A Semantic FAQ System for Online Community Learning

Che-Yu Yang
Department of Information Management
China University of Technology, Taipei, Taiwan
Email: cyyang@cute.edu.tw

Abstract - **Knowledge is a product out of interaction. In online community learning, learners are encouraged to ask questions and discuss answers with each others. The value of participation and sharing during learning is emphasized. However, interactions between learners and between learner and instructor address a problem the instructor cannot be online all the time and it is not possible for the instructor to deal with lots of questions proposed from learners in a timely manner. Therefore, there is a need of an automated FAQ system to support learning efficiency of community learning that everyone in the learning system can share knowledge with and learn from each others.**

This article examines how to adopt online FAQ mechanism to facilitate knowledge share and demonstrates how a FAQ system can help knowledge generation, accumulation and share in an online community learning environment.

Index Terms – **automated frequently asked question, online community learning, knowledge share, semantic retrieval.**

I. INTRODUCTION

Knowledge is a product out of interaction. Current online learning systems haven not reached a high interaction degree between learners. This depresses knowledge generation and share among learners. Community learning especially emphasizes the communication and information share among group members. To leverage the interaction degree during learning, this paper introduces an online frequently asked question (FAQ) system acts like knowledge share platform. Through the processes of question raising and answer discussion, the FAQ knowledge base will be enriched for future question answering. Further, not only the learners can get answers to their questions, but also the instructors could know what problems learners encounter in learning. This would be a great help for both the instructor and the learners in online community learning environment.

The concept of community learning has become a hot topic as Web 2.0 came up. It stresses the importance of cooperation and share, which means learners working together to accomplish common learning goals and to maximize their own and community members' achievements. In [1], authors indicate that learners learn better when they learn together and foster creative thinking as members in a community generated new ideas, strategies, and solutions more frequently than working individually. Indeed, the effectiveness of community learning on the World Wide Web has been identified by various studies. It has been found that a learner's level of involvement and his incentive to learn have increased significantly with a wider and more complete understanding of the subject knowledge [2,3,4].

For an online FAQ system for community learning, the information retrieval model is the critical to success. Textual retrieval systems that utilize automatic indexing techniques to create text representatives from natural language, for better performance, must deal with the problems of polysemy and synonymy. Polysemy, a single word form having more than one meaning, decreases retrieval precision by false matches. While synonymy, multiple words having the same meaning, decreases the recall by missing true conceptual matches. To overcome these problems, textual information can be dealt with by its underlying concepts (using word sense disambiguation technique), rather than the keywords (word forms) [5,6].

The content of this article is arranged as follows. Section 2 presents a survey of the related works on community learning. In section 3, how an FAQ system will facilitate online community learning is demonstrated. In section 4 the architecture of interactive community learning environment based on automated FAQ mechanism is presented. Section 5 demonstrates the interactions between learners and between learn and the system. Finally, a summary of this paper and several conclusions are enumerated.

II. RELATED WORKS

Community learning promotes a type of collaborative learning mode. Several researchers agree that students perform better through group learning than by learning alone [1,7]. In group learning mode students are interested in sharing their knowledge form a learning group. Students may learn through the assistance of other group members. Group members communicate experience and viewpoint, discuss questions, help each others, and teach each others, etc. Therefore, learning is both a group activity and a social process and thus learning performance is strongly affected by peers [8]. In the development of networks, the learning eliminates the obstacles of time and space. Students can participate in community learning by computer at anyplace, at the same time or different time - synchronous and asynchronous, respectively. Researchers have used activity theory to analyze Computer Supported Collaborative Work (CSCW) systems [9]. Group communication relationship [10] refers to the intra-group relationships determined by

the interactions among members. However, how to form a group is a problem in collaborative learning. [11,12] introduce two methods to form a learning group. But, in these methods, new learners cannot participate in the learning group after the group has been formed and teachers must all be online.

This paper brings forward a new FAQ system to form a group model, in which learners can attach their questions to the group when they want to collaborate with others. The method also considers that learners can communication through Q&A interaction to provide support for construction and access of information share.

III. Knowledge Cycle in Online Community Learning

An automated FAQ system in online community learning operates upon the FAQ knowledge base. When a learner has information need, he raises a question through a designed interface to the FAQ system. This question is then processed by the system and its representation (query) will be created. This representation is then compared with the representations (index) of the Q&A sets in the knowledge base. A rank is given to any match between the new question and any existing Q&A sets. The matched answers are then presented to the questioner.

If no answer is provided by the system to a new question, or the learner is not satisfied with the provided answers, it means no proper match with respect to the question is found from the knowledge base. In the circumstance, the question is then presented to the public (learning community) demanding an answer for it. Then other community members will act as instructors to share and reply their own answers or thought to the question. After the community members manually discuss and answer the question, not only the answer will be sent to the learner who asked for the answer, but also a new Q&A set is formed and entered into the FAQ knowledge base.

Besides raising questions when meeting difficulties, even when having no difficulties, a learner can still browse the knowledge base to see what problems other learners have encountered during learning - now and in the past, and see the answers or solutions offered by the community members.

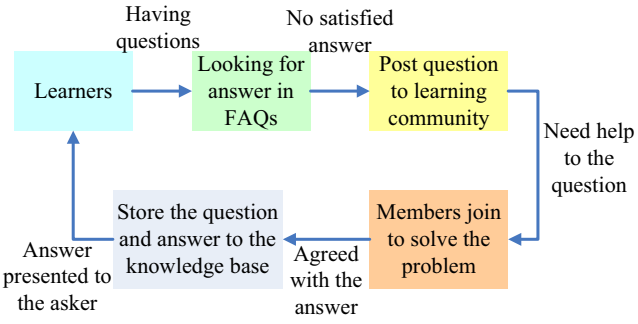

Figure 1. Knowledge generation, accumulation and share cycle formed by the interactions between community members.

The whole process of question discussing and answering in the online community learning environment is as shown in fig. 1. Through the process a learner raising questions and other members discussing for the answers, the knowledge in the database continues accumulating. In the future, this accumulated knowledge will be shared with other learners through raising questions or browsing the knowledge base. The interaction between the learners is actually a cycle of knowledge generation, knowledge accumulation and knowledge share. That is what we need for community learning.

This also conforms to Web 2.0 which includes a social element where users generate and distribute content, often with freedom to share and reuse. This property can result in a rise in the interaction degree of the learning activities on the web. Also this will maximize collective intelligence for each participant by formalized and dynamic information creation and share. The idea can be demonstrated with fig. 2. In this FAQ online community learning environment, each learning member plays two roles during learning – both knowledge requester and provider, depending on the question discussion he participates in. The knowledge flow is not limited to the one-way direction in which the instructor passes knowledge to the learners, which is common in the traditional online learning environment.

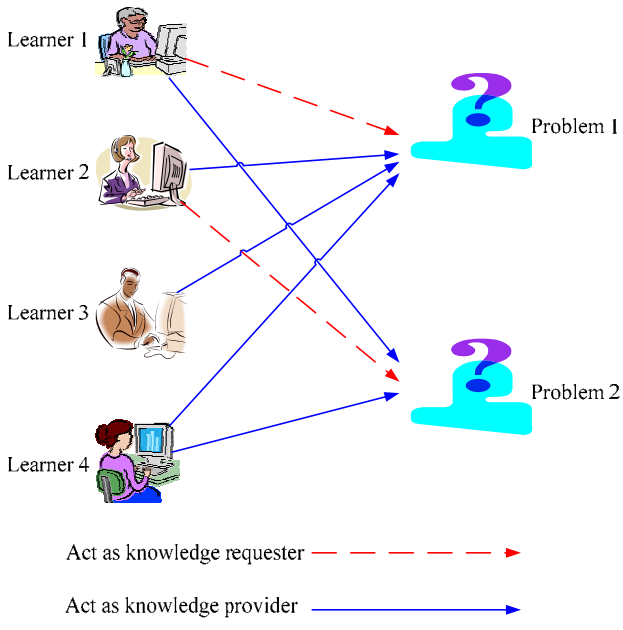

Figure 2. Each learner plays both roles of knowledge requester and provider.

IV. System Architecture

The architecture of the online automated FAQ system is shown in fig. 3. The system consists of five main components — including FAQ knowledge base, Text processor, FAQ module, Knowledge base maintenance module and user interfaces.

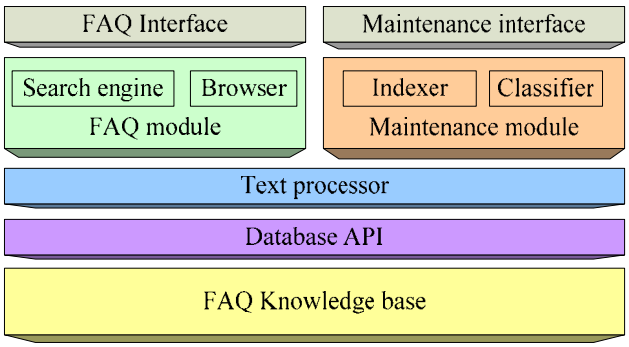

Figure 3. Architecture of the FAQ system.

The functions of the components are described as follows.

A. FAQ Knowledge Base

This is the repository of all Q&A sets and their meta-information. A Q&A set is a pair of question with its corresponding answers. It can be collected on a question-by-question basis through the process of learners working together to figure out a new coming question, or from a batch file consists of existing Q&A sets arranged by the instructor. Through the process of learning community members working together to solve problems for each others, knowledge will be generated and continue accumulating — the knowledge base will be enriched gradually. Not only the system will have more ability to answer various questions from learners, but also the instructor and the learners can know what problems the others encounter during learning.

B. User Interface

The main function of this component is to be the interface between users and the FAQ system. It provides learners with interfaces for raising his questions, searching or browsing knowledge base for existing answers, discuss on the unanswered questions. Also, there is an interface for system maintainer to maintain the knowledge base.

C. Text Processor

This component provides fundamental functions of natural language processing including token extraction, word statistics, syntactic analysis, sentence extraction…etc. These functions are provided in the form of APIs, thus can be easily used by upper layer of the architecture. Specifically, each textual data can be preprocessed in the following steps:

(1) Separate punctuation marks from words;
(2) Remove numbers and punctuation marks;
(3) Convert all words to lower case;
(4) Remove words like prepositions, conjunctions, auxiliary verbs, etc. A version of stop-list can be found in [13].
(5) The fifth step is called "stemming". The idea of stemming is to improve IR performance generally by bringing words sharing a common meaning into one heading variant form. Table 1 is a sample of vocabulary, with the stemmed forms generated by the stemmer.
(6) The remaining words after preprocessing are potential candidates for use as features in the information retrieval model, with each word as one feature in the feature vectors.

TABLE I
SAMPLES OF VOCABULARY WITH THE STEMMED FORMS.

Word	Stem form
consist, consisted, consistency, consistent, consistently, consisting, consists	consist
consolidate, consolidated, consolidating	consolidate
abandon, abandoned, abandoning, abandonment, abandons	abandon
accomplish, accomplished, accomplishes, accomplishing, accomplishment, accomplishments	accomplish

D. Knowledge Base Maintenance Module

This module mainly provides knowledge base maintainer with the function of indexing the textual data in the Q&A sets and the function of classifying the Q&A sets according to some predefined classifications. The index module uses the output (text with their part-of-speech from an asked question or a Q&A set) from the text processor to identify word senses of significant keywords. To extract significant keywords from text, log-entropy weighting scheme assigns minimum weight to terms that are equally distributed over documents, and maximum weight to the terms that are concentrated in a few documents. Entropy takes into account the distribution of terms over documents. The lower half weighted keywords can be eliminated and left the upper half as the significant keywords (which can be adjusted adapting to condition). Equation (1) is the log-entropy formula.

$$Entropy(T) = 1 + \sum_{j=1}^{n} P_j \bullet \log_2^{Pj} \qquad (1)$$

where T is a term in the collection, P_j is "the frequency of term T in a Q&A set j" divided by "the total number of times term T occurs in the whole collection."

To these significant words, the state-of-art TF-IDF weighting schema [14] is used to construct the features of each Q&A set. The following is the definitions in the schema:

- w_i : "ith Term" - a word, stemmed word, or indexed phrase
- D_j: "jth Document" - a unit of indexed text
- C: "The Collection" - the full set of indexed documents
- $TF(w_i, D_j)$: "Term Frequency" - the number of times w_i occurs in document D_j.
- $DF(w_i, C)$: "Document Frequency" - the number of documents from C in which w_i occurs. DF may be normalized by dividing it by the total number of documents in C.
- $IDF(w_i, C)$: "Inverse Document Frequency" - $[DF(w_i, C)\ /\text{size}(C)]^{-1}$. Most often the $\log_2(IDF)$ is used, rather than IDF directly.

In TF-IDF term weighting, in general,

$$TF\text{-}IDF(w_i, D_j, C) = F_1(TF(w_i, D_j) * F_2(IDF(w_i, C)) \quad (2),$$

Usually, $F_1 = 0.5 + \log_2(TF)$ or TF/TF_{max} or $0.5 + 0.5\ TF/TF_{max}$

$$F_2 = \log_2(IDF)$$

The term frequency in the given document shows how important the term is to that document. The document frequency of the term (the percentage of the documents which contain this term in the collection) shows how generally important the term is. A high weight in a TF-IDF ranking scheme is therefore reached by a high term frequency in the given document and a low document frequency of the term in the rest documents in the whole collection.

Textual retrieval systems that utilize automatic indexing techniques to create text representatives from natural language, for better performance, must deal with the problems of polysemy and synonymy. Polysemy, a single word form having more than one meaning, decreases retrieval precision by false matches. While synonymy, multiple words having the same meaning, decreases the recall by missing true conceptual matches. To overcome these problems, we can further index textual information by its underlying concepts [15].

To identify the senses of these significant keywords, we can perform the word sense disambiguation (WSD) process to the keywords (terms) above. The idea of WSD is shown by fig. 4.

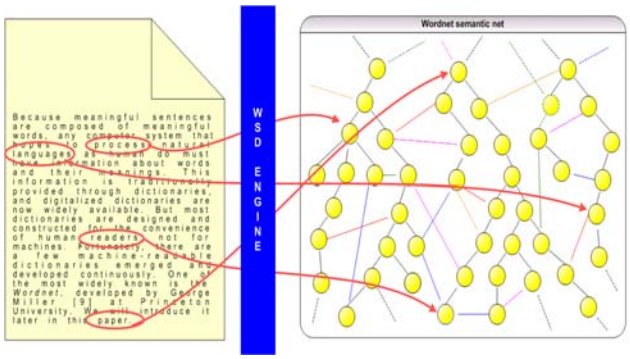

Figure 4. WSD task using Wordnet – to generate mappings between word forms and word senses.

In Wordnet [16] each synset represents a single distinct *sense* or *concept*, and is assigned a unique identification, called *synset-id*. By disambiguating a word in a text, we can get the corresponding concepts/senses defined in Wordnet for that word. Then we can use these senses (actually their synset-ids) as the index for that text. The text here can be a learner's question or a Q&A set. An algorithm for WSD is described in [16].

E. FAQ Module

This module provides the access to the knowledge base to the learners. Besides raising questions when encounter difficulties, even when having no difficulties, a learner can still browse the knowledge base to see what problems other learners have encountered during learning - now and in the past, and see the answers or solutions generated by community members. The browsing activity can be according to author name, question category, date, top viewed…etc. Besides, learners can also directly looking for the question and answer by using search engine module. Learner should be able to search in the fields of a Q&A set such as title, full text, subjects… etc. This component computes the match rank between the semantic index of a new incoming question and the semantic index of the Q&A sets in the FAQ database. Here the cosine measurement is used to retrieve semantically related Q&A sets to learner's question. That is, when a question and a Q&A set have common synset-id with similar weight, then this Q&A set would be judged as semantically related to the user question and then be retrieved.

After retrieve all the related Q&A sets, this answer list should be ordered by some ranking approach, normally according to the relatedness degree between the query and the match items. If a query Q is represented by $Q = (w_{q1}, w_{q2}, \ldots, w_{qt})$, and a Q&A set D_i is represented by $D_i = (wd_{i1}, wd_{i2}, \ldots, wd_{it})$ we have ranking formula (give each of these Q&A set a score) as (3):

$$Similarity(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * wd_{ij}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{dij})^2}} \quad (3)$$

where Q is learner's query, D_j is a matched Q&A set, w_{qk} is the weight of term k to query Q, w_{jk} is the weight of term k to a Q&A set D_j.

The workflow of the proposed semantic FAQ system for online community learning is shown as fig. 5.

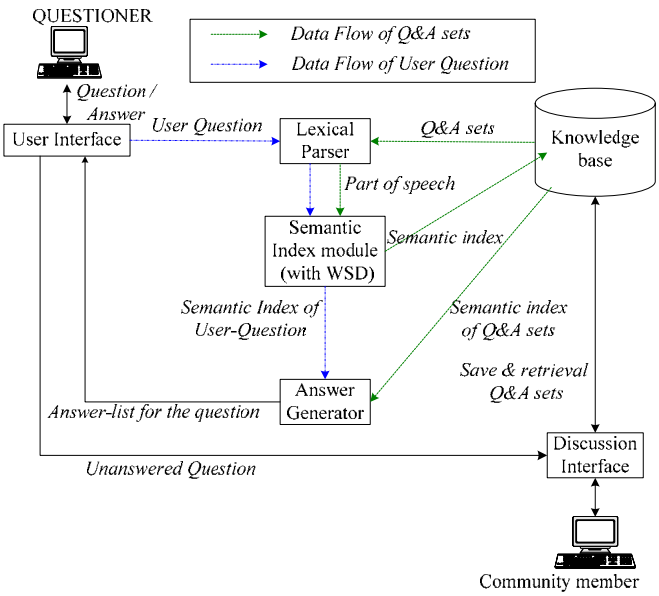

Figure 5. Workflow of the semantic FAQ system.

V. LEARNERS' INTERACTIONS

For the interactions between learners and between

learner and system, several interfaces have been designed. Through the interface in fig. 6, learner can type in an English question in natural language and get answers immediately without waiting for the instructor to get online. A list of answers provided by the system is showed to the questioner, as shown in fig. 7.

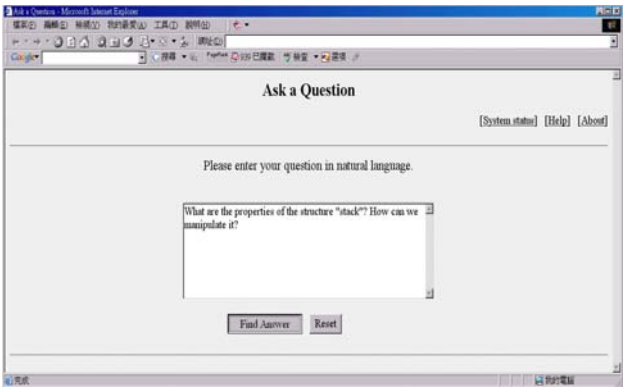

Figure 6. Use FAQ search engine to ask a question.

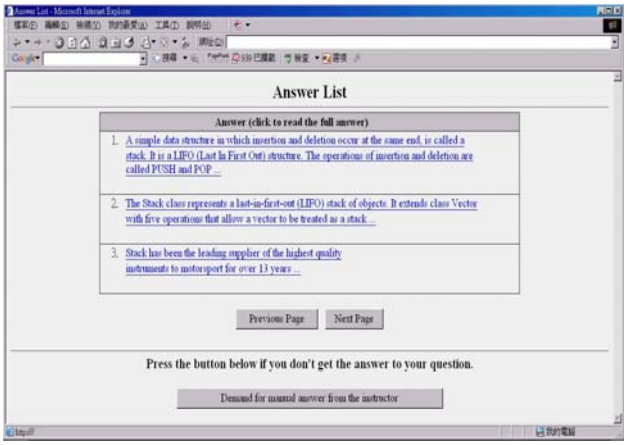

Figure 7. An answer list is provided to the questioner.

If no satisfactory answer is returned, then learner can press a button to send the question to the learning community, and this information need will then become public to the community members demanding for a discussion for the answer. The community members can see a list of unanswered questions and choose to discuss on desired ones, as shown in fig. 8.

When a learner clicks on a question, he can then post his opinion to the question, as shown in fig. 9. Not only the questioner who asked the question will be notified with the new post, but also a new Q&A set will be automatically formed and stored into the question answering knowledge base, if the questioner has figure out the question through the threads of discussion. Day by day, this process is enriching the knowledge base.

Besides raising questions when encountering difficulty, even when having no difficulties, a learner is also free to browse the knowledge base to see what problems others have encountered during learning - now and in the past, and see the discussion. Three kinds of browsing activities are designed – "top viewed Q&A", "recently added Q&A" and "browsing Q&A by category".

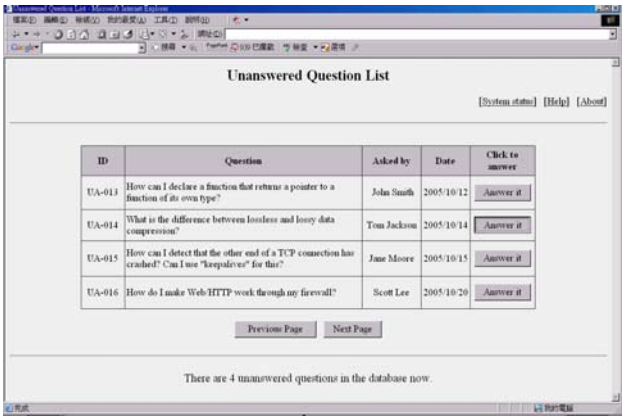

Figure 8. A list of unanswered question is public to the community members.

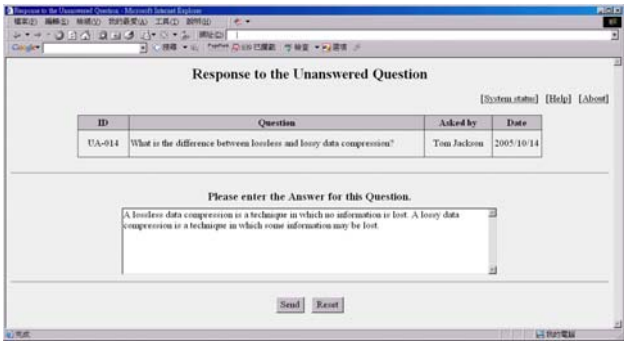

Figure 9. Discussion thread to a question.

VI. Conclusion

In this research, an online automated FAQ system in the community learning environment is designed. The system operates upon the FAQ knowledge base. In the knowledge base, pairs of question with its corresponding answer (Q&A sets) are collected through the process of learners raising questions and the community members discussing on them.

It is very important to have such a system in community learning environment. It benefits both the instructor and the learner – the instructor can release from the heavy load of answering learner's questions, and the learners can looking for the answers to their questions with out the constraint of time and space. Also, through the process of interaction among community members, the knowledge is created and shared in the community. In some way, the proposed system acts like an advanced knowledge share platform, which not only automatically solves the questions based on its knowledge base, but through this process also enriches its own knowledge.

References

[1]. D.W. Johnson and R.T. Johnson, *Cooperation and competition: Theory and research. Edina*, MN: Interaction Book Company, 1989.

[2]. Y.H. Lee and N.S. Chen, "Group composition methods for cooperative learning in web-based instructional systems," in proceedings of 8th International Conference on

Computers in Education/International Conference on Computer-Assisted Instruction, pp.1535-1538, Nov. 2000.

[3]. M. Nagai, Y. Okabe, J. Nagata, and K. Akahori, "A study on the effectiveness of web-based collaborative learning system on school mathematics: Through a practice of three junior high schools," in proceedings of the 8th International Conference on Computers in International Conference on Computer-Assisted Instruction, pp. 279-283, Nov. 2000.

[4]. J. Su, W. Chen, F. Chen and Y. Tsai, "The project-based collaborative learning on Internet—A case study on Geology Education," in proceedings of the 8th International Conference on Computers in International Conference on Computer-Assisted Instruction, pp. 303-308, Nov. 2000.

[5]. C.Y. Yang, M.S. Chiu, C.H. Yang and T.K. Shih, "A Semantic-based Automated Question Answering System for e-Learning", in Proceedings of the Tenth International Conference on Distributed Multimedia Systems (DMS'2004), San Francisco, September 8 - 10, 2004.

[6]. J.C. Hung, C.S. Wang, C.Y. Yang, M.S. Chiu and G. Yee, "Applying Word Sense Disambiguation to Question Answering System for e-Learning", in proceedings of the IEEE 19th International Conference on Advanced Information Networking and Applications, Tamkang University, Taipei, Taiwan, March 28 - March 30, 2005.

[7]. R. Slavin, *Research on cooperative learning and achievement: what we know, what we need to know.* Contemporary Educational Pshchology, 21, pp. 43-69, 1996.

[8]. J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, 1991.

[9]. K. Kuutti, "The concept of activity as a basic unit of analysis for CSCW research", in proceedings of the Second European Conference on Computer-Supported Co-operative Work: EC-CSCW'91 (eds. L.J. Bannon, M. Robinson & K. Schmidt) pp. 249-264, Kluwer, Dordrecht, 1991.

[10]. P. Watzlawick, *Pragmatics of Human Communications: A Study of Interactional Patterns.* Pathologies and Paradoxes. W.W. Norton, New York, 1967.

[11]. T. Supnithi, A. Inaba, M. Ikeda, J. Toyoda, and R. Mizoguchi, "Learning Goal Ontology Supported by Learning Theories for Opportunistic Group Formation", in Proceedings of the 9th International Conference on Artificial Intelligence in Education (AI-ED 99), July, 1999.

[12]. M. Wessner and H.R. Pfister, "Group Formation in Computer-Supported Collaborative Learning", in proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, September 2001.

[13]. David Lewis, *Representation and Learning in Information Retrieval*, University of Massachusetts ,USA. 1992.

[14]. G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, nr. 11, pages 613–620. November 1975.

[15]. C.Y. Yang, J.C. Hung and T.K. Shih, "Word Sense Determination using WordNet and Sense Co-occurrence", in proceedings of the IEEE 20th International Conference on Advanced Information Networking and Applications (AINA2006), Vienna University of Technology, Vienna, Austria, April 18 - 20, 2006.

[16]. C. Fellbaum, *WordNet - An Electronic Lexical Database,* ISBN: 0-262-06197-X, MIT Press, May 1998.

Che-Yu Yang was born in Hsin-chu city, Taiwan, in 1976. Yang received a Ph. D. from the department of computer science and information engineering, Tamkang University in Taiwan in 2006.

He is current an assistant professor in China University of Technology, Taiwan. His research interests include distance learning, question answering, natural language processing, information retrieval and semantic web.