extreme speed nonlinear discriminant analysis](image)

**Fig. 1** flow chart of extreme speed nonlinear discriminant analysis

Topological representation of knowledge cluster network. Network topology is the interconnection of entity elements on the network. After the formation of group intelligence knowledge cluster, the formation of fragmented information is limited to a single domain of knowledge nodes, which uses the internal relationship of information to ultimately achieve the purpose of information exchange. Figure 2 shows the general framework of topological representation of swarm intelligence network by using hypertension diagram.

![Figure 2. Topological representation of hypertension group intelligence network](image)

**Figure 2.** Topological representation of hypertension group intelligence network

### 2.2. Build the Structure Model of Fragmented Knowledge Map

The knowledge map can not only show the network structure of knowledge, but also show its semantic relationship. Taking the fragmented knowledge acquired by the UGC platform of user generated content as the research object, after the group intelligent knowledge cluster is represented by the network topology, the related fragmented knowledge points among the knowledge clusters are connected to form the knowledge map framework, which is regarded as the subject knowledge base. It can point out the aspects of knowledge search, realize visualization, and create good conditions for knowledge sharing, dissemination and use.

The so-called knowledge map is to define the knowledge base from the structural level, and to reflect the definition of the objective world and the relationship between various factors through symbols. The basic component element is “entity relationship entity”, which also includes the relevant attribute values of entities and entities. The network knowledge framework is established by using the relationship of each entity. In this paper, we propose to connect the knowledge clusters reasonably and quantify the relationship between the edge and the node in the structure of the knowledge map based on the similar matrix construction algorithm of connected triplets. The basic idea of connecting triples: set the black center circle to represent the fragmented information node, the solid line big circle to represent different groups of intelligent regions, and the green region to represent the knowledge clusters with similarity through connecting triples.
If knowledge points $x_i \in C_j^1$, there will be an edge between the point and the cluster. For partition $\pi_2$ and $\pi_3$, cluster $C^2_i$ and cluster $C^3_j$ include knowledge points $x_i$ and $x_j$, it can be considered that these two points are similar. For partition $\pi_1$, knowledge points $x_i$ and $x_j$ belong to cluster $C^1_i$ and cluster $C^1_j$ respectively. Algorithm based on join triples, because cluster $C^1_i$ and cluster $C^1_j$ have two connected triples, and cluster $C^1_i$ and cluster $C^1_j$ are the center of the two connected triples, so cluster $C^1_i$ and $C^1_j$ cluster have similarity. For the partition $\pi_1$, it also shows that point $x_i$ and point $x_j$ is similar, and the knowledge similarity is low. The weight of the edge of the connected cluster $C_i$ and cluster $C_j$ is obtained from the number of knowledge points contained by the two clusters:

$$\omega_{ij} = \frac{|x_i \cap x_j|}{|x_i \cup x_j|}$$

(3)

$x_i$ and $x_j$ is the set of knowledge points of cluster $C_i$ and cluster $C_j$ respectively. Adjacency point $C_i$ is the values of connected triples between cluster $C_i$ and cluster $C_j$, so the values of all triples between two clusters can be calculated. The calculation formula of similarity between clusters is

$$Sim(i, j) = \frac{\text{cluster}C_i \text{and} C_j \text{Values of triples}}{\text{The largest of all triples between clusters}} \times DC$$

(4)

$DC$ is a decay factor, we can calculate the similarity between knowledge points

$$S_m(x_i, x_j) = \begin{cases} 1, & C(x_i) = C(x_j) \\ Sim(i, j), & C(x_i) \neq C(x_j) \end{cases}$$

(5)

The computable triple similarity matrix is

**Figure 3.** structure model of knowledge map based on connected triples
\[
S(x_i, x_j) = \begin{cases} 
\frac{1}{M} \sum_{y \in 1} N_y \rho_{y} C(x_y) = C(x_j) \\
\frac{1}{M} \sum_{m=1}^{n} S_m(x_i, x_j), C(x_i) \neq C(x_j) 
\end{cases}
\]

(6)

\(M\) is number of clustering results.the result of \(N_y\) is when knowledge points 2 and 3 belong to the same cluster in \(n\) species division, the value is 1, \(\rho_y\) is weight.

This method ignores the mechanism of low-quality results instead of selecting by generating threshold. Compared with the traditional correlation matrix, the calculation of similarity matrix considers the similarity between two knowledge points when they do not belong to the same cluster. It can be seen that this method expands the potential information between knowledge points and is conducive to the information clustering with complex structure.

3. Research on the Fusion Method of Fragmented Knowledge

3.1 Incomplete Knowledge Fusion Algorithm for Massive Fragmentation

Some incomplete fragmented information samples are selected from the user generated content (UGC) platform, and the extended application of incomplete multi-modal data fusion algorithm based on deep semantic matching is proposed.Assumed incomplete multi-modal fragmentation data set \(X = \{X_{(v)}^{(i)}\}_{i=1}^{V} = \{X_{c}^{v(i)}, X_{f}^{v(i)}\}_{i=v}^{V}\) it contains \(V\) data modes and \(n\) fragmented data belonging to \(k\) knowledge clusters, \(X_{c}^{v(i)} \in R^{c \times n}, X_{f}^{v(i)} \in R^{f \times n}\) represents the \(c\)-th modal eigenmatrix of the \(v\)-th complete modal data and other data in the \(n\) modal. There are \(c + n\) data in a mode, and one data presents \(d\) dimensional attribute characteristics. Different modal data descriptions generally have the same semantic representation,First, each modal data is transformed into depth feature by depth neural network, which is expressed as \(H = \{H^{(i)}\} = \{H_{c}^{(i)}, H_{f}^{(i)}\}\). Second, by using the non negative matrix decomposition model with local invariant graph regularization, all modal representations \(H^{(i)}\) is transformed into the base matrix \(U^{(i)}\) and the same coding matrix \(P^{(i)} = [P_{c}^{i}; P_{f}^{i}]\). Finally, the deep semantic sharing feature of multi-modal fragmented data is obtained. The specific depth semantic matching model of incomplete multi-modal data is

\[
\min \sum_{i=1}^{V} \left\| H_{c}^{(i)}; H_{f}^{(i)} \right\| - U^{(i)}(P_{c}^{i}; P_{f}^{i}) \geq \alpha^{(i)} Tr(P^{(i)} U^{T} P^{(i)})
\]

(7)

\(s.t. U^{(i)} \geq 0, P^{(i)} = [P_{c}^{i}; P_{f}^{i}] \geq 0\)

(8)

\(H^{(i)} = [H_{c}^{(i)}; H_{f}^{(i)}] = f(W_{c}, X^{(i)} + b_{c})\) is feature output of each modal depth network. \(f\) is nonlinear activation function. \(W_{c}, b_{c}\) is the relative weight and bias vector. Then, by connecting all the modal depth learning networks, the base matrix and the same coding matrix, the multi-modal depth semantic shared subspace is obtained, in which the multi-modal data features can be fused and analyzed.

3.2 Redundant Massive Fragmented Knowledge Fusion Method

From the user generated content (UGC) platform, low-quality, inefficient and chaotic fragmented information samples are selected. Based on the fragmentary knowledge map, the correlation matrix \(R\)
and its eigenvalue $\lambda$ of knowledge cluster are calculated by using the feature selection method of correlation information entropy. According to the correlation information entropy formula $H_{R} = -\sum_{i=1}^{n} \frac{\lambda_{i}}{n} \log_{2} \frac{\lambda_{i}}{n}$, the correlation information entropy matrix $H_{Rij}$ is calculated and the redundancy metric $H_{Rel}$ is judged. The lower the degree of feature redundancy, the higher the degree of independence, and the greater the entropy of correlation information of features that are more related to non-cluster features, so as to reduce the redundant information between features, fully consider the multi variable relationship between features of different knowledge clusters, and calculate the information measurement of correlation relationship between multi variables in the range of $[0,1]$, which improves the performance of fragmentation knowledge in classification accuracy.

4. Problems and Challenges
The research on the depth mining of fragmented knowledge is relatively few and the results are limited. Wang Jianji et al opened the curtain of fragmented knowledge processing and networked artificial intelligence, which laid a good foundation for the follow-up research, but also left many scientific problems to be solved, such as whether to propose a theory or framework suitable for dealing with fragmented knowledge? How can we integrate the overload of fragmented knowledge with the metaphor of independent knowledge, eliminate redundant information and ensure the orderly organization and expression of knowledge? How to mine useful knowledge from a large amount of fragmented information to create conditions for AI system to be more intelligent? How to perform machine reasoning and knowledge deduction in the massive fragmented knowledge environment? Due to the widespread and exponential growth of fragmented information, the solution of the above problems is of great significance for fully mining and utilizing fragmented information. In this regard, the following scientific issues and challenges need to be considered and studied by researchers:

(1) The construction of knowledge map of massive fragmentation is the key problem of the research content “structural representation model of massive fragmentation knowledge”. In order to form the effective application of fragmented knowledge, it is necessary to establish a scientific and reasonable knowledge model of fragmentation. The source of fragmented knowledge is disordered fragmented information. How to transform the massive fragmented information with a wide range of points and disorder into the available knowledge model of fragmentation will become a research difficulty, and the fragmented information in various fields will be represented by the knowledge map model. It is a key scientific problem that needs to be solved in fragmented knowledge modeling.

(2) Incomplete and redundant knowledge processing method is the key problem of the research content “integration of massive fragmented knowledge”. After the construction of fragmented knowledge map, the fragmented knowledge in various fields has been effectively aggregated and classified, but the inherent incomplete and redundant characteristics of fragmentation determine that it is not possible to effectively aggregate fragmented information. Therefore, how to improve the availability of fragmented knowledge will be a research difficulty, and how to deal with incomplete and redundant fragmented knowledge is to improve the fragmentation knowledge. The availability of knowledge needs to solve a key scientific problem.

(3) The intelligent reasoning method of fragmented information changing with time and space is the key problem of the research content “deep mining and reasoning of massive fragmented knowledge”. The value of dynamic fragmented information lies in the comprehensive mining of potential and deep-seated implicit knowledge, so as to form a complete and efficient knowledge system. Therefore, how to mine and reason the massive fragmented knowledge is a research difficulty and key point, and how to carry out intelligent reasoning of dynamic fragmented knowledge is the deep mining and reasoning of fragmented knowledge. A key scientific problem.

Acknowledgement
This work was supported in part by the Program for Innovative Research Team (in Science and
Technology) in University of Yunnan Province, and the scientific research funding of the Education Department of Yunnan Province under Grant 2018JS737.

References

[1] Wang Jianji, Ma Yongqiang, Chen Shitao, et al. Fragmented knowledge processing and networked artificial intelligence [J]. Chinese Science (Information Science), 2017, 47 (2): 171-192.

[2] Chen Y, Liu Z Y. The rise of mapping knowledge domain. Stud Sci Sci, 2005, 23: 149–154.

[3] YU T, LI J H, YU Q, et al Knowledge graph for TCM health preservation [J]. Artificial Intelligence in Medicine, 2017, 77: 48-52.

[4] YAN J H, WANG CH Y, CHENG W L, et al. A retrospective of Knowledge graphs [J]. Frontiers of Computer Science, 2018, 12(1): 55-74.

[5] Shao W, Shi X, Yu P S. Clustering on multiple incomplete datasets via collective kernel learning [C]. Proceedings of 13th IEEE International Conference on Data Mining, Dallas: IEEE, 2013: 1181-1186.

[6] Xu C, Tao D, Xu C. Multi-view learning with incomplete views [J]. IEEE Transactions on Image Processing, 2015, 24(12): 5812-5825.

[7] Li S Y, Jiang Y, Zhou Z H. Partial multi-view clustering [C]. Proceedings of 28th AAAI Conference on Artificial intelligence, Quebec City: AI Access Foundation, 2014: 1968-1974.

[8] Shao W, He L, Yu P S. Multiple incomplete views clustering via weighted nonnegative matrix factorization with L2, Irregularization [C]. Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Porto: Springer-Verlag, 2015: 318-334.

[9] Yin Q, Wu S, Wang L. Unified subspace learning for incomplete and unlabeled multi-view data [J]. Pattern Recognition, 2017, 67(67): 313-327.

[10] DIAO Ai-jun. Hybrid cloud redundant data delete algorithm based on feature code compression [J]. Bulletin of Science and Technology, 2015, 31 (8): 42-44.