Knowledge Map Creation and Maintenance for Virtual Communities of Practice

Fu-ren Lin, Chih-ming Hsueh
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan 804
frlin@cc.nsysu.edu.tw

Abstract

This paper proposes knowledge map creation and maintenance approaches by utilizing information retrieval and data mining techniques to facilitate knowledge management in virtual communities of practice. Besides evaluating their performance using synthesized data, the generated knowledge maps for documents collected from the teachers' cyber community, SCTNet, and the master thesis repository at Taiwan's National Central Library, are evaluated by domain experts. Domain experts are asked to revise the obtained knowledge maps, and the proportion of modification is small and acceptable. Therefore, the developed approaches are suitable for support knowledge management of professional communities on the Internet.

1. Introduction

With rapid development of Internet technology, professionals can communicate with others via Internet regardless geographical distance, so that abundant of professional knowledge is kept in virtual communities. Knowledge stored in an organizational information system enables the knowledge sharing process including organization knowledge categorization. However, a virtual community with people from different organizations usually lack of such type of managerial authority. Therefore, the knowledge sharing process in a virtual community needs more technical support to deal with the asynchronous addition of knowledge artifacts and maintain consistent knowledge structure.

In general, explicit knowledge can be expressed in the form of words, speech, reports, etc. In order to shorten learning cycle, an individual can exploit the experience of others to enlarge his or her experiences by sharing explicit knowledge on the Internet. Besides, how knowledge is useful is not only the knowledge itself. How to reveal the relationship among knowledge is truly important [7][18]. Therefore, it is important to create the knowledge map to specify the relationship between knowledge artifacts in order to facilitate learning in virtual communities. In this research, knowledge map is defined as the categorization of documents characterized by concepts contributed to communities of practice. In fact, in a knowledge map, document categories are built explicitly to represent concept hierarchy.

In this study, we develop knowledge map creation and maintenance methods to facilitate knowledge management in virtual communities of practice. We evaluate the proposed knowledge map creation and maintenance methods using documents from the SCTNet, a teachers' professional community (http://sctnet.edu.tw), and the thesis repository at the Taiwan's National Central Library.

We introduce related literatures as background knowledge in Section 2. In Section 3 and 4, we design the knowledge map creation and maintenance functions with the information retrieval techniques respectively. In Section 5 and 6, we evaluate the performance of knowledge map creation and maintenance techniques respectively. Section 7 concludes this study.

2. Literature Review

2.1 Knowledge management for community of practice

One of the emerging phenomena on the Internet era is the emergence of virtual communities. The development of Internet will march from telework, computer supported cooperative work (CSCW), virtual corporation, virtual community to tele-democracy [9]. A virtual community is defined as the gathering of people with common interests to share information and coordinate their works via information technologies,

Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)
0-7695-1874-5/03 $17.00 © 2002 IEEE

specifically for transaction, interest, fantasy, and relationship as indicated in [2]. The rise of information technology (IT) helps the growth of knowledge management to codify, store, and disseminate knowledge**.**

Nonaka [17] proposes the SECI model, which asserts that knowledge creation is a spiral process of interactions between explicit and tacit knowledge through socialization, externalization, combination, and internalization. Based on the SECI model, Lin and Lin [14] proposes a virtual organizational learning model to illustrate how a transactive memory system supports inter-organizational knowledge management. A transactive memory system is composed of three components: *knowledge map, social network,* and *mnemonic function*. Knowledge map represents knowledge objects and their dependency. Social network links individuals and specifies the strength of their relationship. Mnemonic functions perform knowledge allocation, social network updating, knowledge maintenance, and collaborative knowledge retrieval.

A teachers' professional community, called Smart Creative Teachers Network (SCTNet, http://sctnet.edu.tw), has been established for teachers from different schools to share knowledge and collaborate works. SCTNet supported by knowledge map and social network management systems is aiming to facilitate the knowledge sharing process to activate the double-loop learning cycle in the inter-organizational context.

2.2 Techniques for knowledge map management

Knowledge map is viewed by various perspectives from decision making, education, and information retrieval [4][18][7]. Adopting the definition from information retrieval perspective, knowledge map is defined as the categorization of documents characterized by concepts contributed to communities of practice. Knowledge map links various concepts from shared documents contributed by community members. Specifically, document categories are built in the knowledge map to represent concept hierarchy, where learning paths can be traversed associated with the problem solving process.

Since community members contribute documents without pre-specified categories, the creation and maintenance of document categories work in the bottom-up fashion, which needs techniques from fields such as information extraction and document clustering. The following subsections overview these related techniques.

2.2.1 Information extraction

Information extraction techniques can be used for extracting essential concepts to represent unstructured or semi-structured text, e.g., document. Information retrieval has been developed for transforming unstructured or semi-structured text to structured data over twenty years. Although the development of information retrieval in western language is gradually mature, research on retrieving information from documents in Chinese still has difficulties in morphological, syntactic, and semantic levels [26]. Techniques used for transforming documents in Chinese include *word segmentation*, *text indexing*, and *feature selection*.

In the data transformation phase, the first step is word segmentation, which divides each article into the minimum units, words. This is considered the major barrier to text retrieval, especially for Asian languages [6][27][21]. Prior research of word segmentation is summarized into three categories: *dictionary* [13][5], *linguistic* [27], and *statistical approaches* [24][11][3][8][16][5].

After allocating word segmentations for all documents, the next task is to identify key terms which represent the main concepts of each document. The process of differentiability assignment to each document is known as feature selection in machine learning, and is equivalent to the weighting system in information retrieval. In the term-weighting system, it uses the occurrence of each keyword in documents as clue to assess the differentiability of each keyword. In the information retrieval research, the fully weighted system $\frac{tf \cdot \log \frac{N}{n}}{\sqrt{\sum_{vector} (tf_i \cdot \log \frac{N}{n_i})^2}}$ obtains the best effects among the term-weighting systems [21].

2.2.2 Document clustering

The recent researches in the information science are tended to facilitate the tasks of the automatic documents categorization. However, the clustering task is more challenging than the documents categorization, since there is no pre-existing set of categories created by human experts. Decision making in the clustering task includes not only assigning documents to categories but also determining the number of clusters [20]. Document representation and clustering technique are two major issues in clustering documents. The vector space model is

usually adopted to represent documents, where a document is represented as a multidimensional vector, and each dimension corresponds to a unique key term.

Clustering techniques can be classified as partitioning and hierarchical clustering approaches [10]. Two-stage clustering combines the partitioning and hierarchical clustering [19]. The partitioning clustering outperforms the hierarchical clustering if the parameters for the partition are not generated randomly. *K*-means clustering technique may have better noise tolerance, and its performance is independent from the distance measure. Another clustering technique is self-organizing map (SOM) neural network. SOM neural network automatically organizes the documents onto a two-dimensional grid, and similar documents are projected close to each others [12][15][23].

3. Knowledge Map Creation

It is difficult to have the well pre-defined knowledge structure to classify knowledge in an autonomous virtual community. Knowledge map creation in a virtual community requires the bottom-up document categorization based on documents' key terms. Information extraction techniques can be used for transforming the unstructured documents into the structured data, and data mining techniques can be applied to discover the relationships among documents. These techniques include text indexing, keyword extraction, term-weighting, and document clustering.

(1) *Text Indexing.* Unstructured documents in semi-infinite strings (*sistring*) are first indexed in a PAT-Tree to enable the efficient access of data source, where the occurrence of words is calculated.

(2) *Keyword Extraction.* In this stage, a hybrid keyword extraction method is employed. The combination of mutual information and natural language processing method attempts to obtain more concise extracted keywords. First, we use the *auto_tag* program developed by CKIP projects in Academia Sinica to identify the morphological information of each word phrases, and prune all words but noun phrases. Then, the keyword extraction algorithm proposed by [5] is employed to extract keywords from continuous noun phrases. A statistics, *AE*, is responsible to judge whether a word is a keyword or not. After this stage, unstructured document is transformed into structured data, and many value analysis methods can be applied to the transformed data.

(3) *Term Weighting.* Representative keywords for each document are identified by the term-weighting system, where normalized *TF* is responsible for this task. Besides, the *DF* is used as a threshold to prune those non-representative keywords.

(4) *Clustering.* Clustering techniques are used for specifying document relationships based on features extracted from documents. In this study, we use two-stage clustering to implement the knowledge map creation and knowledge map maintenance functions.

The hierarchical clustering expresses the hierarchical relationships of documents. In order to identify the relationship of the concepts extracted from knowledge artifacts, such as documents, we adopt the two-stage clustering approach as the core technique. In the first stage, the preliminary binary hierarchical relationships are explored by the hierarchical clustering. From the preliminary structure, we determine the best number of clusters to replace the binary structure. In the second stage, the *k*-means clustering is employed to physically partition documents into the best number of clusters. The two-stage clustering procedure is described in Figure 1.

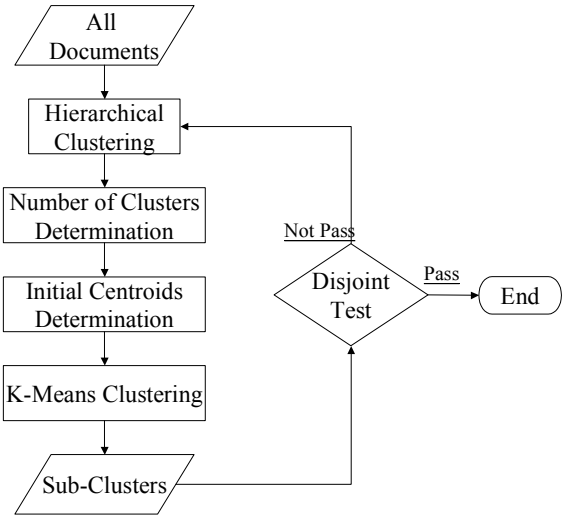

Figure 1. The procedure of knowledge map creation

3.1 Stage 1 (Hierarchical clustering)

We firstly apply the hierarchical clustering to obtain the hierarchical relationship of documents. The binary structure is the nature of the hierarchical clustering method and the binary knowledge structure limitedly supports the knowledge navigation activity because of its complex category structure. The complexity of the binary structure generated from the hierarchical clustering can be reduced by choosing the cutting threshold to determine the best umber of clusters in

order to perform the physical partitioning with *k*-means clustering. The best number of clusters is determined by the silhouette coefficient (*sc*) [10].

Silhouette coefficient is defined as follows. Assume that the cluster to which object *i* is assigned is denoted as *A*. Let *a*(*i*) be the average dissimilarity of *i* to all other objects of cluster *A*. For any cluster *C* different from *A*, let *d(i,C)* be the average dissimilarity of *i* to all objects of *C*. After computing *d(i,C)* for all clusters *C*, the smallest one among them denoted as $b(i) = \min_{C \neq A} d(i, C)$ is selected. The silhouette coefficient of object *i*, *s*(*i*), is then obtained by combining *a*(*i*) and *b*(*i*) as follows:

$$s(i) = 1 - \frac{a(i)}{b(i)}, \quad if\ a(i) < b(i)$$
$$= 0 \quad , \quad if\ a(i) = b(i) .$$
$$= \frac{b(i)}{a(i)} - 1, \quad if\ a(i) > b(i)$$

Thus, we can choose the best number of clusters by calculating the silhouette coefficient of each number of clusters. It shows that higher the silhouette coefficient value stronger the cluster structure.

3.2 Stage 2 (*k*-means clustering)

In general, the *k*-means clustering method obtains much better the performance than hierarchical clustering in most of the cases if the number of clusters and initial seeds are chosen properly. After calculating silhouette coefficient at Stage 1, we obtain initial seeds from the resulting clusters of the hierarchical clustering. We then use the *k*-means clustering method to physically partition the document set starting with these initial seeds. To construct the knowledge map with hierarchical relationship among concepts, the structure discovery procedure should be performed iteratively between Stage 1 and Stage 2 in each layer. If documents in a cluster belong to one concept, we say that these documents are similar and their distances are relatively small; that is, the cohesion of this cluster is high (*i.e.*, the disjoint of the cluster is low). Thus, the disjoint test on a cluster will identify if the cluster contains sub-clusters. For any two objects *i* and *j* in cluster *A*, let *d*(*i, j*) be the dissimilarity between object *i* and *j*. The disjoint of cluster *A* is defined as

$$D_A = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} d(o_i, o_j)}{N(N-1)/2}$$, where *N* denotes the number of objects in cluster *A*.

The disjoint of cluster *A,* denoted as D_A, represents the average distance within cluster *A*. A high C_A value indicates that documents in cluster *A* are highly similar. Otherwise, it implies that documents in cluster *A* are quite different.

4. Knowledge Map Maintenance

Members of a virtual community contribute their experiences continuously and the knowledge map of the community may vary dynamically. Therefore, how to allocate the new incoming documents in the existing knowledge map is very important, especially in the virtual community, which lacks of the managerial authority embedded in a physical organization. The function of knowledge map maintenance is designed to refresh the knowledge map by modifying categories in a small range of knowledge map. Besides reducing the computational cost of re-clustering the whole knowledge map, the incremental knowledge map maintenance approach will greatly reduce the community members' cognitive loading on adapting the new knowledge structure. The major steps of the knowledge map maintenance approach are depicted in Figure 2, and elaborated in the following subsections.

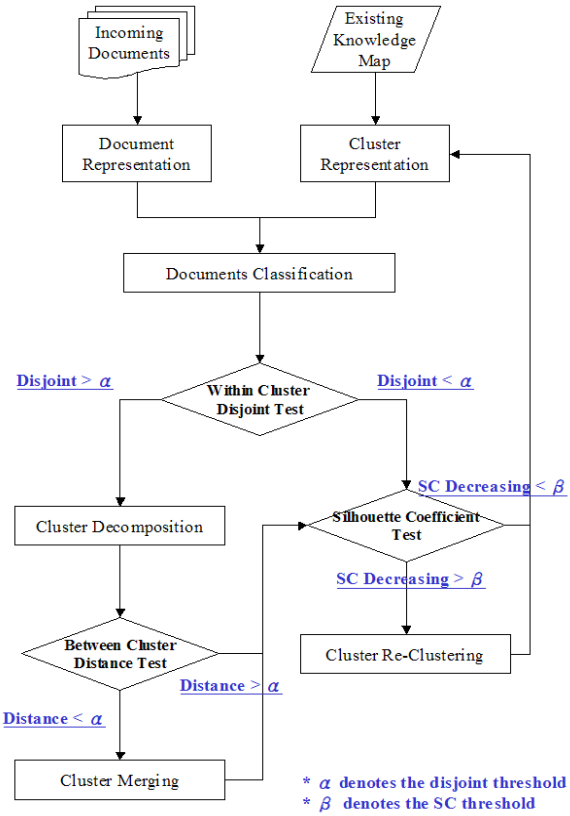

Figure 2. The procedure of knowledge map maintenance

4.1 Document and cluster representation

A cluster resulting from the knowledge map creation is represented by the vector space model, and the centroid vector of a cluster is the average of vector values of objects in the cluster. An incoming document is represented by the vector space model, and it is classified into a cluster by calculating the distance between the document vector and cluster centroid vector. In the vector space model, a vector values denotes the occurrences of keywords in a document, and the centroid vector of a cluster denotes the average occurrences of keyword within the cluster. In this step, both the new documents and existing clusters are transformed into vector space representation, and both have equivalent dimensions. Thus, we can compare the similarity between a new document and the existing clusters in keywords.

4.2 Document classification

In this step, we assign a new document into the existing clusters by comparing the document vector with the cluster centroid vector. By calculating the Euclidean distance between the two vectors, we obtain the dissimilarity between the new document and the cluster. The document is inserted into the cluster between it having the smallest distance.

After the incoming document is classified into the most similar cluster, the disjoint test within cluster is performed. The disjoint within cluster is defined as the average distance between documents within a cluster. If documents within a cluster are similar, the distance between documents is relatively small, which results in low disjoint value. Suppose that the disjoint threshold is set to α. If the result of disjoint test is greater than α, the similarity of documents within the cluster is low, and the cluster is split into sub-clusters. If the disjoint is smaller than α, the similarity within the cluster is high, and no further decomposition is needed.

4.3 Cluster decomposition and merging

If the disjoint is greater than α, the two-stage clustering is applied to decompose the cluster, and the number of sub-clusters is determined by silhouette coefficient. The cluster decomposition will replace the original cluster with the new sub-clusters regardless the hierarchical relationship of the original cluster and new sub-clusters.

New sub-clusters generated in the cluster decomposition phase will be compared with the existing clusters in distance. If the distance between a new sub-cluster and an existing cluster is the smallest and within the disjoint threshold α, these two clusters will be merged into a single cluster. The merged clusters may originally belong to different super-clusters. The candidate locations to insert the new merged clusters are chosen from these super-clusters. For example, the distance between the existing cluster C_1 and the new sub-cluster C_2 is within the distance threshold α, so that they can be merged into the same cluster C_{new}, where C_1 and C_2 belong to the original super-clusters C_{S1} and C_{S2} respectively. After the merging action, we can choose two locations, C_{S1} and C_{S2}, to be the final mounting point of the new cluster C_{new}. In deciding the mounting point, we compare the distance of the new cluster centroid vector $< C_{new}>$ with the original super-cluster centroid vector $<C_{S1}>$ and $<C_{S2}>$, and the closest super-cluster will be chosen as the final mounting point. This indicates that the content of the new clusters are much similar to the content of the closest super-cluster.

4.4 Re-clustering

As mentioned above, *sc* is used for determining the number of clusters. In this step, we again test *sc* of the given super-cluster. Comparing the original *sc* with the new *sc*, if the *sc* decreasing rate is greater than a given threshold β, the number of sub-clusters of the given super-cluster is inaccurate after updating the cluster structure. Therefore, we should perform the re-clustering action to discover the new knowledge map in the super-cluster. If the *sc* decreasing rate is smaller than a given threshold β, the cluster structure of the given super-cluster is still acceptable, and no further re-clustering action is needed. If the *sc* decreasing rate is larger than a given threshold β, we will perform the re-clustering action to the whole branches of knowledge map to find the better structure. Re-clustering knowledge map follows the creation procedure. In the re-clustering process, the knowledge structure is demolished, and then documents originally containing in sub-clusters are gathered and re-clustered.

5. Evaluation of the Knowledge Map Creation

We conducted experiments to evaluate the quality of the created knowledge map by inviting sixteen domain experts to revise knowledge map generated by the knowledge map creation system.

5.1 Document sets

The document sets we used for experiments are from two domains: natural science in elementary education and information management in master thesis abstracts. There are 254 natural science documents (NSD) (419,196 Chinese characters), including teaching materials and research articles, chosen from the SCTNet. There are 281 master thesis abstracts (267,288 Chinese characters) of departments of information management (TAIM) chosen from the thesis repository of the National Central Library.

5.2 Evaluation criteria

In Chinese information retrieval research, there are few standard test corpuses as the evaluation benchmarks. Thus, we adopt the evaluation method proposed in [20], which conducts experiments to evaluate the outcomes by invited domain experts. By comparing the result between the automatic clustering and experts' revision, the quality of the knowledge map is evaluated. Precision and recall are two measures to evaluate the quality of the built knowledge map. We define precision and recall as $precision = \frac{N_{A \cap M}}{N_A}$ and $recall = \frac{N_{A \cap M}}{N_M}$, where A denotes the document set containing in a cluster generated by the knowledge map creation method, M denotes the document set in a cluster modified by domain experts, N_A denotes the number of documents in document set A, N_M denotes the number of documents in document set M, and $N_{A \cap M}$ denotes the number of documents both in A and M.

5.3 Experimental design

Sixteen subjects as domain experts were asked to evaluate the knowledge map generated from documents in the two corresponding domains in the same experimental settings. Eight graduate students from the department of information management of a national university in Taiwan evaluated the knowledge map generated from TAIM document set. Eight elementary school teachers from two elementary schools in a Taiwan city evaluated the knowledge map created from NSD document set in nature science.

Although silhouette coefficient is employed to choose the number of clusters in the process of document categorization, we do not expect large number of clusters to express the hierarchical relationship among concepts. Therefore, three strategies of determining the number of clusters were adopted and three different knowledge maps are generated. The first strategy, called "deep", decides the number of clusters ranging from 1 to 10 in choosing the maximum silhouette coefficient. In the deep strategy, we expect to obtain a smaller number of clusters in each layer and the concept hierarchical tree looks deep. The second strategy, called "wide", chooses the maximum silhouette coefficient in setting the number of clusters ranging from 10 to 20. In the wide strategy, we expect to obtain a larger number of clusters in each layer, and the concept hierarchical tree looks wide. The third strategy, called "mix", chooses the number of clusters ranging from 1 to 20.

Each subject spent 40 minutes as one section to finish the revision of one type of knowledge map generated by one strategy, deep, wide, or mix, and the coordination between subjects was not allowed. There is ten minute break between two sections. In total, three sections for one document set took about two hours. A subject verified and modified the knowledge map on a desktop computer along with the printed document abstracts which he or she can reference. Meanwhile, we collected the reference materials after each section to prevent the side-effects on the next experiment. We analyzed the differences of knowledge maps before and after subjects' revision.

5.4 Evaluation results

In the experiments, precision and recall defined in Subsection 5.2 are chosen as the evaluation criteria. The precision and recall of the TAIM document set are shown in Table 1 and the evaluation of the NSD document set is shown in Table 2. We found that the precision and recall in these evaluation sessions are very high (91% ~ 93%). However, we found that the modified knowledge maps among different experts vary greatly. It is insightful to investigate the degree of consensus of experts among different experts. We modify precision and recall measures to reflect the consensus of knowledge maps. The modified precision and recall are defined as $\frac{N_{U_1 \cap U_2 \ldots U_n}}{N_A}$ and $\frac{N_{U_1 \cap U_2 \cap \ldots \cap U_n}}{N_{U_1 \cup U_2 \cup \ldots \cup U_n}}$ respectively, where A denotes the document set containing in a cluster by the knowledge map creation method, U_i denotes the document set containing in a cluster modified by domain experts i, N_A denotes the number of documents in document set A, $N_{U_1 \cup U_2 \ldots \cup U_n}$ denotes the number of distinct documents in the cluster modified by expert *1, 2, ..., n,*

and $N_{U_1 \cap U_2 \ldots \cap U_n}$ denotes the number of common documents in the cluster modified by expert *1, 2, …, n.*

If we replace the numerator of the original precision with the union of documents in a cluster, which modified by eight human experts, we obtain the modified precision 0.79 for wide, 0.75 for deep, and 0.72 for mix knowledge maps in TAIM document set. The modified precision denotes the consensus to the correct relationships of documents in a knowledge map.

If we replace the denominator of the modified precision with the union of the documents in a cluster, which modified by eight human experts, we obtain the modified recall 0.66 for wide, 0.60 for deep, and 0.58 for mix knowledge maps in TAIM document set. The modified precision of NSD document set is 0.61 for wide, 0.59 for deep, and 0.60 for mix respectively. The modified recall of NSD document set is 0.72 for wide, 0.73 for deep, and 0.68 for mix respectively.

Table 1. Precision/recall of TAIM document set

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Average
Wide	0.94/0.94	0.95/0.95	0.94/0.95	0.95/0.95	0.96/0.96	0.94/0.94	0.89/0.88	0.91/0.91	0.93/0.93
Deep	0.92/0.92	0.93/0.93	0.94/0.94	0.93/0.94	0.97/0.96	0.88/0.91	0.87/0.87	0.87/0.88	0.91/0.92
Mix	0.95/0.96	0.97/0.97	0.90/0.91	0.98/0.98	0.96/0.94	0.90/0.89	0.82/0.82	0.92/0.91	0.93/0.92

Table 2. Precision/recall of NSD document set

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Average
Wide	0.90/0.91	0.85/0.85	0.91/0.91	0.90/0.92	0.93/0.93	0.95/0.94	0.92/0.92	0.92/0.92	0.91/0.91
Deep	0.86/0.88	0.91/0.88	0.91/0.93	0.97/0.97	0.96/0.97	0.98/0.98	0.91/0.92	0.92/0.91	0.93/0.93
Mix	0.95/0.95	0.91/0.88	0.89/0.91	0.98/0.98	0.98/0.98	1.00/1.00	0.91/0.91	0.93/0.91	0.95/0.94

From the evaluation results of modified precision and recall, the common consensus evaluation of the TAIM and the NSD document sets is much lower than those using original precision and recall measures. It implies that the bottom-up knowledge map discovery method enacts a high degree of freedom in category allocation, which is more difficult to reach a consensus opinion in categorization.

One-way ANOVA test was conducted to test if the results of three types of knowledge maps, wide, deep, and mix, have significant differences in precision and recall. The result shows that there are no significant differences between these three types of knowledge maps either in the NSD or TAIM document sets. In summary, the knowledge map creation function can be implemented for various knowledge map structures, and its performance in terms of precision and recall is encouraging.

6. Evaluation of the Knowledge Map Maintenance

We use synthesized and real world data to obtain thorough understanding of the proposed maintenance method. A document set with a high dimension of keyword vector, *e.g.*, NSD or TAIM in this study, takes much time in the computation. It is more efficient to conduct experiments on documents with only 24 keywords than taking the real world documents. Besides the efficiency issue, we can tune parameters to obtain more comprehensive understanding of the proposed method. In this section, the NSD and TAIM document sets are evaluated.

6.1 Evaluation criteria

Three performance criteria, purity, diversity, and specificity are applied to evaluate the status of knowledge map after executing the knowledge map maintenance process [1][25]. They are defined as follows.

(1) $Purity = \sum_{i=1}^{m} Purity(i) \times \frac{N_i^U}{N}$, where *purity*(*i*) is defined as $\frac{n_i^U}{N_i^U}$, and N_i^U denotes the total number of documents in the updated cluster *i*, n_i^U denotes the maximum number of documents that belong to the same category in the updated cluster *i*, and *N* denotes the number of documents.

(2) $\text{Diversity} = \frac{t_U}{T_O}$, where T_O denotes the number

of original category, and t_U denotes the number of true categories covered by the updated categories.

(3) $Specificity = \frac{t_U}{T_U}$, where T_U denotes the number of updated categories, and t_U denotes the number of true categories covered by the updated categories.

In this study, we aim to discover the hierarchical relationship among documents. However, the criteria described above do not reflect the structural change between original and updated knowledge maps. Therefore, we define the structure adaptation indicator (SAI) to express the structural change. SAI is defined as $\mathrm{SAI} = \sum_{l=1}^{L} \frac{1}{\log l} \left| N_l^O - N_l^U \right|$, where l denotes the level number in the concept hierarchy. At the root node, l is 1, and as the layer descends, the level number increments. A leaf node has level number L. N_l^O denotes the number of clusters containing in level l in the original knowledge map, and N_l^U denotes the number of clusters containing in level l in the updated knowledge map. With the SAI design, $\frac{1}{\log l}$ enables the structural differences in top layers to have big effects on the SAI value. It means that the structural difference occurring at the top level of (*i.e.* small level number) results in high SAI value.

6.2 Experimental Design

At first, 210 documents in the synthetic document set are used to generate the category label of each document. Secondly, cross validation method is applied by taking one-third of documents (*i.e.*, 70 documents) as the newly added documents. We take out one-third of the documents and cluster the remaining documents to obtain the knowledge structure. Then, we insert the one-third documents into the initial knowledge map and evaluate the purity, diversity, specificity, and SAI of the resulting clusters. The resulting knowledge map is compared with the knowledge map generated from the total documents, *i.e.*, 210 documents.

In the evaluation, we use two disjoint threshold α values 0.40 and 0.50, two merging threshold α values 0.40 and 0.50, and two re-clustering threshold β values –20% and –100%. A synthetic document uses five keywords as the category pattern and the remaining keywords are randomly assigned value ranging from 1 to 10. These remaining keywords are treated as noises, and will decrease the purity of the synthetic documents. To generate different noise levels, the values of the remaining keywords are assigned with probability of 30%, 40%, and 50% respectively, resulting three different document sets with three different noise levels, 30%, 40%, and 50%. Thus, four experimental settings are listed as follows: setting *A* with α=0.40 and β=-20%, *B* with α=0.50 and β= –20%, *C* with α=0.40 and β= –100%, and *D* with α=0.50 and β= –100%.

6.3 Evaluation results from synthetic documents

The performance of the knowledge map maintenance is evaluated by four criteria, purity, diversity, specificity, and SAI, as shown in Figure 3, 4, 5, and 6. From the evaluation results, the knowledge map maintenance function performs less efficiently and effectively as the noise level of document sets increases. In contrast of two α threshold level, 0.40 and 0.50, we found that the strict threshold of within cluster disjoint test, such as setting *A* and *C* achieved better performance in terms of purity and diversity with noise level 40% and 50% during the knowledge map maintenance. We found that experiments with α=0.40 produced more categories, and had higher purity and diversity than experiments with α=0.50. On the contrary, the experiment setting with α=0.40 results in lower specificity than α=0.50 because it generates much more categories, and the strict merging criterion sustains a high purity within a category. The SAI increases greatly as the degree of noise increases as shown in Figure 6.

The knowledge map maintenance is not a global optimal categorization method, so that the structure in general is quite different from the original knowledge map. The re-clustering action is the most costly action to modify the knowledge map, thus we are reluctant to perform the re-clustering action frequently. From the evaluation results, the re-clustering action performs at most 8.33 out of 70 times of the document insertion. It shows that the probability of the re-clustering action having been triggered is at most 12%. The number of documents affected by the re-clustering action is about 22.73 out of 210 documents.

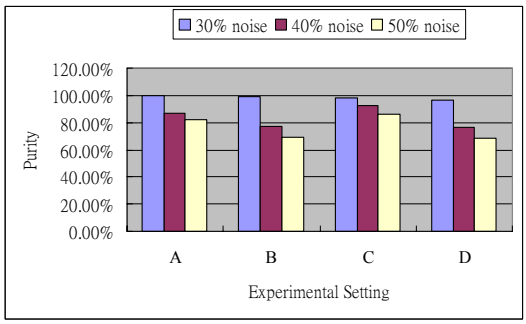

Figure 3. Evaluation of purity

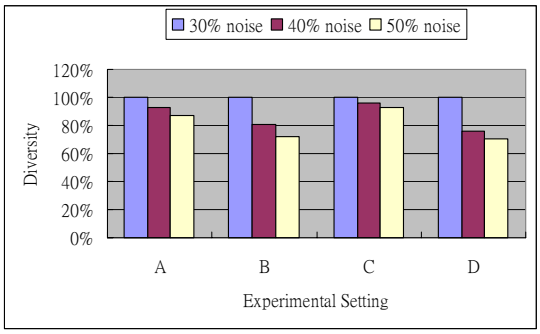

Figure 4. Evaluation of diversity

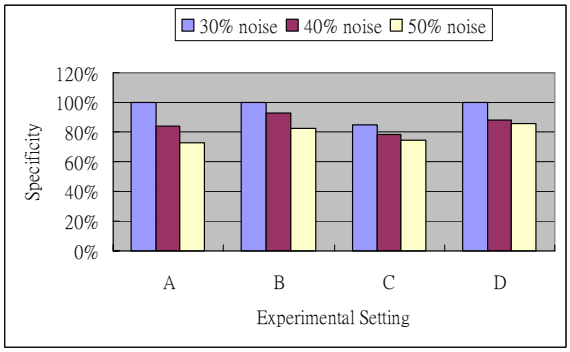

Figure 5. Evaluation of specificity

6.4 Evaluation results from real world documents

We follow the same procedure to evaluate the performance of knowledge map maintenance function for the NSD and TAIM document sets with α=0.84 and β=-10%. The performance in purity, diversity, and specificity, the incremental knowledge map updating process achieves highly acceptable performance as shown in Table 3.

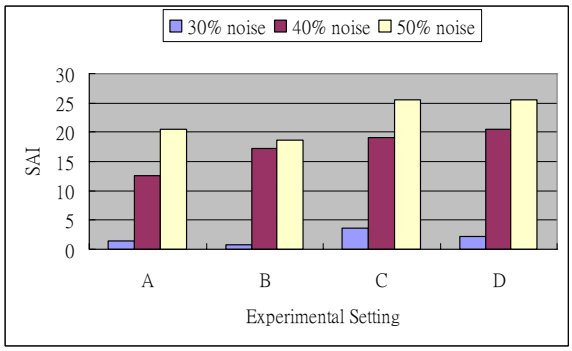

Figure 6. Evaluation of SAI

The incremental document insertion action used for updating the knowledge map presents a high degree of structural difference from that by the one-shot knowledge map creation. The knowledge map maintenance approach modifies the knowledge structure, which is affected by the insertion of new documents. It is not a global optimal approach to discover the knowledge structure. The resulting knowledge map generated by the knowledge map maintenance function will be naturally different from the global optimal approach. However, it is a tradeoff between the knowledge map quality and computational complexity. In the knowledge map management system, it is desirable to have an efficient knowledge map maintenance process to deal with the random insertion documents. In summary, the proposed knowledge map maintenance approach achieves the acceptable purity, diversity, and specificity. The computationally costly re-clustering action is rarely triggered; therefore, the incremental knowledge map updating process can be performed in a real time basis. The re-clustering of the whole documents can be triggered periodically depending on the frequency and quantity of new arrival documents.

Table 3. Evaluation results of the real-world document sets

	Purity	Diversity	Specificity	SAI	# of re-clustering
NSD (α=0.84,β=-10%)	64.56%	74.28%	81.25%	55.20	4 (15.75)*
TAIM (α=0.84,β=-10%)	75.08%	72.13%	88%	38.88	2 (12.5)

* 4(15.75) means that 4 times of re-clustering occur, and 15.75 documents in average are re-categorized.

7. Conclusions

In this paper, we have developed knowledge map creation and maintenance techniques to keep document categories up-to-date in order to facilitate the knowledge sharing activities in a virtual community. The knowledge map creation function employs the automatic categorization technique to discover the knowledge structure from documents in a bottom-up fashion. The knowledge map maintenance function helps the community manager to update knowledge structure efficiently.

In evaluating the knowledge map creation performance, sixteen subjects as domain experts evaluated the knowledge maps generated from two document sets in precision and recall aspects. The

results show that the proposed knowledge map creation method achieves 91%~ 93% precision and recall rate. The modified precision and recall measures to evaluate the consensus of knowledge maps from different subjects obtain relatively low rates. These results indicate that, individually, these experts generally accept the proposed bottom-up knowledge map discovery method, but their changes on the knowledge maps vary due to individual different in domain backgrounds. People are used to classify knowledge in specific knowledge structure based on their educational background, and they may not align with the knowledge map created by the proposed method. It also explains the challenges of the bottom-up knowledge map discovery method to fit accurately to individual knowledge structures.

The incremental knowledge map maintenance approach obtains about 86.14% purity with respect to the knowledge map formed by the one-shot knowledge map creation. Although the knowledge map maintenance may also perform the re-clustering action to modify the knowledge structure, from the experimental results, merely 12% of document insertions trigger the re-clustering action, which affects at most 22.73 out of 210 documents. Thus, the knowledge map maintenance approach is acceptable in considering knowledge map quality and computation complexity of knowledge structure modification.

References

[1] R. Agrawal,, R. Bayardo, and R. Srikant, "Athena: Mining-based Interactive Management of Text Databases," *Proceedings of the Seventh Conference on Extending Database Technology*, July 1999.

[2] A. Armstrong, and J. Hagel, "The Real Value of On-Line Community", *Harvard Business Review*, May-June 1996.

[3] R. Baeza-Yates, and G. Gonnet, "Fast Text Searching for Regular Expressions or Automaton Searching on Tries," *Journal of the ACM* (43:6), 1996, pp. 915-936.

[4] G. Browne, S. Curley, and P. Benson, "Evoking Information in Probability Assessment: Knowledge Maps and Reasoning-Based Directed Questions," *Managements Science* (43:1), 1997, pp. 1-14.

[5] L.-f. Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," *Proceedings of the 1997 ACM SIGIR*, Philadelphia, PA, USA, pp. 50-58.

[6] L.-f. Chien, and H.-t. Pu, "Important Issues on Chinese Information Retrieval," *Computational Linguistics and Chinese Language Processing* (1:1), 1996, pp. 205-221.

[7] C. Chou, and H. Lin, "The Effects of Navigation Map Types and Cognitive Styles on Learners' Performance in a Computer-Networked Hypertext Learning System," *Journal of Educational Multimedia and Hypermedia* (7), 1998, pp. 151-176.

[8] G.H. Gonnet, and R. Baeza-Yates, "New Indices for Text: Pat Trees and Pat Arrays," *Information Retrieval Data Structures and Algorithms*, Prentice Hall, pp. 66-82.

[9] L. Kaufman, and P. J. Rousseeuw, *Finding Group in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, Inc., 1990.

[10] D. E. Knuth, *The Art of Computer Programming: Sorting and Searching*, Vol. 3. Addison-Wesley, Mass., 1973.

[11] T. Kohonen, S. Kaski, K. Lagus, J. Salojvi, V. Paatero, and A. Saarela, "Self Organization of a Massive Document Collection," *IEEE Transactions on Neural Networks* (11:3), May 2000, pp 574-585.

[12] Z. Li, and L. Xing, "Search the Chinese Web — Design and the Operation of Net-Compass," *Proceedings of the First Asia Digital Library Workshop*, 1998, pp. 42-46.

[13] F.-r. Lin, and S.-c. Lin, "A Conceptual Model for Virtual Organizational Learning," *Journal of Organizational Computing and Electronic Commerce*, 11(3), 2001, pp.155-178.

[14] D. Merkl, and A. Rauber, "Automatic Labeling of Self-organizing Maps for Information Retrieval," *Proceedings of ICONIP '99. 6th International Conference*, 1999, pp. 37-42.

[15] D. Morrison, "PATRICIA: Practical Algorithm to Retrieve Information Coded in Alphanumeric," *Journal of ACM*, 1968, pp. 514-534.

[16] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science* (5:1), 1994, pp. 14-37.

[17] R. Paolucci, "The Effects of Cognitive Style and Knowledge Structure on Performance Using a Hypermedia Learning System," *Journal of Educational Multimedia and Hypermedia* (7), 1998, pp. 123-150.

[18] G. Punj, and D. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research,* May 1983, pp 134-148.

[19] D. Roussinov, and H. Chen, "Document Clustering For Electronic Meetings: An Experimental Comparison Of Two Techniques," *Decision Support Systems,* 27 (1-2), 1999, pp. 67-79.

[20] G. Salton, and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management* (24:5), 1988, pp. 513–23.

[21] G. Salton, *Automatic Text Processing. Reading*, *Addison-Wesley*, MA., 1989.

[22] H. Yang, and C. Lee, "A Text Data Mining Approach Using a Chinese Corpus Based on Self-Organizing Map," *The Fourth International Workshop on Information Retrieval with Asian Languages*, 1999.

[23] C. Yang, J. Yen, S. Yung, and A. Chung, "Chinese Indexing using Mutual Information," *Proceedings of the First Asia Digital Library Workshop*, 1998, pp. 57-64.

[24] C.-p. Wei, and Y.-x. Dong, "A Mining-based Category Evolution Approach to Managing Online Document Categories," *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[25] K.-f. Wong, and W.-j. Li, "Intelligent Chinese Information Retrieval - Why Is It So Difficult?" *Proceedings of the First Asia Digital Library Workshop*, 1998.

[26] Z. Wu, and G. Tseng, "Chinese Text Segmentation for Text Retrieval: Achievements and Problems," *Journal of the American Society for Information Sciences* (44), 1993, pp. 532-542.