Information Visualization from the Perspective of Big Data Analysis and Fusion

XIANG LIN
Research Center for Advertising and Media Economics, Wuhan Technology and Business University,
Wuhan 430065, CHINA

Abstract—In the big data environment, the visualization technique has been increasingly adopted to mine the data on library and information (L&I), with the diversification of data sources and the growth of data volume. However, there are several defects with the research on information association of L&I visualization network: the lack of optimization of network layout algorithms, and the absence of L&I information fusion and comparison in multiple disciplines, in the big data environment. To overcome these defects, this paper explores the visualization of L&I from the perspective of big data analysis and fusion. Firstly, the authors analyzed the topology of the L&I visualization network, and calculated the metrics for the construction of L&I visualization topology map. Next, the importance of meta-paths of the L&I visualization network was calculated. Finally, a complex big data L&I visualization network was established, and the associations between information nodes were analyzed in details. Experimental results verify the effectiveness of the proposed algorithm.

1. Introduction

LIBRARY and information (L&I) visualization breaks the shackles of knowledge-based solutions. This technique changes the rules, technologies, or service forms of book management [1-3]. In the big data environment, digital L&I information resources increase and age at a growing pace, with the diversification of data sources and the growth of data volume. Meanwhile, the user demand for discovery services is changing rapidly [4-6]. The visualization technique has been increasingly adopted to mine the data on L&I [7-11]. Therefore, it is of certain practical significance to study the L&I visualization from the perspective of big data analysis and fusion.

Currently, the research on L&I visualization mainly focuses on the visualization methods and visualization analysis tools of information resources and information retrieval [11-15]. Mo et al. [16] analyzed the status quo of L&I visualization in terms of annual number of published papers, authors, journals, and keywords, and explained the utilization of visualization software CiteSpace with an actual case. To deeply mine potential knowledge and disclose the deep correlations between L&I, Rowley et al. [17] introduced co-occurrence analysis to visualize L&I, and quantified the co-occurrence information in L&I information carriers. Borrego [18] sorted out and summarized the current research of L&I data-driven knowledge discovery, clarified the core ideas of library knowledge discovery and L&I visualization service, and detailed the innovative data environment, driving mechanism, and pattern application, which are necessary for digital library knowledge discovery and L&I visualization service. Johnson [19] reviewed the results of cross disciplinary research on L&I in China, visualized the relevant research results with automated valuation models (AVMs), and discussed the differences between various methods, namely, cluster analysis, survey method, and strength, weakness, opportunity, and threat (SWOT) analysis.

The information isolated island and overloading of book data hinder users from acquiring and sharing of L&I [20-21]. L&I information is commonly organized in two modes: information fusion and information aggregation [22-23]. Paramonova [24] enhanced the semantics of L&I information through information classification, ontology analysis, and data association method, and relied on the semantics for knowledge organization and visual metering. In this way, knowledge organization was integrated with econometric analysis to mine the resource associations, through the analysis of co-occurrence relationships, coupling relationships, and social network analysis.

Through the above analysis on domestic and foreign research, there are several defects with the research on information association of L&I visualization network: the lack of optimization of network layout algorithms, and the absence of L&I information fusion and comparison in multiple disciplines, in the big data environment [25-27]. To overcome these defects, this paper explores the visualization of L&I from the perspective of big data analysis and fusion. Section 2 analyzes the topology of the L&I visualization network, and calculates the metrics for the construction of L&I visualization topology map, node degree distribution, clustering coefficient, and mean path length. Section 3 calculates the importance of

meta-paths of the L&I visualization network. Section 4 establishes a complex big data L&I visualization network, and analyzes the associations between information nodes. The visualization results of the proposed algorithm were obtained through experiments, which confirm the effectiveness of our algorithm.

2. Topology Analysis

L&I information is a dataset composed of books, literature information, and archival information on the library management platform. Let V_1, V_2, …, V_m be the information of a book or document in the L&I group information; AS_m be the association between multiple books or documents. Then, the set QX of L&I group information can be expressed as:
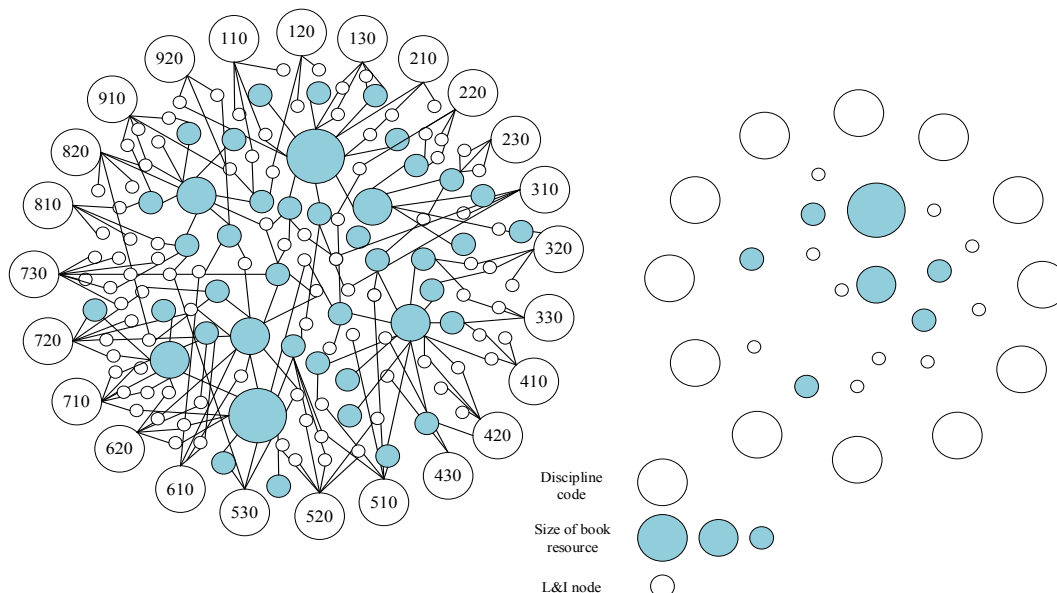
$$QX = \{V_1, V_2, ..., V_m\} \tag{1}$$

Fig. 1 Visualization of L&I network

Fig. 1 visualizes the L&I network. On the left side are the visualized information of different disciplines and books, as well as L&I nodes. On the right side are the visualized codes of the visualized information on the left. In Figure 1, the L&I nodes are coded in multiple dimensions by their respective size and color, according to the type of book resources. From the structural features of the visualized network, the following will calculate the metrics of the topology of L&I visualization network, including node degree distribution, clustering coefficient, and mean path length.

Let l be a node with the degree of l in the L&I visualization network; M be the total number of nodes. Then, the node degree distribution $ND(l)$ can be described by:

$$ND(l) = l / M \tag{2}$$

If the constructed network is a scale-free network, the degree distribution of nodes can be obtained by the power law distribution features:

$$ND(l) \sim l^{(-e)} \tag{3}$$

The mean clustering coefficient AC^* that measures the closeness between multiple book or document nodes can be expressed as:

$$AC^* = \sum_l ND(l)AC(l) \tag{4}$$

Let $N_{mn}(l)$ be the mean number of connects between a node

with the degree of l and its adjacent nodes. Then, the local clustering coefficient in formula 4 can be calculated by:

$$AV(l) = N_{mn}(l) / l(l-1) \tag{5}$$

Network closeness NC, which represents the associations between book or document nodes in the L&I visualization network, can be calculated by:

$$NC = \sum_{r,o} b_{r-o} / M(M-1) \tag{6}$$

Let QX_{r-o} be the number of the shortest paths between nodes r and o; $QX_{r-o(q)}$ be the number of the shortest paths between nodes r and o, which pass through node q. Then, the betweenness centrality BE, which reflects the influence of a node in the L&I visualization network, can be calculated by:

$$BE = \sum_{q<o} QX_{r-o(q)} / QX_{r-o} \tag{7}$$

Formula 7 shows that, the greater the BE, the more likely that the corresponding node will be visualized in the information network.

In essence, the information query in the L&I visualization network is to find the L&I nodes with similar semantic relationships as the input node in the query. This paper abstracts the multiple paths between two nodes into a meta-path, and describes the semantic association between nodes in an advanced level. Fig. 2 provides the examples of the paths and
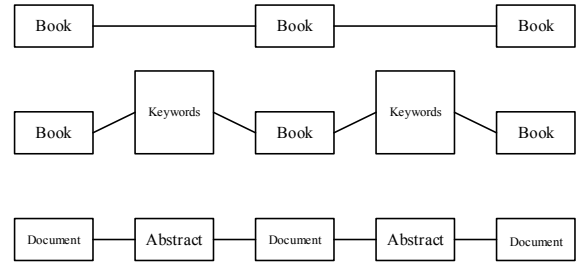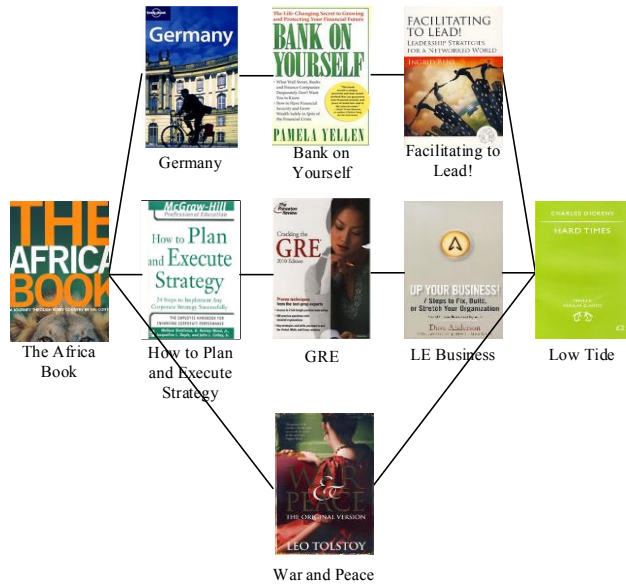
meta-path between nodes.

Fig. 2 Examples of the paths and meta-path between nodes

For the L&I visualization network, the minimum mean path length can be denoted as K_{AV}, and the maximum clustering coefficient can be denoted as CO_{max}. Suppose all the nodes in the network are connected by meta-paths, then the network is fully coupled, i.e., the network state satisfies $K_{AV}=1$ and $CO_{max}=1$. The length of meta-paths changes regularly with the distances between nodes. In the fully coupled state, the L&I visualization network has a fixed number of nodes and a fixed number of meta-paths. M nodes need to be connected by $M(M-1)/2$ meta-paths.

In the network, a node r is connected to its adjacent nodes, and the nodes within the distance of l/2 on the left and right. Let l be the number of nearest neighbors of the node. Then, the clustering coefficient of node r can be calculated by:

$$CO_r = \frac{3(l-2)}{4(l-1)} \approx \frac{3}{4} \tag{8}$$

When the number of nodes is infinitely large, the mean path length of the L&I visualization network can be calculated by:

$$K_{AV} \approx \frac{M}{2l} \to \infty (M \to \infty) \tag{9}$$

Suppose the L&I visualization network is a star network, in which any node is only connected to one node. Then, the clustering coefficient of the M-node network can be calculated by:

$$CO_{ST} \approx \frac{M}{M-1} = 1(M \to \infty) \tag{10}$$

The mean path length of the network can be calculated by:

$$K_{ST} = 2 - \frac{2(M-2)}{M(M-1)} = 2(M \to \infty) \tag{11}$$

If the number and distribution meta-paths are random, then the L&I visualization network is a stochastic network. As long as the mean node degree l remains unchanged, the presence/absence of each meta-path in the random network is uncertain. Then, the node degree distribution of the meta-paths

in the network can be described by the Poisson's distribution below:

$$PO(l) = CO_{M-1}^l PO^l (1-s)^{M-1-l} \tag{12}$$

3. Calculation of Meta-Path Importance

Without considering the limitations on L&I contents, the first step to find the optimal meta-path that depicts node relationships is to determine the meta-path length, path number, and path rarity. The optimal meta-path refers to the meta-path with the greatest importance. Let $LU_{r \to o}$ be the set of meta-paths between nodes r and o; $R_{r-o}(LU)$ be the importance support function of meta-path LU; $E_{r-o}(LU)$ be the rarity of LU; $PE(/LU/)$ be the length attenuation function of LU. Then, the importance $ZY_{r,o}(LU)$ of meta-path LU connecting the given node pair <r, o> can be calculated by:

$$ZY_{r-o}(LU) = R_{r-o}(LU) \cdot E_{r-o}(LU) \cdot PE(|LU|)(LU \in LU_{r \to o}) \tag{13}$$

Traditionally, the word frequency of L&I contents is calculated by counting the term frequency–inverse document frequency (TF-IDF) value. However, the TF-IDF-based method performs poorly on relatively short L&I texts, which has a few words and lacks context information after word segmentation. In this paper, short L&T texts are firstly segmented into words, and high-quality Chinese word vectors are generated by the directional skip-gram (DSG) model. Let ST be the input short L&I text; $[w_0, w_1, \ldots w_i \ldots w_L]$ be the word sequence corresponding to the short text; L be the length of the word sequence; w_i be the i-th word; U_{wi} be the corresponding word vector. From the word vector, the sentence vector U_{ST} can be derived as:

$$U_{ST} = \frac{1}{m} U_{w_i} \tag{14}$$

Compared with short L&I texts, the abstracts of books and documents offer clear contextual relationships and many word

vectors. To acquire the weighted mean sentence vector, the weight coefficients of words can be determined through statistical method. Let Φ be the abstract of a book or a document; $[w_0, \ldots w_i \ldots w_n]$ be the word sequence segmented from the abstract; $[\omega_0, \ldots \omega_i \ldots \omega_n]$ be the weight sequence of the segmented words. Then, a word vector $U_{\Delta wi}$ can be constructed for each word Δw_i segmented from the abstract. Then, the weighted mean sentence vector U_Φ can be calculated by:

$$U_\Phi = \frac{1}{n} \omega_j \bullet u_{\tau_j} \tag{15}$$

The similarity between the sentence vectors of two texts, namely, the input short text and the abstract of a book or document can be measured by cosine similarity:

$$S(ST, \Phi) = 1 - \frac{U_{ST} \cdot U_\Phi}{\|U_{ST}\| \cdot \|U_\Phi\|} \tag{16}$$

Based on the similarity of the short text of abstract, the importance of meta-paths $R_{r-o}(LU)$, $E_{r-o}(LU)$, and $PE(/LU/)$ can be measured. Let $\alpha \in (0, 1)$ be the attenuation coefficient. Then, a length penalty function $PE(/LU/)=\alpha^{-/LU/}$ can be configured to effectively lower the importance of meta-path LU, when it is relatively long. The length attenuation of the length penalty function could be adjusted by changing the value of α.

The rarity function $E_{r-o}(LU)$ aims to evaluate how rare is the meta-path LU in the set of node pairs WC_{r-o} similar to node pair $<r, o>$. WC_{r-o} can be defined as:

$$WC_{r-o} = WC_o \cup WC_r \tag{17}$$

where, set WC_o can be expressed as:

$$WC_o = \{(r,u) | u \in U, \rho(u) \cap \rho(o) \neq \phi\} \tag{18}$$

Set WC_r can be described as:

$$WC_r = \{(u,o) | u \in U, \rho(u) \cap \rho(r) \neq \phi\} \tag{19}$$

Since o does not belong to set WC_o, then r does not belong to set WC_r. Drawing the idea of IDF calculation, the rarity function $E_{r-o}(LU)$ can be described as:

$$E_{r-o}(LU) = log \frac{|WC_{r-o}|}{|\{(v,u) \in WC_{r-o}, LU \in LU_{v \to u}\}|} \tag{20}$$

To rank the meta-paths by importance, the meta-path importance function $ZY_{r-o}(LU)$ must be monotonically decreasing and have a maximum. Since the length penalty function $PE(/LU/)$ is a strict monotonically decreasing function, formula 20 must satisfy:

$$\frac{|WC_{s-t}|}{|\{(v,u) \in WC_{r-o}, LU \in LU_{v \to u}\}|} \geq 1 \tag{21}$$

Formula 21 shows that $E_{r-o}(LU)$ is monotonically increasing, with a minimum of 0. According to formula 13, for $ZY_{r-o}(LU)$ to be monotonically decreasing and have a maximum, $R_{r-o}(LU)$ must have the same properties. To ensure that $R_{r-o}(LU)$ have these properties, this paper adopts the minsize of path that can represent the minsize of each type of nodes in LU. The meta-path of a given node pair $<r, o>$ can be described by:

$$LU = D_1 \xrightarrow{E_1} D_2 \xrightarrow{E_2} \ldots \xrightarrow{E_k} D_{k+1} \tag{22}$$

Let IC_i be the number of instances for the i-th node on the meta-path. Then, the minsize of LU can be calculated by:

$$MU_r(LU) = \min_{1 < i < j+1} |\{IC_i | IC \in LU\}| \tag{23}$$

Due to the existence of the following inequality:

$$\min_{1 < i < j+1} |\{IC_i | IC \in LU\}| \leq \min_{1 < i < j+2} |\{IC_i | IC * E_i\}| \tag{24}$$

The minsize MU of LU is monotonically decreasing. In the L&I visualization network, the relationships do not take the same form. Instead, imbalanced relationships might exist, such as one-to-many or one-to-one relationships. To solve the problem, this paper resorts to the calculation of enhanced minsize, which is obtained by applying an intensity factor to the minsize. The intensity factor can be calculated by:

$$STR(LU) = \frac{1}{min(UZ(D), RZ(D))} \tag{25}$$

Let D be the nodes with minsize obtained by formula 23; $UZ(D)$ and $RZ(D)$ be the out-degree and in-degree of D, respectively. If node D is a book or document, $UZ(D)$ can be calculated by:

$$If \ D = MO, UZ(D) = \sum_{n \in IC^D} S(U_n, U_{ST}) \tag{26}$$

Formula 26 shows that $UZ(D)$ is the sum of similarities between the short L&I text and each node vector in instance set D. Through the above analysis, MU_{r-o} and $STR(LU)$ can be combined to get the support function $R_{r-o}(LU)$ of meta-path importance:

$$R_{r-o}(LU) = STR(LU) \cdot MU_{r-o}(LU) \tag{27}$$

Meta-paths are extensible. The importance of an extended meta-path equals the product between the current $R_{r-o}(LU)$ and the intensity coefficient $STR(E_i)$ of the extended edge E_i.

4. Network Construction and Association Analysis

4.1 Network construction

Let CP_{r-o} be the probability for an edge to link up two nodes at the same time. The probability indicates how many of the edges to node o also link up node r. The conditional probability for the connection from node r to node o can be defined as:

$$CP_{r-o} = \frac{LI_{r-o}}{LI_r} \tag{28}$$

where, LI_{r-o} be the number of other nodes connected to nodes r and o; LI_o be the number of other nodes connected to node o. Then, the following inequality holds:

$$LI_o \leq \sum_r LI_{r-o} \tag{29}$$

Some edges could link up more than 2 nodes. Hence, the correlations vary from node to node:

$$CP_{r-o} = \frac{r \xrightarrow{LI} o}{LI_r} \tag{30}$$

Let $r \xrightarrow{LI} o$ be the number of edges from node r to node o; LI_o be the change in the number of edges to node o. Because no node records the same value repeatedly, duplicate values should

be removed from network construction and parameter selection. The similarity between two nodes in the L&I visualization network can be calculated by:

$$COR_{r-o} = \frac{LI_{r-o}LI - LI_r LI_o}{\sqrt{(LI-LI_r)(LI-LI_o)LI_r LI_o}}$$ (31)

The calculation of network instances shows that the correlation between most information nodes equals 1. When there are only a few nodes, the linear relationship between variables LI_r and LI_o cannot be judged by correlation coefficient alone. Table 1 lists the calculation results of some parameters for the part of the network with fewer than 2 nodes.

Table 1. Some calculation results of scale-free network

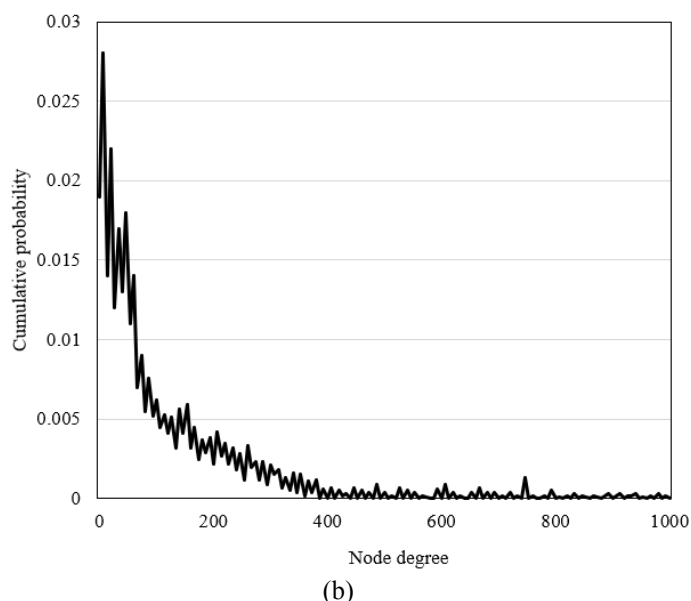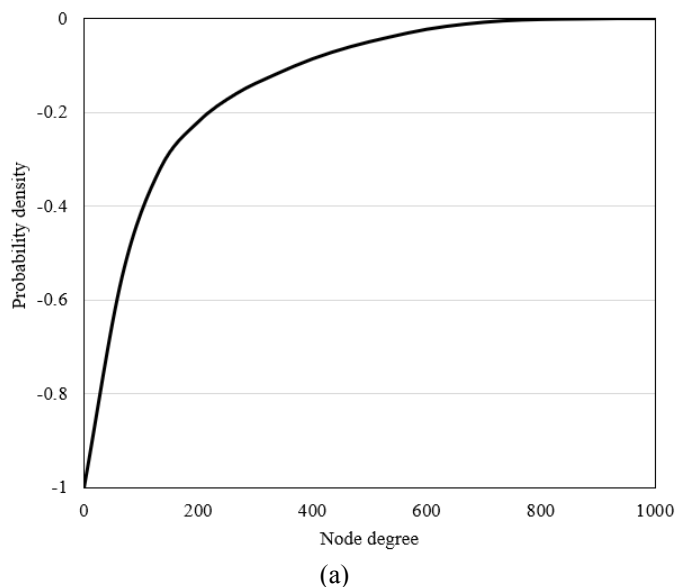| Serial number	Node r	Node o	Number of nodes	Correlation
1	4.023	4.172	31	1
2	4.176	4.36	31	1
3	5.38	5.37	11	1
4	10.82	10.58	32	1
5	10.49	10.53	42	1
6	11.374	11.372	12	1
7	16.275	15.981	14	1
8	10.73	10.76	26	1
9	15.072	15.492	10	1
10	4.127	4.39	31	1

(a)

(b)

Fig. 3 Power law distribution features of node degree

Fig. 3 presents the calculated node degree distribution of the established scale-free L&I visualization network. Only 3.6% of all network nodes had a correlation greater than 0.8, which satisfies the construction principle of complex network and the theoretical power law distribution of node degree.

4.2 Association analysis

To further measure the association between L&I nodes, this paper analyzes the similarity of individual books and documents, and that of a group of books and documents in L&I visualization network. The most important aspect of the analysis is the modeling of network nodes. The L&I nodes were analyzed and modeled from three aspects: the basic information BA, the text information TE, and the relationship information RE.

Let $IM(r, o)$ be the similarity between two nodes r and o. For node r, a model can be established as $US(r)=\{BA(r), TE(r), RE(r)\}$; For node o, another model can be established as $US(o)=\{BA(o), TE(o), RE(o)\}$. Based on the data of BA, TE, and RE, the similarity between nodes could be calculated. Firstly, the similarity of nodes was calculated in terms of the three attributes. The similarity $S(r_\mu, o_\mu)$ between the basic information $BA(r)$ and $BA(o)$ of nodes r and o can be described by:

$$S(r_\mu, o_\mu) = S(BA(r), BA(o))$$ (32)

The similarity between the text information $TE(r)$ and $TE(o)$ of nodes r and o can be described by:

$$S(r_\sigma, o_\sigma) = S(TE(r), TE(o))$$ (33)

The similarity between the relationship information $RE(r)$ and $RE(o)$ of nodes r and o can be described by:

$$S(r_\gamma, o_\gamma) = S(RE(r), RE(o))$$ (34)

The basic information of network nodes mainly includes the publisher, publication date, and number of pages. The basic information of nodes r and o can be expressed as $BA(r)=\{LO(r), AG(r), TA(r)\}$, and $BA(o)=\{LO(o), AG(o), TA(o)\}$,

respectively. Let δ_i be the weight of the similarity for each attribute of the basic information. Then, the similarity between the basic information can be calculated by:

$$S(r_\lambda, o_\mu) = \delta_1(LO(r), LO(o)) + \delta_2(AG(r) \quad (35)$$
$$+AG(o)) + \delta_1(TA(r) + TA(o))$$

where, $\delta_1 + \delta_2 + \delta_3 = 1$. This paper uses the text eigenvector to characterize the text information of books and documents. The similarity between the text information can be calculated by:

$$S(r_\sigma, o_\sigma) = \frac{\sum_{i=1}^{m}(\delta_{ri} + \delta_{oi})}{\sqrt{\sum_{i=1}^{m}\delta_{ri}^2}\sqrt{\sum_{i=1}^{m}\delta_{oi}^2}} \quad (36)$$

The relationship information between nodes can be divided into primary and secondary levels. The relationship information of nodes r and o can be expressed as $RE(r) = \{FO(r), FA(r)\}$, and $RE(o) = \{FO(o), FA(o)\}$, respectively. The similarity analysis of relationship information is equivalent to the similarity analysis of the primary and secondary relationship information. Let η_1 and η_2 be the weights of the similarity between primary relationship information and secondary relationship information, respectively. The similarity between the relationship information can be calculated by:

$$S(r_\gamma, o_\gamma) = \delta_1 S(FO(r), FO(o)) + \delta_2 S(FA(r) + FA(o)) \quad (37)$$

where, $\eta_1 + \eta_2 = 1$. The similarity between the primary relationship information can be calculated by cosine similarity:

$$S(FO(r), FO(o)) = \frac{FO(r) \cdot FO(o)}{\|FO(r)\| \cdot \|FO(o)\|} \quad (38)$$

The similarity between the secondary relationship information can be calculated by cosine similarity:

$$S(FA(r), FA(o)) = \frac{FA(r) \cdot FA(o)}{\|FA(r)\| \cdot \|FA(o)\|} \quad (39)$$

Then, the weights θ were assigned to the three kinds of information BA, TE, and RE. The relationship between the three weights can be expressed as $\theta_\mu + \theta_\sigma + \theta_\gamma = 1$. The similarity between two nodes r and o ca be calculated by:

$$S(r,o) = \delta_\mu S(v_\mu, u_\mu) + \delta_\sigma S(v_\sigma, u_\sigma) + \delta_\gamma S(v_\gamma, u_\gamma) \quad (40)$$

Based on the topology of the L&I visualization network, the similarity between network nodes was analyzed according to the primary and secondary relationships between groups of books and documents. The L&I visualization network can be described as a directed graph $QX(H, K)$, where H and K are the set of nodes and set of edges in the network, respectively. Let B be the path connection matrix of the directed graph; $\Omega = \{SS(r, o)\}^M_{r, o=1}$ be the similarity matrix of nodes r and o, with $SS(r, o)$ be the structural similarity between r and o:

$$SS(r,o) = \sum_{l=1}^{M} RZ(D_{vl}=1) \cdot RZ(D_{ul}=1) \cdot \frac{1}{\log RZ_l} +$$
$$\sum_{l=1}^{M} RZ(D_{lv}=1) \cdot RZ(D_{lu}=1) \cdot \frac{1}{\log UZ_l} \quad (41)$$

where, $1/\log RZ_l$, and $1/\log UZ_l$ are the weights of primary relationship node l and secondary relationship node l,

respectively; RZ_l, and UZ_l be the in-degree and out-degree of node l, respectively.

5. Experiments and Results Analysis

All the associations between L&I nodes in each type of disciplines were captured under the JavaScript programming environment. However, it is impossible to visualize so many types of disciplines one by one. Fig. 4 and Fig. 5 present the L&I visualization networks for science and engineering disciplines, respectively. The light blue and light pink clusters are located in the center of the networks, respectively.

Table 2 presents the descriptive statistics (number of nodes, number of edges, node degree, and mean clustering coefficient) on the primary disciplines of the top 8 disciplines, whose L&I visualization network density is greater than 0.5. It can be seen that, the L&I networks of disciplines with node degree greater than 0.5 have similar features: the networks of science, engineering, medicine, agronomy, and management have a few nodes, but the nodes boast a high node degree, possess a high mean clustering coefficient, and belong to the fully connected state. To better describe the distance between nodes from the planar perspective, Fig. 6 visualizes the L&I networks of different disciplines with node degree smaller than 1. The figure provides a visual display of the cross influence of node distribution density and mean clustering coefficients on the L&I of these disciplines, as well as the status of every secondary discipline network under each primary discipline L&I network. The information in the figure fully demonstrates the superiority of the visualization technology that cannot be reflected by columns of numbers.
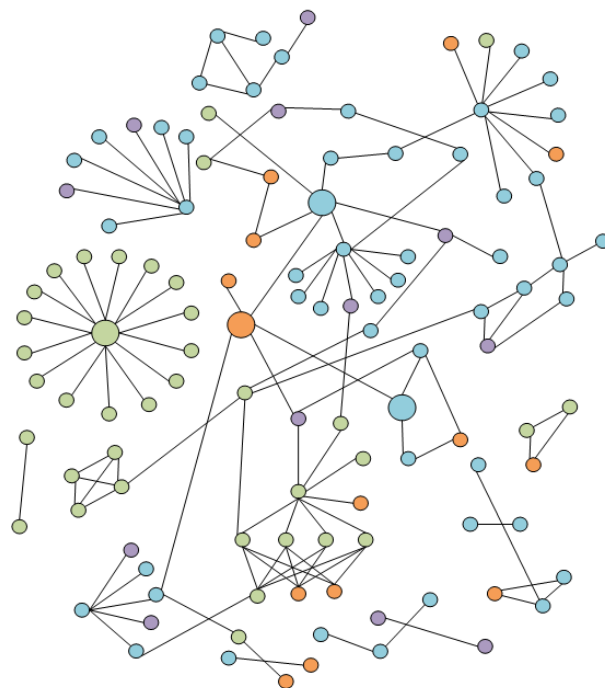

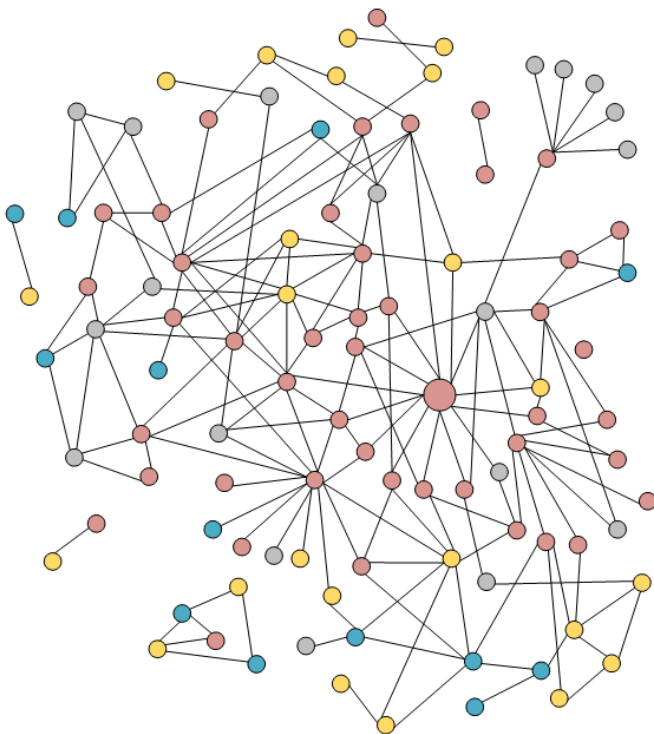Fig. 4 L&I visualization network for science disciplines

Fig. 5 L&I visualization network for engineering disciplines

Table 2. Descriptive statistics on the primary disciplines with node degree greater than 0.5

Serial number	Primary discipline	Number of nodes	Number of edges	Node degree	Mean clustering coefficient
1	Science	7	16	1	1
2	Engineering	5	7	1	1
3	Engineering	6	11	1	1
4	Medicine	7	16	1	1
5	Science	15	79	0.98624531	0.986243138
6	Science	6	8	0.82356213	0.828856904
7	Agronomy	12	37	0.63577895	0.812524782
8	Management	7	9	0.52225542	0.567790876

Table 3. Descriptive statistics on 24 disciplines with node degree smaller than 0.5

Serial number	Name of discipline	Number of nodes	Number of edges	Network density	Mean clustering coefficient
1	Philosophy	25	28	0.46321458	0.69849156
2	Economics	13	131	0.40961562	0.68595231
3	Law	66	129	0.38915615	0.66241782
4	Sociology	65	23	0.30897034	0.72420423
5	Education	12	792	0.30741159	0.60225465
6	Journalism and Communication	19	613	0.29459192	0.41460262
7	History	10	15	0.26489414	0.42163075
8	Art	27	41	0.24563437	0.21784579
9	Math	45	10	0.24419193	0.59712653
10	Physics	58	82	0.24556232	0.72345632
11	Chemical	65	231	0.21849248	0.60245620
12	Astronomy	53	339	0.20113861	0.62162664
13	Geography	15	400	0.19841563	0.49423346
14	Biology	36	238	0.18771669	0.47312304
15	Mechanical engineering	59	16	0.16895613	0.17692613
16	Material science	59	106	0.15489156	0.40465792
17	Electrical engineering	100	305	0.14518965	0.72332643
18	Architecture	45	276	0.13192632	0.61976221
19	Civil Engineering	55	683	0.12231568	0.55343579
20	Hydraulic engineering	61	133	0.12052491	0.53143923
21	Mining engineering	132	184	0.11891562	0.58432634
22	Veterinary science	95	220	0.10247959	0.48323316
23	Clinical medicine	97	931	0.10137563	0.66312978
24	Business management	98	462	0.10112315	0.59463123

Table 3 presents the descriptive statistics on 24 disciplines with network node degree smaller than 0.5. From the statistics on the number of nodes, it can be inferred that this number is negatively correlated with network node degree. There is also a negative correlation between the number of nodes and the sum of the initial cross influence between disciplines. This law can be verified by the number of edges. If a few nodes are connected by many edges, the clustering coefficient and node degree of the visualization network will increase. If many nodes are connected by a few edges, the clustering coefficient and node degree of the visualization network will decrease. This further confirms the cross influence of L&I in terms of discipline.

To measure the influence of discipline integration on the structure of the L&I visualization network, this paper compares the structural evolution of the network before and after discipline integration, and analyzes the dynamic evolution law of the interactive operation on the form of L&I visualization. Fig. 7 presents the structure of the visualization network for each primary discipline of liberal arts.

Table 4. Nodes of visualized paths on different levels of importance

Number of shortest paths	Start points	End points	Nodes on the shortest paths
3	16.723	16.21	16.723, 16.21
3	16.12	16.849	16.12, 16.23, 16.849
4	18.372	16.314	18.372, 17.25, 16.953, 16.314
4	2.135	16.756	2.135, 4.623, 7.968, 15.562, 16.756
5	16.74	16.23	16.74, 17.25, 18.31, 15.26, 17.48, 16.23

5	16.37	16.95	16.37, 18.25, 17.42, 18.93, 17.25, 18.84, 16.95
6	3.82	16.12	3.82, 5.67, 7.24, 9.85, 7.52, 12.79, 14.56, 15.19, 16.12
6	5.94	16.71	5.94, 6.08, 7.67, 9.52, 10.24, 11.75, 12.43, 14.69, 15.72, 16.71
7	12.35	16.45	12.35, 12.75, 13.18, 13.84, 14.23, 15.37, 16.75, 18.92, 18.05, 16.45
7	9.51	16.53	9.51, 10.72, 14.35, 16.21, 18.79, 17.63, 16.54, 18.31, 17.25, 18.35, 16.53

Table 4 displays the nodes of visualized paths on different levels of importance, aiming to verify the effectiveness of path importance on L&I visualization tasks. Considering the calculation results on importance levels and the meta-paths for the typical paths in the table, it is possible to generate the node pairs corresponding to the paths, and then measure the similarity between nodes by the methods specified in Subsection 4.2.

Further, our algorithm is expected to correctly guide the similarity measurement of nodes in information flow query tasks. To verify if our algorithm meets the expectation, this paper compares the experimental data under different weight distributions, similarity metrics, and edge properties. Figure 8 compares the L&I query accuracies under different visualization encoding schemes. There is a total of eight curves in the figure, including the curved edges with no weight and no similarity metric CUR1, the curved edges with no weight yet similarity metric CUR2, the curved edges with weight and no similarity metric CUR3, the curved edges with weight and similarity metric CUR4, the straight edges with no weight and no similarity metric SL1, the straight edges with no weight yet similarity metric SL2, the straight edges with weight and no similarity metric SL3, and the straight edges with weight and similarity metric SL4. It can be observed that, under any encoding scheme, the nodes with similarity metric had higher accuracies than those without similarity metric. Regardless of the presence or absence of weight, the similarity metric and modeling difficulty both greatly affected the L&I query accuracy.
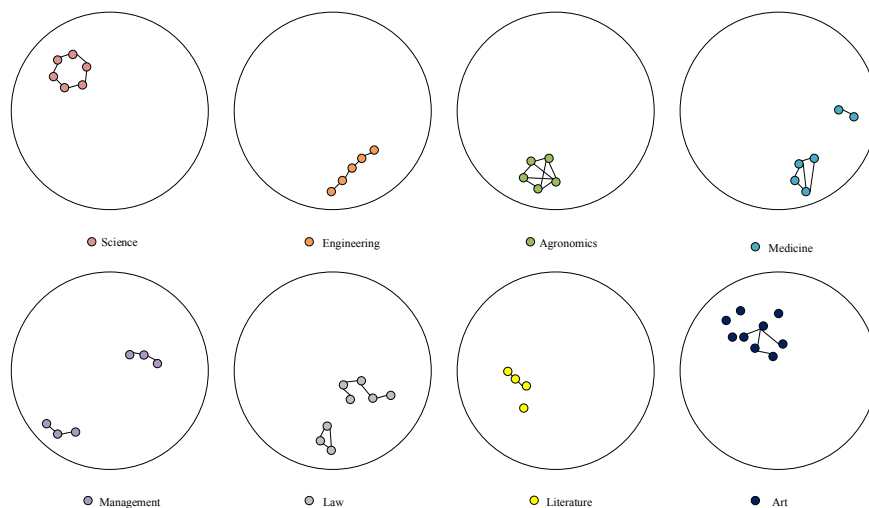

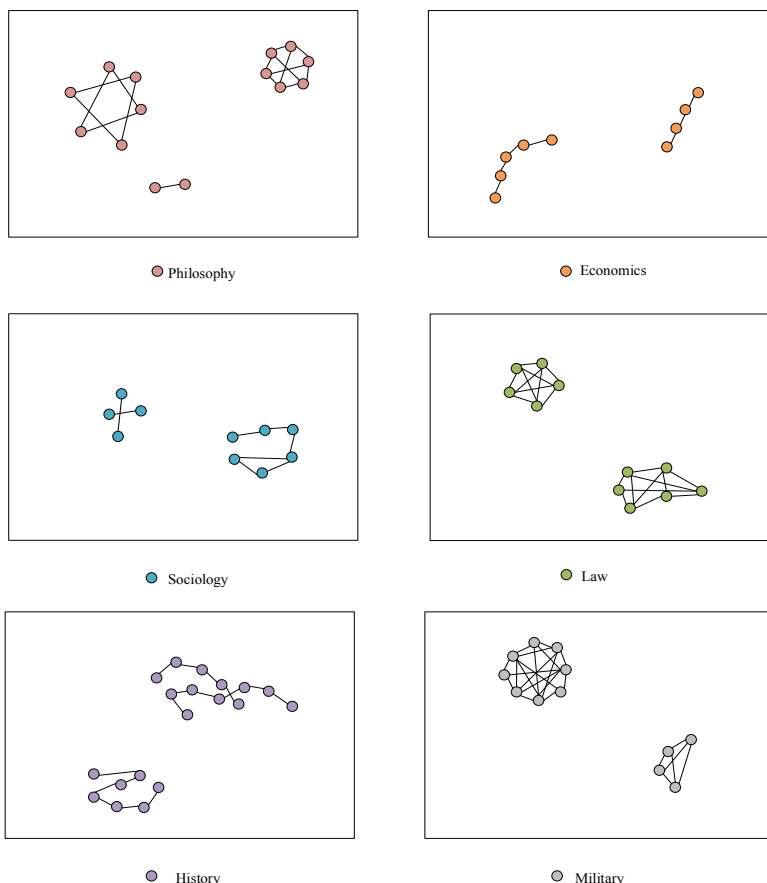Fig. 6 L&I visualization networks of different disciplines with node degree smaller than 1

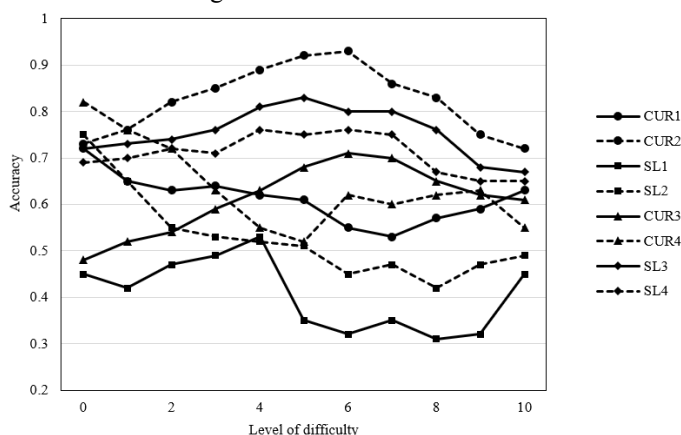Fig. 7 Structural evolution of L&I visualization network before and after discipline integration

Fig. 8 L&I query accuracies under different visualization
encoding schemes

6. Conclusions

This paper probes into the visualization of L&I from the perspective of big data analysis and fusion. The first step is to analyze the topology of the L&I visualization network, and calculate the metrics for the construction of L&I visualization topology map, as well as the importance of meta-paths of the L&I visualization network. After that, a complex big data L&I visualization network was established to analyze the associations between information nodes. Through experiments, the disciplines with network node degree greater or smaller than 0.5 were described statistically, and the structural evolution of the L&I visualization network before and after discipline integration was demonstrated to confirm the effectiveness of our visualization strategy. Further, the nodes of the visualized paths were obtained under different levels of difficulty, which verify that the path importance is effective in L&I visualization tasks. It is also proved that, during the information flow query, the node similarity metric derived by our algorithm can correctly guide the query process.

References

[1] M. A. Islam, N. K. Agarwal, "What is the impact of library and information science on knowledge management research," VINE Journal of Information and Knowledge Management Systems, Vol. 51, No. 1, pp. 1-26, 2020.

[2] H. Alrubaye, M. W. Mkaouer, A. Ouni, "On the use of information retrieval to automate the detection of third-party java library migration at the method level," In 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC), pp. 347-357, 2019.

[3] D. Hienert, D. Kern, K. Boland, B. Zapilko, P. Mutschke, "A digital library for research data and related information in the social sciences," In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 148-157, 2019.

[4] N. Kalenov, I. Sobolevskaya, A. Sotnikov, "Hierarchical representation of information objects in a digital library environment," In Russian Conference on Artificial Intelligence, pp. 93-104, 2019.

[5] B. D. Lund, T. Wang, D. E. White, "University program rankings in the United States: Failures in assessing important factors for library and information science students," Education for Information, Vol. 35, No. 4, pp. 419-438, 2019.

[6] L. Salisbury, A. O. Omolewu, J. J. Smith, "Technology use for non-educational purposes during library instruction: Effects on students learning and retention of information," Science & Technology Libraries, Vol. 37, No. 3, pp. 274-289, 2018.

[7] C. Ruchon, R. Grad, "Evidence reversal: Towards awareness of the phenomenon in library and information science," Education for Information, Vol. 34, No. 4, pp. 301-304, 2018.

[8] C. G. Zhang, D. H. Zha, S. Zhou, H. X. Zhou, H. D. Jiang, "3D visualization of landslide based on close-range photogrammetry," Instrumentation Mesure Métrologie, Vol. 18, No. 5, pp. 479-484, 2019.

[9] H. Chen, J. Yao, Y. Fu, Y. Pang, J. Wang, Y. Huang, "Tagmentation on microbeads: restore long-range DNA sequence information using Next Generation Sequencing with library prepared by surface-immobilized transposomes," ACS applied materials & interfaces, Vol. 10, No. 14, pp. 11539-11545, 2018.

[10] R. Han, D. Shao, L. H. Wang, M. J. Wang, "A novel communication and monitoring system for integrated utility tunnel," Instrumentation Mesure Metrologie, Vol. 18, No. 2, pp. 195-200, 2019.

[11] X. Li, "Young people's information practices in library makerspaces," Journal of the Association for Information Science and Technology, Vol. 72, No. 6, pp. 744-758, 2021.

[12] S. Adhikari, "Method for Intelligent Information Processing Using Software Reuse Library and Blockchain Architecture," In AIAA Scitech 2021 Forum, pp. 1919, 2021.

[13] G. Retscher, A. Leb, "Development of a Navigation and Information Service for a University Library," In Proceedings of the 2021 International Technical Meeting of The Institute of Navigation, pp. 229-242, 2021.

[14] J. Ren, "An Empirical Study of Evaluation of Library WeChat Micro Service by Mobile Information Network Technology," In Journal of Physics: Conference Series, Vol. 1952, No. 3, pp. 032018, 2021.

[15] Q. Xiao, "Resource classification and knowledge aggregation of library and information based on data mining," Ingénierie des Systèmes d'Information, Vol. 25, No. 5, pp. 645-653, 2020.

[16] Y. Mo, E. Seon, G. Park, H. Kim, "Course Analysis of Library and Information Science in Korea," Information, Vol. 11, No. 1, pp. 19, 2020.

[17] E. Rowley, L. Kuryloski, K. Moore, "Extending the Role of the Library and Librarian: Integrating Alternative Information Literacy into the Engineering Curriculum," In 2020 ASEE Virtual Annual Conference Content Access, 2020.

[18] Á. Borrego, "The impact of MOOCs on library and information science education," Education for Information, Vol. 35, No. 2, pp. 87-98, 2019.

[19] Q. Johnson, "Privacy considerations for library and information professionals," Information Services & Use, (Preprint), pp. 1-4, 2020.

[20] C. C. Okeji, O. Mayowa-Adebara, "An evaluation of digital library education in library and information science curriculum in Nigerian universities," Digital Library Perspectives, Vol. 37, No. 1, pp. 91-107, 2020.

[21] F. Peset, F. Garzón- Farinós, L. M. González, X. García- Massó, A. Ferrer- Sapena, J. L. Toca- Herrera, E. A. Sánchez- Pérez, "Survival analysis of author keywords: An application to the library and information sciences area," Journal of the Association for Information Science and Technology, Vol. 71, No. 4, pp. 462-473, 2020.

[22] J. T. Mickel, "Digital Culture and Information: A Library-Based Minor Focusing on How Modern Information Access Impacts Society," In Proceedings of the 20th Annual SIG Conference on Information Technology Education, pp. 151-151, 2019.

[23] J. Griffin, "Leadership to advance data and information science at Virginia Tech library," Information Services & Use, Vol. 39, No. 3, pp. 177-181, 2019.

[24] I. E. Paramonova, "Information Interactions: The Criteria of the Choice of Communication Channels in a Scientific and Technical Library," Scientific and Technical Information Processing, Vol. 46, No. 3, pp. 181-186, 2019.

[25] H. A. Pradana, P. P. G. P. Pertama, M. S. Mayasari, Y. Andrika, F. P. Juniawan, D. Y. Sylfania, "Audit of Provincial Library Information System based on COBIT 4.1," In 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), Vol. 1, pp. 18-22, 2019.

[26] R. Haunschild, L. Leydesdorff, L. Bornmann, "Library and Information Science papers as Topics on Twitter: A network approach to measuring public attention," ISSI, 2019.

[27] G. Retscher, A. Leb, "Development of a smartphone-based university library navigation and information service employing wi-fi location fingerprinting," Sensors, Vol. 21, No. 2, pp. 432, 2021.