A Novel Grouping Method of Learning Community Based on Interests

Yan Cheng and Yongchun Miao

***Abstract*—To solve the problem that the distance education inevitably produced the "isolated" learners, the learners with the same interest are organized into the same community for collaborative learning. In view of neglecting the semantic relevance between terms of the traditional vector space model, the vector space model based on ontology is proposed to calculate the learner's interest eigenvector, and the corresponding explicit express can be obtained according to the recessive expression, which improves the accuracy of the interest similarity comparison. At the same time, a self-organization algorithm based on the similarity match-degree and matching concentration of learner's interest for the learning community is put forward. Great dimensions would take place with the ontology to construct vector space, thus Concept Indexing method is adopted to reasonably reduce the dimensionality of interest characteristic matrix so that greatly reduces the computational complexity. Finally, an experimental analysis of online education cases indicates that the algorithm has high efficiency and good extensibility.**

***Index Terms*—Ontology, vector space model, concept indexing method, interest similarity match-degree.**

I. Introduction

In order to solve the "isolated" learner on account of the distance education inevitably, learners with the same interest should be organized into the same learning community to help them for collaborative learning. Community is a group of people with the common sense, benefit and interest in a certain geographical area. If a network space is substituting for geographical area, it will be called virtual community. Virtual community is composed of the participants with common interests, who interact through the network, find a group of like-minded partners and are able to discuss a certain extent and significance of topics.

Currently, research of the virtual community of education mainly focus on theoretical and practical research exploring the basic principle of virtual community, and subject covers knowledge sharing and management, information resources management, information society, social network, search community, electronic groups, online chat groups and so on. Anita summarized the related research of virtual community [1]. However, more researches of virtual community technology development are tools used in virtual community and its technical potential, and it is hot topic to discuss how to build a virtual community. Jin discussed the design and development of a prototype system of the virtual community based interactive learning environment [2]. Lin *et al*. proposed knowledge map creation and maintenance approaches by utilizing information retrieval and data mining techniques to facilitate knowledge management in virtual communities of practice [3]. Lee *et al*. discussed how to build the core subject vocabularies for community-oriented subject gateways by analyzing their attended three library projects [4]. Zurada *et al*. discussed the issue of building a large-scale virtual organization for individuals and institutions that are associated with the field of computational intelligence (CI) and machine learning (ML) [5]. Vernet *et al*. [6] proposed a new approach of intelligent tutoring of virtual learning communities.

The space model based on vector occupies an important position in current text processing and text mining. It can be used in the comparison of the space of document and paragraph and the comparison between sentences and sentences [7]. Many search engines have used this model to classification of web documents. Due to the lack of what the semantic relevance among terms was neglected for the traditional space model based on vector, many scholars took the semantic correlation among concepts into consideration, such as Latent Semantic Indexing [8]-[10], Linguistic Conceptualization, a method based on the dictionary or classification [11] and so on, but these methods took simple and rare relationship of the concept into consideration. The semantic web technology is used to provide a sharing knowledge background for the information extraction, namely ontology. With the support of it, the semantically related terms in the interest-based descriptive item are no longer regarded as key words in isolation, but have certain semantic relationship each other. The semantic relativity between concepts of the limited ontology graph overcomes its shortcoming, which takes into account that these methods of the latent semantic relation among terms are able to improve the accuracy of the similarity comparison of interests. The space dimension of the ontology-based information extraction system's vector depends on the number of entities in the ontology. With the increase of ontological concepts, the high dimension of space vector brings the difficulty and complexity in calculation. Existing solutions include Random Map (RP) [12], [13], Latent Semantic Analysis (LSA) [14], [15], Concept Index (CI) Dimension Reduction [16] and so on. In the thesis, the improvement of the Concept Index (CI) Dimension Reduction will used to reduce the dimension of the interest-based feature and to reasonably reduce the dimension of the interest-based characteristic matrix, then the matching

Manuscript received September 1, 2015; revised October 22, 2015. This work was supported by National Natural Science Foundation of China (NSFC, Grant No. 61262080), Jiangxi Province Science and Technology Support Major Project (20151BBE50121).

Yan Cheng is with Tongji University. She is also with Jiangxi Normal University, Nanchang, Jiangxi, China (e-mail: chyan88888@jxnu.edu.cn).

Yongchun Miao is with Jiangxi Normal University, School of Computer Information Engineering, Nanchang, China.

calculation is able to greatly decrease the computational complexity and improve the efficiency of the algorithm.

This paper discusses how to build an interest-based learning community and to make learners collaborative learning within the community for learners on the network. A basic idea of building interest-based learning community is to firstly calculate interest-based vectors according to the related data of learners, namely to obtain corresponding explicit show according to implicit show of interest, then to calculate an interest-based matched-degree between learners on the basis of above vectors and a matched-concentration of learners, finally to build a learning community according to a certain method. The article proposes the space model based on the ontological vector through the concrete analysis of the characteristics of the learning community and puts forward a self- organization algorithm of learning community based on similar matched-degree and matched-concentration of learners' interest. Finally example analysis shows that the model algorithm has high efficiency and good extensibility. The following will introduce the relevant concepts and algorithms in detail.

II. Related Work

A. Semantic Distance and Semantic Relativity

The semantic distance is used to evaluate the semantic similarity among objects (concept, vocabulary and sentence), and its value is a real ranging 0 to ∞. The definition of the semantic distance is determined according to the system model, but a definition of it isn't still universal so that everyone can give a definition of the semantic distance according to their needs to the actual application, therefore there are many methods to give a definition of it, such as virtual distance, hierarchical distance, etc. The semantic distance and semantic relativity have an inverse relationship, for example, the longer the semantic distance between two concepts, the lower their semantic relativity, conversely, the shorter the distance, the greater the relativity. Semantic relativity, the concept of a strong subjectivity, is related to the specific application, so it is hard to give a unified definition. Considering that the ontology can be expressed as the limited ontological graph, the semantic distance can be calculated by the shortest path method based on graphics.

There is a one-to-one correspondence between nodes on the limited ontological graph and the classes or concepts in the ontology, so to collection of concepts $O_c = \{c_1, c_2, \cdots, c_n\}$ in the ontology, the function of the semantic distance is a binary mapping $d : O_c \times O_c \to R$, and satisfying the following conditions: (1) non-negativity: $d(c_x, c_y) \geq 0$; (2) identify: $d(c_x, c_y) = 0$, if and only if $c_x \equiv c_y$; (3) symmetry: $d(c_x, c_y) = d(c_y, c_x)$; (4) triangle inequality: $d(c_x, c_y) \leq d(c_x, c_z) + d(c_z, c_y)$. Therefore the space of the semantic distance constitutes a metric space, and d is called a metric function (O_c, d) on the O_c . The matrix of the semantic distance is used to describe the metric space.

Definition 1: To arbitrary node $\boldsymbol{V}(c_x)$ and $\boldsymbol{V}(c_y)$ in the limited ontological graph, if a path exists between them, the value of the semantic distance is equal to the length of the shortest path between two nodes:

$$d(\boldsymbol{V}(c_x), \boldsymbol{V}(c_y)) = \min(pathLength_i(\boldsymbol{V}(c_x), \boldsymbol{V}(c_y)))$$

Definition 2: The value of the semantic distance from any nodes to itself on the graph of the limited ontological space is 0, $d(\boldsymbol{V}(c_x), \boldsymbol{V}(c_y)) = 0$.

Definition 3: To arbitrary node $\boldsymbol{V}(c_x)$ and $\boldsymbol{V}(c_y)$ in the limited ontological graph, the direct connection exists between them $\exists edg(\boldsymbol{V}(c_x), \boldsymbol{V}(c_y))$, and if the relationship between two nodes is a parent-child relationship (*SubClassOf*), the value of the semantic distance is defined as $d(V(c_x), V(c_y)) = 1$; If the relationship between two nodes is an object property relationship (*ObjectProperty*), the value of the semantic distance is defined as $d(\boldsymbol{V}(c_x), \boldsymbol{V}(c_y)) = 2$.

According to the above calculation method based on the semantic distance on graphics, the matrix is used to express the semantic distance among concepts in the ontology, giving the matrix of the semantic distance:

$$\boldsymbol{DisM}(O_c) = \begin{bmatrix} 0 & d(c_1, c_2) & \cdots & d(c_1, c_N) \\ d(c_2, c_1) & 0 & \cdots & d(c_2, c_N) \\ \vdots & \vdots & \vdots & \vdots \\ d(c_N, c_1) & d(c_N, c_2) & \cdots & 0 \end{bmatrix} \quad (1)$$

The dimension N of the matrix of the semantic distance is equal to the number of classes in the ontology, and the element $d(c_i, c_j)$ showed the value of the semantic distance between c_i and c_j . Semantic distance and semantic relativity have an inverse relationship, namely the longer the semantic distance between two concepts, the lower their semantic relativity. The formula calculating semantic relativity based on the semantic distance, $r(c_i, c_j) = e^{-\alpha \cdot d(c_i, c_j)}$, $d(c_i, c_j)$ corresponds to the value of the semantic distance between c_i and c_j in the formula (1), and α show the steepness coefficient. The formula meets the corresponding relationship between the semantic distance and the semantic relativity: (1) when the value of the semantic distance between two objects is equal to 0, the value of the semantic relativity between two objects is equal to 1; (2) when the value of the semantic distance between two objects is equal to infinity, the value of the semantic relativity between two objects is equal to 0;(3) the longer the semantic distance between two objects, the lower their semantic relativity (monotonically decreasing).

According to the formula calculating semantic relativity, the matrix of the semantic distance $\boldsymbol{DisM}(O_c)$ can be converted into the matrix of the semantic relativity $\boldsymbol{R}(O_c)$.

$$\boldsymbol{R}(O_C) = \begin{bmatrix} 1 & r(c_1, c_2) & \cdots & r(c_1, c_N) \\ r(c_2, c_1) & 1 & \cdots & r(c_2, c_N) \\ \vdots & \vdots & \vdots & \vdots \\ r(c_N, c_1) & r(c_N, c_2) & \cdots & 1 \end{bmatrix} \quad (2)$$

B. Ontology

In recent years, Ontology, an originally philosophical

concept, has been used by scholars in the field of artificial intelligence to express the concept of a higher level. Although there is no a precise definition of ontology, but on the connotation of it, researchers' understanding of the ontology is unified, which consider the ontology as conceptual semantics to contribute to communicating between human and machine, namely ontology provides the sharing knowledge explicitly defined in order to let the machine automatic processing and integrate the network information.

Information (query statements based on natural language, web documents, paragraphs, sentences, etc.) on the Internet is made up of the collection of informational objects. $D=\{d_i \mid 1 \leqslant i \leqslant M\}$, M as a total of informational objects. According to the vector space model, d_i a message can be represented by an eigenvector $\boldsymbol{v}(s)=[s_1, s_2, \cdots, s_N]$. Vector s_i corresponds to concepts c_i of the ontology, which express c_i term's weight of some informational object. Eigenvectors expressed all the informational objects constitute the vector space. The dimension of the vector space is equal to the dimension N of the knowledge universe space made up of the ontology.

C. Limited Ontology Graph

The ontology, a structural knowledge, has the natural hierarchy. From the perspective of graph theory, the ontology can be represented as a graph $G=\{V, Edg\}$, the collection of nodes V corresponds to the entity set $\{c_1, c_2, \cdots, c_N\}$ of the ontology, and there is a one-to-one mapping relationship: $v_i \in V \leftrightarrow c_i, i=1,2,\cdots,N$, between nodes and entities, so $V(c_i)$ can be used of the representation of the node v_i in the graph. The edge $edg_i=(u,v)$ use the connection of the node $u \in V$ and node $v \in V$, and in the ontology, edge set corresponds to relationship set $R: edg \leftrightarrow r_i \in R_i, i=1,2,\cdots, N$, therefore $edg(V(c_i), V(c_j))$ can be used to express the edges in the graph. The above graph is defined as ontological graph.

The limited graph of ontology can be extracted by a constraint for ontological graph, such as the constraint of relationship. Putting aside minor problems, the main problems are focused on through an appropriate constraint. The term in the vector space model based on ontology corresponding to concepts in the ontology, its semantic relevance of concepts is evaluated by calculating the distance of terms in the limited ontological graph. Ontology can be expressed as the limited ontological graph which is a directed acyclic graph of topological order and shows a hierarchical structure. Therefore on the basis of the "Breadth -First" algorithm of traversing graphical level, similar to Breadth-first traversal methods, the limited ontological graph is scanned, then the index table of concepts in the ontology is created. The hierarchical relation in the index table of concepts is obvious that the index number of a high-level concept is in front of the low level and connected to the concept of brothers are arranged in together, which simplifies the calculation method of semantic distance, reduce to scan concept graph, is advantageous for the text processing of natural language based on ontology.

III. Ontology-Based Quantificational Design of System

In the paper, the feature vector of learners' interest is calculated based on the vector space model in the ontology, and the quantitative system of the feature vector of user's interests based on the ontology is designed(Fig. 1), which is a kind of calculation model considering the semantic relevance between terms, then according to the learners to describe their interests and on the basis of the learners' common knowledge-ontology, after they are disposed by two function modules of the preprocessing and the feature quantification and the text information is expressed as the feature vector of a unified scale, the correlation between the vectors can be readily calculated. For example, the cosine method is used to carry out the information matching and extraction.

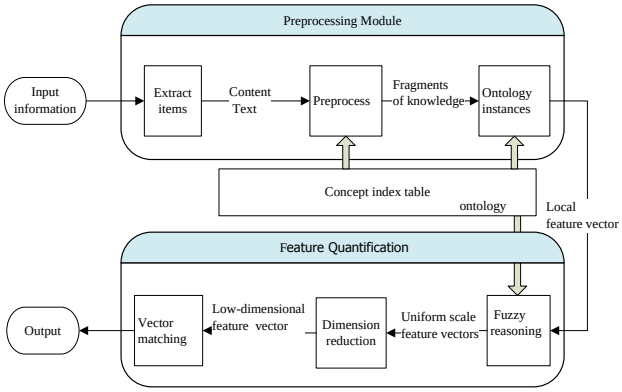

Fig. 1. Designing frame of the quantitative system.

A. Preprocessing Module

The preprocessing module mainly aims to recognize words in a sentence from the descripted information, extract items belonging to the epistemology domain, and reconstruct the ontological instances. Users' personal knowledge background and the understanding of knowledge are potentially included in the interest information released the dispersive learners, so the local knowledge of dispersive users can be extracted from the descripted information. Fragments of these local knowledge may be regarded as instances of ontology, and the ontology depicts the domain knowledge and can be expressed as the ontological graph, so the users' local knowledge of the ontological instances also can be given by the local concept map.

Definition 4: According to the vector space model based on the ontology, $\boldsymbol{v}(s)=[s_1, s_2, \cdots, s_N]$ expresses users' local feature vector, and s_i corresponds to concept c_i of the ontology:

$$s_i = \begin{cases} e^{-\alpha Dis(e_i, root)} \\ 0 \end{cases} \tag{3}$$

$Dis(e_i, root)$ shows the value of the semantic distance from $e_i \in O_c$ to root in the local concept graph, $s_i \in [0,1]$, and the shortest path method based on graphics is used ; α is the

gradient estimation and $-7/\mathrm{MAX}(\mathrm{Dis})$ is often chosen as its value in the fuzzy modeling due to $e^{-7} \approx 0$.

Specific steps of reconstruct the ontological instances: 1) with the concept of ontology as terms of the vector space, terms contained in the information are extracted. If an extracted concept from the ontology can't find the corresponding item, but it is a part of some compound words in the ontology, all corresponding compound words should be added. 2) Local concept map, consist of extracted concepts, shows the user's knowledge fragments, as an instance of the ontology, and has the same features with limited ontology graph, namely, they are a directed acyclic graph. Reconstruction of ontology instances follows the rules as: a collection of knowledge which an item of information contains is regarded as the root; the relative relationship between extracted concepts is consistent with the relative relationship of ontology graph, i.e. hyponymy and offshoot relationship remain unchanged. Hypothetically, using terms of learners of the community with common domain knowledge is random, but potential structural relationship between terms should match the relative relationship between concepts of the domain knowledge. Fragments of knowledge need to be extracted, reconstructed, and ontology instances are saved to the knowledge database. Results which preprocessing module outputs is the input of feature quantization module that is core module of the quantitative system, and then the feature vector of the local knowledge can be gotten.

B. Concepts Index (CI) Dimension Reduction

Ontology-based information extraction system adopts the vector space model to express the information text feature of interest and show the huge dimension, which leads to complexity of the process in computation. Therefore, dimension reduction need to be done for the feature matrix. Concept Index (CI) is an effective dimension reduction method: $\boldsymbol{v}(s)=[s_1, s_2, \cdots, s_N]$ represents a local knowledge feature vector of learners' interest. W is set an $M \times N$ matrix, which M is the number of learners, and N is the dimension of knowledge domain space constituted by the ontology, thus, $\boldsymbol{v}_1(s)=[s_{11}, s_{12}, \cdots, s_{1N}]$, $\boldsymbol{W}=[\boldsymbol{v}_1(s), \boldsymbol{v}_2(s), \cdots, \boldsymbol{v}_M(s)]^T$, and $\boldsymbol{W}_i$ expresses an interest vector space of the i-th behavior of the i-th learner in the $\boldsymbol{W}$, and then set r is the number of dimensions to be reduced. Firstly, concept index adopts some simple clustering algorithm (k-means or hierarchical algorithm, etc.) and execute r times clustering for the $\boldsymbol{W}$, and the feature vector set is divided into r disjoint subsets $s_1, s_2, \cdots s_r$. Then, centroid vector $\boldsymbol{C}_i$ is calculated for each set s_i, and they are normalized to a unit vector $\boldsymbol{C}_i^{'}$. If each unit vector $\boldsymbol{C}_i^{'}$ is as a coordinate axis of space after the dimension reduction, an r-dimensional subspace will be gotten. The vector of each learner can be mapped to obtain its r-dimensional subspace; therefore, interest vector space $\boldsymbol{W}^{'}$ after the dimension reduction is obtained by the matrix operation of the formula 7:

$$\boldsymbol{W}_{m\times r}^{'}=\boldsymbol{W}_{m\times n}\times\boldsymbol{C}_{n\times r} \tag{4}$$

wherein, $\boldsymbol{W}_{m\times r}^{'}$ expresses a matrix of local feature vector of user interest after the dimension reduction, and i-th row represents the local feature vector of the i-th user' interest. The dimension of local feature vectors reduced from N to r, which can greatly simplify the calculation.

C. Information Unified Metrics

From the local knowledge graph (ontology instance graph), the randomness and subjective cognition of terms of learners lead the term relationship of local knowledge to depend on the user's relative relationship. Thus the obtained feature vector is relative to the quantization of individual local knowledge structure. On the other hand, concepts of the same field are not isolated, but have a certain semantic relevance, and the used concepts of ontology instances often implicitly contain the information of the conjoint concept. To be able to uniformly measure dispersed feature vectors, local knowledge needs to be transformed into a unified metric space. The fuzzy reasoning is used in the paper, based on the correlation between concepts, and local feature vectors are transformed into feature vectors with the unified metric scale for subsequent matching comparison. The formula is:

$$\boldsymbol{V}(s)\circ\boldsymbol{R}(o)\Rightarrow\overline{\boldsymbol{V}}(s) \tag{5}$$

$\boldsymbol{V}(s)$ represents a feature vector of local knowledge, and is calculated according to the formula (3). Fuzzy relation matrix $\boldsymbol{R}(o)$ has been calculated by the formula (2). Each element r_{ij} of $\boldsymbol{R}(o)$ reflects the degree of similarity between terms. So, the semantic relationship of those terms that are not explicitly appeared in the description of interest will also be considered. In the formula, " o " represents the inner product of fuzzy relations, and represents *MAX-MIN* operation. Dispersed feature vectors are transformed into feature vectors with the unified metric space via the fuzzy reasoning.

$$s_i=Max(\min(s_1, r(1,i)), \min(s_2, r(2,i)), \cdots, \min(s_N, r(N,i))) \tag{6}$$

The value of s_i, an element of feature vector $\bar{v}(s)$ with unified metric space, is obtained by the above formula (6). After the above calculation, descriptive information from the personal interest of learners has been transformed into feature vectors with a unified measure after considering the semantic. Feature vectors with a uniform scale can be matched. Matching can be achieved by calculating the similarity of semantic feature vectors. Reference vector space model, current methods that calculate vectors' similarity include: Inner Product method (Cosine function), Nearest Neighbor method (Max), Minkowski distance, Euclidean distance and so on. One of the most intuitive and effective method is the inner product.

The feature vector of user 1 is represented as the space vector $\vec{\boldsymbol{v}}(d_1)=\{x_{11}, x_{12}, \cdots, x_{1n}\}$, and the feature vector of user 2 is represented as the space vector $\vec{\boldsymbol{v}}(d_2)=\{x_{21}, x_{22}, \cdots, x_{1n}\}$. Then closeness between vectors is measured by calculating the angle between $\vec{\boldsymbol{v}}(d_1)$ and $\vec{\boldsymbol{v}}(d_2)$ (inner product).The larger

the value of the inner product is, the smaller the angle between two vectors is. Similarity is calculated as follows:

$$sim(d_1, d_2) = \vec{v}(d_1) \cdot \vec{v}(d_2) \Big/ \left|\vec{v}(d_1)\right| \left|\vec{v}(d_2)\right|$$

The similarity of feature vectors is calculated based on the inner product method (Cosine). The interest feature vector $\vec{v}_i = [s_{i1}, s_{i2}, \cdots, s_{iN}]$ of user i, and $\vec{v}_j = [s_{j1}, s_{j2}, \cdots, s_{jN}]$ of user j, the match-degree between vectors is calculated by the formula (6) according to the Cosine method:

$$MatchDegree(i, j) = \left|\vec{v}_i \times \vec{v}_j\right| \Big/ \left|\vec{v}_i\right| \left|\vec{v}_j\right| \quad (7)$$

Closeness between vectors is measured by the inner product of $\vec{v}_i$ and $\vec{v}_j$, and larger the value of the inner product is, the smaller the angle between two vectors is, so the closer two users' interest is.

IV. Self-organization Algorithm for Learning Community

The community should be a group of learners with the same or similar interests, so learners with great similarity should be divided into the same community when establish the virtual learning community. According to the guidance of the idea, the process is proposed to build the learning community as follows:

Step 1: Interest match-degree between every two learners is calculated by the formula (7) after the dimension reduction of interest feature vectors.

Step 2: A threshold value T_1 is set, and calculated the matching concentration of each learner (the level of concentration is evaluated by the number of learners whose interest match-degree is higher than the learner). The number of a university network learner is supposed m, and the concentration θ_i of learner i is calculated by the formula $\theta_i = \sum_{i=1}^{m} a_{ij} / m$, wherein $a_{ij} = \begin{cases} 1 & MatchDegree(i, j) \geq T_1 \\ 0 & otherwise \end{cases}$ and T_1 is a presupposed threshold value.

Step 3: The highest of learners' matching concentration is selected as a center learner for the community. T_2 is a presupposed threshold value, and if interest match-degree between the center learner and him is higher than T_2, the learner will be divided into the community.

Step 4: To remaining learners, the value is recalculated according to the order of the first step to the third step until the maximum matching concentration of learners is lower than T_0. For the learner whose matching concentration is lower than T_0, the similarity between him and every center learners of community is calculated and he is assigned into the closest learning community.

V. Case Analysis

This paper presents an ontology-based self-organization algorithm of the learning community. Learners with similar hobbies automatically and effectively are organized into the corresponding learning communities, which is able to share resources, and collaboratively learn. Experimental results demonstrate that the algorithm has high efficiency and good grouping scalability.

A. Experimental Procedure of Algorithm

1) According to the semantic distance method, semantic distance between concepts of ontology can be expressed by the semantic matrix.

The domain ontology of programmers is established based on expert support, which provides shared knowledge background for distributive learners in the community network. Considering the depth of the concept hierarchy tree and density of semantic distance, for two nodes with the same path length, if it located at the lower conceptual level, the node' semantic distance should be smaller, and if it located at the higher density area of the concept hierarchy tree, the node' the semantic distance should less than in the lower density area. Therefore, the shorter the distance from a conceptual level to the root is, the greater the weight of distance is. Domain knowledge is encoded, and a figure (Fig. 2) can be gotten with concept of the ontology.

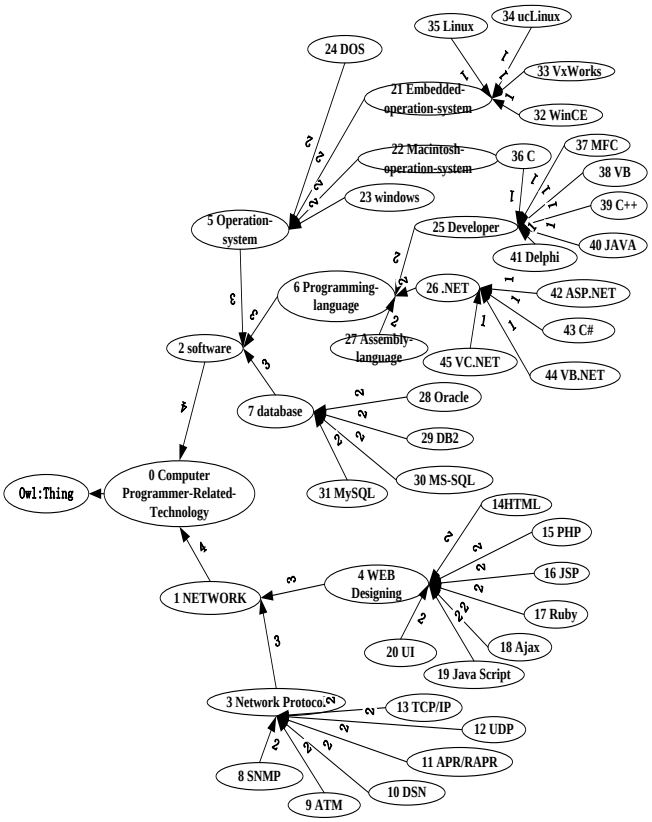

Fig. 2. Domain ontology graph of programmers.

From the ontology diagram (Fig. 2), knowledge can be hierarchically represented, and there is some semantic relevance between concepts of ontology. Ontology concepts are showed by the number, and limited ontology graph is a directed acyclic graph and is analyzed according to the Breadth-First algorithm with the level traversing. The index table with corresponding concept ontology graph in Fig.2 is represented as follows: {0 Computer Programmer-Related -Technology,|1 NETWORK, 2 software,|3 Network Protocol,4 WEB Designing,5 Operation-system,6

Programming-language,7 database,|8 SNMP,9 ATM,10 DSN,11 APR/RAPR,12 UDP,13 TCP/IP,14 HTML,15 PHP,16 JSP,17 Ruby,18 Ajax,19 Java Script,20 UI,21 Embedded-operation-system, 22 Macintosh-operation-system,23 windows,24 DOS,25 Developer,26 .NET,27 Assembly-language,28 Oracle,29 DB2,30 MS-SQL,31 MySQL,|32 WinCE,33 VxWorks,34 ucLinux,35 Linux,36 C,37 MFC,38 VB,39 C++,40 JAVA,41 Delphi,42 ASP.NET,43 C#,44 VB.NET,45 VC.NET}.

Vertical lines in the index list indicate the level classification of the concept of the limited ontology graph. Here, the label is not randomly assigned corresponding to concepts, but is closely related to hierarchical arrangement of concepts, label in the graph corresponding to serial number of the index list. Considering the ontology can be expressed as limited ontology graph, and then the semantic distance will be calculated by the shortest path method based on the graph. According to the definition of the semantic distance matrix, The 46 × 46 semantic distance matrix can be obtained (Fig. 3).

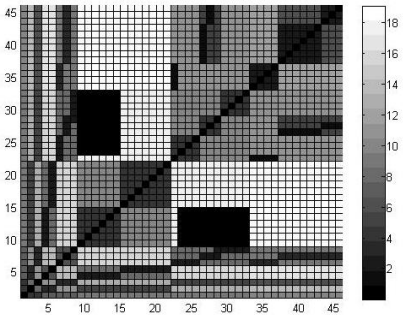

Fig. 3. Semantic distance matrix graph of domain ontology of programmers.

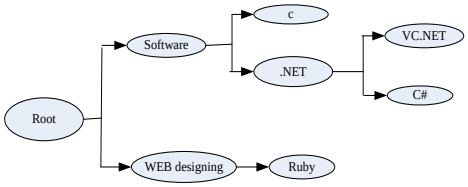

Fig. 4. Ontology concept graph of learner *A*.

In the Fig. 3, the semantic distance matrix is a symmetric matrix, and gray areas measure the size of the semantic distance between entities. The label of *X*, and *Y* axis corresponds to serial number of the index list of ontology concepts.

2) Local knowledge of learners' interest information is quantified, and is transformed into feature vectors with a unified metric after fuzzy reasoning.

For example, "*Software Engineer, experiences in C language,.NET such as VC.NET, C#, interested in WEB designing with Ruby*" is an item of interest information described by learner *A*, "*Software,C,.NET, VC.NET, C#, WEB designing, Ruby*" terms can be extracted from knowledge space via the method mentioned in the 3.1 section. According to the relative structure of terms in the knowledge space, the upwardly extracted knowledge fragments can be represented as a concept graph (Fig. 4). "*Operation-system*, *Embedded-operation-system*, *Linux*, *Database*, *MySQL*" terms are again extracted in the knowledge space from an interest item described by learner *B*, adopting the same method to get its concept graph (Fig. 5).

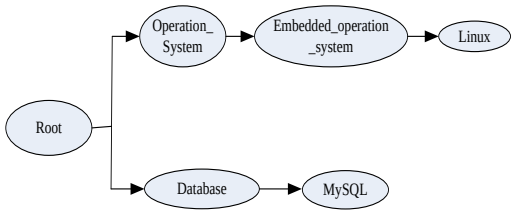

Fig. 5. Ontology concept graph of learner *B*.

Local knowledge of learner *A*' and learner *B*' interest information is quantified as local feature vectors as follows:

$$v_1(s):\left[0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-3\alpha},0,e^{-3\alpha}\right]$$

$$v_2(s):\left[0,0,0,0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots e^{-2\alpha},0,0,0,e^{-3\alpha},0,\cdots\right]$$

Elements of the feature vector correspond to items of the concept index list. *M*=12 learners want to find the most appropriate learning community via interests they have described. Similarly, Interest feature terms can be extracted from interest information described by learners, and reconstruct them, then get local knowledge vectors of interest information items from other learners' description as follows:

$$v_3(s):\left[0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,0,0,0,0\right]$$

$$v_4(s):\left[0,0,0,\ e^{-\alpha},e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-3\alpha},0,\cdots,0\right]$$

$$v_5(s):\left[0,0,e^{-\alpha},0,\cdots,0,e^{-\alpha},e^{-\alpha},0,\cdots,e^{-2\alpha},e^{-2\alpha},0,\cdots,0\right]$$

……

3) Through the fuzzy reasoning, the learners' local knowledge is transformed into unified metric feature vectors. Then, the CI method can be adopted to get lower- dimensional interest feature vectors on the basis of the dimension set by the user.

Hypothetically, the dimension *r* equals 3, and there are *M* = 12 learners. Therefore, their interest local knowledge vectors can be expressed as an *M* × *N* matrix (*M* is the number of learners and *M* equals 12. *N* is the dimension of knowledge domain space that is composed of the ontology, and *N* equals 46), namely, $\boldsymbol{W}'_{12\times3}=\boldsymbol{W}_{12\times46}\times\boldsymbol{C}_{46\times3}$ on the basis of the formula (4).

Wherein, $\boldsymbol{W}'_{12\times3}$ is a matrix of learners' interest local feature vectors after the dimension reduction, and the *i*-th row represents an interest local feature vector of the *i*-th learner. Dimension of local feature vectors matrix is reduced from *N* = 46 to *r* = 3, which is able to greatly simplify the calculation.

4) The learning community is automatically built according to the similarity between the feature vectors. Interest match-degree between *M* = 12 learners are calculated by the formula (6) and form a matrix as follows:

user1	1	0.22933	0.22933	0.22933	0.22933	0.22933	0.22933	0.69647	0.64377	0.65392	0.46811	0.22933
user2	0.22933	1	0.29816	0.24005	0.69939	0.69939	0.29816	0.22933	0.22933	0.22933	0.22933	0.69939
user3	0.22933	0.29816	1	0.24005	0.29816	0.29816	0.57877	0.22933	0.22933	0.22933	0.22933	0.29816
user4	0.22933	0.24005	0.24005	1	0.24005	0.24005	0.24005	0.22933	0.22933	0.22933	0.22933	0.24005
user5	0.22933	0.69939	0.29816	0.24005	1	0.70074	0.29816	0.22933	0.22933	0.22933	0.22933	0.70074
user6	0.22933	0.69939	0.29816	0.24005	0.70074	1	0.29816	0.22933	0.22933	0.22933	0.22933	0.74846
user7	0.22933	0.29816	0.57877	0.24005	0.29816	0.29816	1	0.22933	0.22933	0.22933	0.22933	0.29816
user8	0.69647	0.22933	0.22933	0.22933	0.22933	0.22933	0.22933	1	0.64377	0.65392	0.46811	0.22933
user9	0.64377	0.22933	0.22933	0.22933	0.22933	0.22933	0.22933	0.64377	1	0.64377	0.46811	0.22933
user10	0.65392	0.22933	0.22933	0.22933	0.22933	0.22933	0.22933	0.65392	0.64377	1	0.46811	0.22933
user11	0.46811	0.22933	0.22933	0.22933	0.22933	0.22933	0.22933	0.46811	0.46811	0.46811	1	0.22933
user12	0.22933	0.69939	0.29816	0.24005	0.70074	0.74846	0.29816	0.22933	0.22933	0.22933	0.22933	1

Threshold value is set 0.3. For the fuzzy matrix, the fuzzy

coefficient is greater than a given threshold value, and it is set to 1, but is set to 0,

$$\begin{array}{c} user1 \\ user2 \\ user3 \\ user4 \\ user5 \\ user6 \\ user7 \\ user8 \\ user9 \\ user10 \\ user11 \\ user12 \end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Finally, learning communities are built via the matching concentration method mentioned in the 4 section. User1, user8, user9, user10, user11 are assigned into the community 1. User2, user5, user6, user12 are assigned into the community 2. User3, user5, user7 are assigned into the community 3. User4 are assigned into the community 4.

B. Result Analysis

1) When reduce dimension of feature vectors, the smaller the value of r is, the smaller the number N of corresponding to concepts is, the smaller the complexity of the algorithm is. Dimension reduction via CI method, the computational complexity will be directly affected by the dimension user sets.
2) The threshold value T_1 is set according to the interest match-degree between learners, calculate every learners' matching concentration and find a learner with the most 1 in the matrix (i.e., the highest concentration), and then he or she is as a center to build the learning community. If the threshold is too small, the number of communities will be small, and the number of each community learners will be more, which may make learners with different interests into the same communities. On the contrary, if the threshold is too big, the number of communities will be more, which may make learners with the same interests into the different communities. Therefore, the threshold should be set according to past experience, and the threshold should not be too large or too small.
3) Ontology-based vector space model takes into the dependency between concepts account. In the paper, students from Tongji University Network Education College as experimental subjects, experiments showed that the ontology-based self-organization algorithm of learning community has higher precision and full rate. Precision and full rate is calculated by the formula as follows:

$$\text{Precision} = M/N \qquad \text{Full rate} = N/L$$

wherein, M is the right number of learners in the community; N is total learners in the community; L is total learners participate in the test.

The method with artificial judgment to return items is used to calculate precision. For example, for 20 students to build learning communities, right items that accord with students' hope is n through artificial judgment, precision=n/20. For the algorithm in the paper, the number M of learners is set as 20, 50, and 100 in the test. When M=20, the average precision equals 83%; When M=50, the average precision equals 95%; When M=100, the average precision equals 98%. Therefore, experiments show that the more experimental subjects is, the higher precision is. For the learner whose matching concentration is lower than T_0, the similarity between him and every center learners of community is calculated and he is assigned into the closest learning community. So, full rate of the algorithm is nearly 100%. Students can enter into their most suitable learning communities according to their interests.

VI. Conclusion

Ontology-based vector space model, a calculation model considering the semantic relevance between terms, is used to calculate learner's interest feature vector. On the basis of the ontology with learners' common knowledge background, implicit interest described by learners is transformed into the corresponding explicit representation (i.e. interest vectors). The semantic relativity between concepts is calculated in the limited ontology graph, which overcomes a shortcoming that the semantic relativity between terms is neglected in the traditional vector space modals, and improves the precision of interest similarity comparison. A self-organization algorithm of learning community based on similar matched-degree and matched-concentration of learners' interest is proposed according to the idea of interest similarity. Dimension of the vector space that is constructed by ontology concepts in the ontology-based vector space model is huge, so CI method is improved so that it is applied to dimension reduction of interest feature matrix. Dimension of interest feature matrix is reasonably reduced, and then matching calculation is executed, which greatly decreased the amount of computation and computational complexity and improve the efficiency of algorithm.

References

[1] A. L. Blanchard, "Testing a model of sense of virtual community," *Computers in Human Behavior*, vol. 24, no. 5, pp. 2107-2123, Sep. 2008.
[2] Q. Jin, "Design of a virtual community based interactive learning environment," *Information Sciences*, vol. 140, no. 1–2, pp. 171-191, Jan. 2002.
[3] 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, March 2006.
[4] W. Lee and S. Sugimoto, "Toward core subject vocabularies for community-oriented subject gateways," *China New Technology of Library and Information Service*, no. 01, pp. 25-32+82, Feb. 2006.
[5] J. M. Zurada, M. A. Mazurowski, R. Ragade, A. Abdullin, J. Wojtudiak, and J. Gentle, "Building virtual community in computational intelligence and machine learning," *Computational Intelligence Magazine, IEEE*, vol. 4, no. 1, pp. 43, 54, Feb. 2009.
[6] D. Vernet, X. Canaleta, and G. Pallas, "Intelligent tutoring of virtual learning communities," *International Symposium on Computers in Education (SIIE)*, pp. 1, 6, 29-31, Oct. 2012.
[7] B. Y. Ricardo and R. N. Berthier, *Modern Information Retrieval*, New York: ACM Press, ch. 5, 1999.
[8] T. A. Letsche and M. W. Berry, "Large-scale information retrieval with latent semantic indexing," *Information Sciences*, vol. 100, no. 1–4, pp. 105-137, Aug. 1997.
[9] T. Jaber, A. Amira, and P. Milligan, "Enhanced approach for latent semantic indexing using wavelet transform," *Image Processing, IET* , vol. 6, no. 9, pp. 1236-1245, Dec. 2012.
[10] D. Thorleuchter and D. V. D. Poel, "Quantitative cross impact analysis with latent semantic indexing," *Expert Systems with Applications*, vol. 41, no. 2, pp. 406-411, Feb. 2014.

[11] J. Kalina, "Classification methods for high-dimensional genetic data," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 10-18, 2014.
[12] N. Ailon and B. Chazelle, "Faster dimension reduction," *Communications of the ACM*, vol. 53, no. 2, pp. 97-104, Feb. 2010.
[13] A. P. Verbyla, J. D. Taylor, and K. L. Verbyla, "RWGAIM: An efficient high-dimensional random whole genome average (QTL) interval mapping approach," *Genetics Research*, vol. 94, no. 6, pp. 291-306, Dec. 2012.
[14] B. Hendrickson, "Latent semantic analysis and fiedler retrieval," *Linear Algebra and its Applications*, vol. 421, no. 2–3, pp. 345-355, March 2007.
[15] J. C. Valle-Lisboa and E. Mizraji, "The uncovering of hidden structures by latent semantic analysis," *Information Sciences*, vol. 177, no. 19, pp. 4122-4147, Oct. 2007.
[16] A. Artemiou and B. Li, "Predictive power of principal components for single-index model and sufficient dimension reduction," *Journal of Multivariate Analysis*, vol. 119, pp. 176-184, Aug. 2013.

Yan Cheng was in February, 1976 in Wu Yuan City. She got the bachelor's degree in 1998, the master's degree in 2005, and the Ph.D. degree in 2010 from Tongji University in Shanghai and she is now a postdoctor of Tongji University in 2015, majored in computer science, engaged in e-learning, virtual learning community and educational technology innovation. From July 2013 to March 2014, she went to the California University as a visiting scholar. Cheng's main research interests include intelligent computer aided education, educational data mining, virtual learning community and e-learning.

Currently she is a postdoctor of Tongji University and a professor in Jiangxi Normal University in Nanchang, Jiangxi province. As a researcher, currently she is the leader of 2 national NSFC (Natural Science Foundation of China) Plan Project and more than 5 provincial projects on IT and education innovations. She is the author of more than 30 scientific papers, won Guanghua Ph.D. scholarship, 3 talks in the global intelligent automation conference, 1 national doctoral academic best paper award. She published 1 academic monograph (Beijing, Science Press, 2014) on educational data mining, 5 computer professional teaching materials.
Dr. Cheng is the senior member of the Systems Engineering Association of China, the review expert of Computer Application Study, one member of Chinese Intelligent Automation, the consulting evaluation expert of small and medium-sized enterprise in Jiangxi province. Dr. Cheng is also the direction leader of the ministry of education "software engineering" discipline and the expert in education informationization.

Yongchun Miao received the B.Sc. degree in software engineering from Jiangxi Normal University, Nanchang, in 2013. Her mains research interests include educational data mining, complex social networks, images analysis and processing and virtual learning community.