Toward a Semantic Forum for Active Collaborative Learning

Yanyan Li¹, Mingkai Dong² and Ronghui Huang¹

¹School of Educational Technology, Beijing Normal University, China // liyy1114@gmail.com // Huangrh@bnu.edu.cn

²Corporate Technology, Siemens Limited China, Beijing, China // mingkai.dong@siemens.com

ABSTRACT

Online discussion forums provide open workspace allowing learners to share information, exchange ideas, address problems and discuss on specific themes. But the substantial impediment to its promotion as effective e-learning facility lies in the continuously increasing messages but with discrete and incoherent structure as well as the loosely-tied learners with response-freeness. To motivate and facilitate active collaborative learning, this paper describes the design of a semantic forum with semantic link networking on discussion transcripts. Based on domain ontology and text mining technologies, messages are automatically processed for structural modeling with semantic association and special interest groups are automatically discovered for topic-centric social context measurement, which lays the foundation for the fulfillment of distinctive functionalities in the semantic forum (i.e. semantic search, relational navigation and recommendation). Compared with traditional forums, the semantic forum has three outstanding features. First, it deals with the structural incoherence and content isolation within forums. Second, it enables active learning by providing learners with relational navigation to meet their learning demands. Third, it supports social context based ranking to recommend learning companions or transcripts for collaborative problem-solving.

Keywords

Semantic forum, Social context, Semantic link network, Relational navigation

Introduction

Proponents of collaborative learning claim that learners in cooperative teams achieve higher levels of performance and retain information longer than learners who work individually (Webb, 1995). Asynchronous discussion forum plays an important role in supporting collaborative learning, allowing learners to communicate at large to ask questions, articulate their thoughts, explain and justify their opinions, share ideas and resources, with collaborative contribution to the knowledge building within online discussion forum. The accumulated discussion transcripts contain a great deal of deposited human knowledge. So, almost all e-learning platforms or environment provide discussion forums, expecting to support flexible collaborative learning and maximize resource sharing and reuse.

However, there exist some challenges and difficulties that hinder online discussion forums being as an effective e-learning facility. Firstly, the messages being delivered are arbitrarily structured and the content is dynamically generated through multiple learners’ inputs, interactions, and annotations, thus learners have to spend a lot of time to go through nonlinear messages for locating useful messages. Second, messages are incoherent and disordered without semantic association. So, learners cannot search for the specific topic by retrieving certain discussion transcripts in separation. Last but not least, learners usually feel isolated and are easily tended to become disengaged and inactive when they are outside of the social context of the classroom (Amy et al., 2002). The literature in education research suggests that learners who are actively engaged in learning will be more likely to achieve success (Bonwell, 1996)(Richards, 1995)(Rubin et al., 1998). But investigators have come to recognize that asking learners to interact and discuss on forums does not necessarily lead to high-quality discourse. So, how to motivate learners to be actively engaged in collaborative learning via scaffolding intelligent facilities in forums becomes an important issue.

Therefore, by incorporating domain ontology and text mining technologies, this paper proposes a semantic forum that focuses on the semantic organization of discussion transcripts with semantic associations. By analyzing the content of messages as well as interaction between learners, an approach is also presented to discover special interest groups (SIGs) on the same or similar themes for topic-centric social context measurement. Meanwhile, a semantic search with relational navigation support is also provided to guide the learners through well-structured and coherent messages relevant to their learning demands. With such facilities, online discussion forum not only supplies communication space, but also serves as full-fledged learning environment with evolved structured knowledge repository to better support highly active and peer-to-peer in-depth collaborative learning.
Related Works

Analysis of the relationship among participants in discussion forums draws a lot of attention from researchers (Hewitt et al., 2007). Social Network Analysis (SNA) is an established method to derive person-person relations in the form of sociograms from "traces" of communication in a networked community (Wassermann et al., 1994). It is widely used to study the way people participated and interacted with each other in discussion boards (Laat et al., 2005), which provides information about the activities of such a community and the way they learn collaboratively. Based on writing and responding to messages, researchers have concentrated particularly on the constructed social networks and properties such as the small-world property and power-law degree distributions. This method is simply based on the information flow between learners but ignores the content of messages, and is therefore unable to discover interest groups to support theme-centered in-depth communication and learning. Wang et al. (2006) propose an analysis framework composed of forum level, thread level, and member level for assisting investigations on forum communities, in which the member level analysis is to segment members into different categories and identify their roles. Kato and Akahori (2005) study the familiarity among participants on a BBS by conducting an experiment with focus on self-disclosure of participants.

As a matter of fact, content of messages in the discussion boards are important for studying the relationship between learners. Researchers in the field of Computer Supported Collaborative Learning (CSCL) focus upon using content analysis technique to analyze transcripts of asynchronous discussion (Laat et al., 2005)(Schrire, 2006). Wever (2005) summarizes various content analysis schemes to analyze transcripts of online asynchronous discussion. Researchers usually annotate the messages one by one either by manual-coding or depending on assistance of content analysis software, which cost much time and efforts. So, this method only applies to small amount of messages.

Alternatively, some researchers put forward automatic method to discover useful information from online discussion forums (Dringus et al., 2005). Murakami et al present a method to extract information from web discussion boards by summarizing threads (Murakami et al., 2001). They use quote and comment relationships to extract a thread summary, which indicate there are topic bindings between messages. But as the messages on the forum are discrete and incoherent, so how to discover the topic-related messages and structure them orderly has to be solved. Xi et al. (2004) propose a specialized ranking function for Usenet by using linear regression and support vector machine techniques. Their approach is based on metadata, such as prior knowledge about the message author or the depth of the message. They do not address short message problem. Kim et al. (2005) propose a topic segmentation approach to discover the topic divergence within one thread. This knowledge can be used in segmenting the content in coherent units and guiding the learners through segments relevant to their navigational goals. But actually, the average thread depth is actually low in many discussion forums, so we emphasize the semantic organization of large amount of messages and support for semantic navigation instead of analyzing one discussion thread.

The Framework for Designing a Semantic Forum

The designed framework of a semantic forum illustrated in Figure 1 includes the following core modules:

- **User Interface**, which allows the geographically distributed learners to access the forum via a friendly interface. After successfully logging in, learners can asynchronously discuss with other learners via interaction facilities.
- **Text Corpus**, which saves all the original discussion transcripts delivered by the participating learners.
- **Knowledge Map (KM)**, which expresses the concept map of the discourse (i.e. concepts and the relations between concepts) by referring to domain ontology. Herein knowledge map can be drawn by experts or advanced learners via the authoring tool.
- **Transcripts Processing and Organization**, which is in charge of analyzing the discussion transcripts based on text mining technologies. Via such module, the transcripts are structured with semantic association.
- **Knowledge Engineer**: who is responsible for assisting transcripts processing and maintaining the knowledge repository.
- **Knowledge Repository**, which saves the processed messages with annotations of topic and type in a structured and semantic-associated manner. It dynamically evolves and serves as an objectification of the community’s advancing knowledge.
- **Intelligent Facilities**, which offers various functions to support active collaborative learning.
Figure 1. The framework for designing a semantic forum

Processing Discussion Transcripts for Semantic Organization

Message Topic Recognition

Assuming that the messages in a discussion thread represent the same topic, we combine them into a summary document and then process it. Each document is represented as a weighted term vector \( d = (d_1, d_2, \ldots) \) with the standard TFIDF function.

\[
d_i = TF(w_i, d) \times \log\left(\frac{|E|}{DF(w_i)}\right)
\]

Where the term frequency \( TF(w_i, d) \) is the number of times word \( w_i \) occurs in document \( d \), \( |E| \) denotes the total number of documents in the training set and the \( DF(w_i) \) is the number of messages containing the word \( w_i \) at least one time.

We consider the text in title field and body field of messages separately but discriminatively. Usually, title is the outline of body contents, so words in title field are more descriptive and discriminative in contrast to the words in body field. So, words in title field are assigned larger weights to reinforce their stronger impact. For \( TF(w_i, d) \), one time appearance in title field equals to \( t \) times appearances in body field. The cosine method is adopted to compute the similarity between the document vector and the concept vector in the knowledge map, and thus the documents belong to the certain concept with the maximum similarity value.

Message Type Identification

Diverse content analysis schemes have been proposed to classify transcripts of online asynchronous discussion (Wever et al., 2005), which reflects a wide variety of approaches and differ in their level of detail and the type of analysis categories used. Particularly, researchers (Gunawardena et al., 1997)(Kanuka et al., 1998)(Lally, 2001) argue that learners' discussion comprises a series of phases in terms of collaborative knowledge building (e.g. information sharing and comparing, concept exploring and discovering, and negotiation of meaning and construction of knowledge). Nevertheless, according to large amount of experimental study, most of learners' discussion transcripts actually fall into the first phase of information sharing and comparing that includes observations, opinions, statements of agreements, examples, clarifications, and identifications of problems. Accordingly, we
classify the messages in the forums into six types, which is useful to recognize the continuation of the topic regarding information sharing and comparison. Herein the messages types include Question, Opinion, Suggestion, Recommendation, Request and Citing. Thus, learners can search for messages with the specified type to cater for their individual demand. For example, if a learner wants to obtain some learning resources on the topic “semantic web”, he can input the word “semantic web” and select recommendation type. Question-type messages are the most common ones, and we further classify it into four categories corresponding to the four levels of knowledge, aiming to provide question-initiating messages accumulated from the former learners to help address later learners’ same or similar problems:

- **Concept-level questions.** The concept-level questions are the simplest ones, which mainly concerns the explanation or definition of a concept as well as some attributes of a concept.

- **Axiom-level questions.** This type of question is relevant to some common sense and the relationship between concepts, and the answer is “yes” or “no” with validating documents. Moreover, in some cases, the questions began with “what”, “who”, “name” and “which” interrogatives are relevant to the axiom. Such questions are the most ambiguous when it comes to determining the question type. This type of questions can be formulated as an embedded expression Relation (concept, concept) after being pre-processed.

- **Rules-level questions.** Rules mainly indicate the logical relationship between axioms, and causality relationship is typical, so this type of questions includes the questions began with “why” interrogative and other semantics similar questions that ask the reason for some axioms.

- **Method-level questions.** This kind of question takes into account finding a solution to a given problem, which is mainly related to the how-term and other semantics similar questions such as “The approach of …”. Obviously, this type of questions is the most complicated and the solutions to them will probably involve the reasoning according to the other three levels of knowledge.

<table>
<thead>
<tr>
<th>Table 1. Message types and corresponding features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Messages</strong></td>
</tr>
<tr>
<td>Question</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Rule-level (A^R)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Method-level (Q^M)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Opinion</td>
</tr>
<tr>
<td>Suggestion</td>
</tr>
<tr>
<td>Recommendation</td>
</tr>
<tr>
<td>Request</td>
</tr>
<tr>
<td>Citing (for reference)</td>
</tr>
</tbody>
</table>

By analyzing a large amount of messages on the forums, we define the most usual patterns and keywords for each type. Table 1 lists some of the patterns and hint-keywords for each type of messages, and more details can refer to (Zhuge et al., 2002). We currently parse each initiating-message to identify the various types of messages in Chinese. Herein we treat the title distinctively, and we process it by following a three-step process: Chinese word segmentation, hint-keywords matching and pattern matching. The Chinese word segmentation is based on the bilateral maximum matching algorithm, and the pattern matching is based on shallow syntax parsing (Fung et al., 2007). At the same time, we adopt the HowNet (Dong et al. 2007), a large lexical database of Chinese, to determine the synonyms and hypernyms for keywords matching. If the title is too short or there is not any matching hint-keyword or pattern, we further analyze the content of the initiating-message to identify its type. Regarding the case of multiple matching hint-keywords or patterns for one message, some rules will be executed to select the final type.
For example, the type of question or request will be considered the most probable because of their typical distinctiveness.

**Semantic Association of Messages**

A semantic link network (SLN) is a model to intuitively represent the semantic relationships between document fragments or documents (Zhuge, 2003). In this paper, we use the SLN to organize the messages with semantic associations. The semantic associations represent the binary relationships between messages, which is denoted as $M_i \rightarrow \alpha \rightarrow M_j$, where $\alpha$ is a type of semantic relationship while $M_i$ and $M_j$ are messages. The relationship can be one of the following types.

- **Sequential**, denoted as $M_i \rightarrow seq \rightarrow M_j$, which defines that $M_i$ is the prerequisite to $M_j$. A single message may have multiple prerequisite messages, and can also be a prerequisite to multiple messages.
- **Part-of**, denoted as $M_i \rightarrow par \rightarrow M_j$, which defines that $M_j$ is semantically a part of $M_i$.
- **Similar-to**, denoted as $M_i \rightarrow sim \rightarrow M_j$, which defines that $M_j$ is similar to $M_i$. The similar-to link is intransitive.
- **Reference**, denoted as $M_i \rightarrow ref \rightarrow M_j$, which means that $M_j$ is related to $M_i$, e.g. $M_j$ can be an annotation of $M_i$.
- **Cause-effect**, denoted as $M_i \rightarrow ce \rightarrow M_j$, which means that $M_i$ is the cause of $M_j$, and the $M_j$ is the effect of $M_i$.
- **Contrast**, denoted as $M_i \rightarrow con \rightarrow M_j$, which means that $M_j$ is in contrast to $M_i$. Unlike the similar-to link, this contrast link emphasizes the differences between $M_i$ and $M_j$ in a context. The contrast link is intransitive.
- **Build-on**, denoted as $M_i \rightarrow bo \rightarrow M_j$, which means that $M_j$ serves as the supplementary or additional content to $M_i$.
- **Rise-above**, denoted as $M_i \rightarrow ra \rightarrow M_j$, which means that $M_j$ is the summary or conclusion to $M_i$. Actually, a single message is rise-above on multiple messages.
- **Empty**, denoted by $M_i \rightarrow \emptyset \rightarrow M_j$, which represents that $M_i$ and $M_j$ are semantically unrelated.
- **Null**, denoted by $M_i \rightarrow N \rightarrow M_j$, which says that no semantic relationship between $M_i$ and $M_j$ is known for certain.

Semantic links between messages can be manually defined by forum participants, and automatically be discovered and derived under heuristic rules. The reasoning rules can be used for chaining the semantic relationships and obtaining the reasoning result from the chaining. A simple case of the reasoning is that all the semantic relationships have the same type, which is called single-type reasoning. According to the transitive characteristic of the semantic relationships, we have the following reasoning rule: $M_i \rightarrow \alpha \rightarrow M_j$, $M_j \rightarrow \alpha \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow \alpha \rightarrow M_k$, where $\alpha \in \{seq, par, ce, ref, bo, ra\}$. The heuristic rules suitable for connecting different types of semantic links are listed in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule1</td>
<td>$M_i \rightarrow par \rightarrow M_j$, $M_i \rightarrow par \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow par \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule2</td>
<td>$M_i \rightarrow par \rightarrow M_j$, $M_i \rightarrow ref \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow ref \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule3</td>
<td>$M_i \rightarrow par \rightarrow M_j$, $M_j \rightarrow bo \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow bo \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule4</td>
<td>$M_i \rightarrow par \rightarrow M_j$, $M_j \rightarrow ce \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow ce \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule5</td>
<td>$M_i \rightarrow ce \rightarrow M_j$, $M_j \rightarrow par \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow ce \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule6</td>
<td>$M_i \rightarrow bo \rightarrow M_j$, $M_j \rightarrow par \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow bo \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule7</td>
<td>$M_i \rightarrow bo \rightarrow M_j$, $M_j \rightarrow ra \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow ra \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule8</td>
<td>$M_i \rightarrow seq \rightarrow M_j$, $M_j \rightarrow bo \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow seq \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule9</td>
<td>$M_j \rightarrow par \rightarrow M_j$, $M_j \rightarrow ra \rightarrow M_k$ $\Rightarrow$ $M_j \rightarrow ra \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule10</td>
<td>$M_i \rightarrow par \rightarrow M_j$, $M_i \rightarrow con \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow con \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule11</td>
<td>$M_i \rightarrow N \rightarrow M_j$, $M_j \rightarrow \alpha \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow N \rightarrow M_k$</td>
</tr>
<tr>
<td>Rule12</td>
<td>$M_i \rightarrow \emptyset \rightarrow M_j$, $M_j \rightarrow \alpha \rightarrow M_k$ $\Rightarrow$ $M_i \rightarrow \emptyset \rightarrow M_k$</td>
</tr>
</tbody>
</table>
After recognizing the topics and types of messages, the transcripts are structured based on the heuristic rules by referring to the knowledge map. Figure 2 intuitively shows a part of Message-SLN. As it illustrates, the ellipse represents a topic-SLN, the square represents a threaded message, and the single-line arrow with label represents various relationships between messages within a topic-SLN. In this sample, there are only three kinds of relationship: reference, build-on, similar-to. The double-line arrow with label represents relationships between different topic-SLNs.

![Figure 2. An example of Message-SLN](image)

**Topic-centric Social Context Measurement by Discovering SIGs**

Social context is the identical or similar social positions and social roles as a whole that influence the individuals of a group. One important aspect of an individual’s social context is the people with whom the person interacts. A given social context is likely to create a feeling of solidarity amongst its members, who are more likely to keep together, trust and help one another (http://en.wikipedia.org/wiki/Social_environment). Therefore, by computing peer similarity based on topic association and browsing activities, this paper aims to discover special interest groups for supporting participants with friendly social context. Special interest group brings together persons focusing on the same interests, and they communicate at large to address concerns and raise awareness about some problems. They are relatively tightly organized, providing opportunities for sharing expertise and first-hand knowledge on the latest development trends while fostering and prompting the progress in the specific field. So, the kernel to discover SIGs within discussion forums is to detect the learners’ relationship as well as their interests according to their delivered messages.

**Combining SNA with Semantic Relations for Constructing SIGs**

Social Network Analysis relies usually on homogenous networks with one type of actors, namely persons, which are also called one-mode-networks. In such networks, centrality, prestige of a person or the overall centralization and density of a network can be computed by well-defined formulas. This kind of network utterly depends on the reply-to message relation in the forum and neglects the potential semantic relationship between the topics. Yet computer-supported interaction and cooperation typically involves mediating artifacts, such as written messages within discussion forums. These “tangible” products can be seen as another type of entity to be included in the network. Exchange on the level of the original artifacts can also be provided by similarity-based search in a shared collection of objects (Hoppe et al., 2005). If we restrict our analysis, in this sense, to “communication through artifacts” (Dix et al., 2004) this will result in networks only with relations spanning between elements of the two different categories (i.e. person-artifact). This network structure corresponds to a bi-partite graph.
Figure 3 illustrates the framework to detect special interest groups. The normal way to analyze the discussion transcripts corpus is to use SNA to count the reply-to relationship between learners, which results in a one-mode network. By adding the topics to which the messages belong, the one-mode network can be easily transformed into bi-partite network. On the other hand, community can define the knowledge map to express the domain knowledge. By building the semantic mapping from the topics in bi-partite network to concepts in KM, a theme-centered network can be constructed to indicate the persons gathered around one concept. In this way, the theme-centered network denotes the potential interests of the persons, and by adding the reply-to relationship, a special interest group can be formed with respect to each concept in the KM. Persons might be interested in other persons related to an object of their interest, to contact them for discussion and exchange.

After recognizing message topics with the above-mentioned method, the bi-partite network is transformed into the thematic-centered network by mapping the threaded messages to the concepts in the knowledge map. In the thematic-centered network, learners are linked to the themes if their delivered messages belong to the theme. To select the learners with centralized interest in one theme to form a special interest group, formula 2 is used to compute the energy of the $i^{th}$ learner for the $k^{th}$ theme. Only when the learner’s energy is larger than the threshold, can he or she belong to a special interest group.

Energy:

$$E_i = \frac{w_1 I_k^i + w_2 R_k^i}{N_i}$$

(2)

Where $N_i$ denotes the total number of messages posted by the $i^{th}$ person, $I_k^i$ denotes the number of $k^{th}$-theme-related messages initiated by the $i^{th}$ person, and $R_k^i$ denotes the number of $k^{th}$-theme-related replying messages posted by the $i^{th}$ person. $w_1$ and $w_2$ are parameters that adjust the relative impact of messages. The threshold is set as $\sum \frac{E_i}{M_k}$, where $M_k$ denotes the total number of learners linking to the $k^{th}$ theme in the thematic-centered network.

After discovering special interest groups within discussion forums, the following step is to compute several criteria for a SIG member, including participation, mutuality and activity.

Participation:

$$P_i = \beta \cdot \sqrt{\frac{L_i}{\sum_{j \neq i} L_j}} + (1 - \beta) \cdot \sqrt{\frac{R_i}{\sum_{j \neq i} R_j}}, \beta \in [0,1]$$

(3)
Where $I_i$ denotes the number of messages initiated by the $i$th person, $R_i$ denotes the number of messages replying to others posted by the $i$th person, $l$ denotes the total number of members in a special interest group.

**Mutuality:**

$$M_{ij} = \frac{S_{i-all} \times S_{j-all}}{S_{i-j} \times S_{j-i}}$$

Where $S_{i-all}, S_{j-all}$ are the number of messages that $i$ and $j$ reply to others’ messages and sent to others by means of private short-message, respectively. $S_{i-j} (S_{j-i})$ is the number of messages that $i$ ($j$) replies to $j$ ($i$) and sent to $j$ ($i$) by means of private short-message. So in the network for a SIG, the shorter the length between the nodes is, the more close the relationship between the persons is, and vice versa.

**Activity:**

$$AC_i = \frac{N_i}{\Delta t_p} \times \frac{1}{(t - t_d) + \tau}$$

Where $N_i$ represents the number of messages posted by the $i$th person during the period $\Delta t_p$, $t$ is the current date and $t_d$ is the date the $i$th person posted the latest message in the forum. $\tau$ is the adjust parameter to avoid the denominator is zero, and it is initially assigned 1.

Once a person belongs to a special interest group, he or she will be informed of other learning companions to enhance their in-depth communication and learning. Also, the discussion board can proactively push the newly emerged messages on the specific theme or other related information to the member by means of email or other instant messenger. Such pushing service guarantees the members can keep track of the timely progress on the theme even if they do not login to the forum.

**Collaborative Knowledge Building for Evolving KM**

It is well known that the forum provide convenience for learners to discuss on some common interesting topics, and this constitute a cycle of personal and social knowledge-building. During the collaborative learning process, some new concepts will emerge while some concepts will vanish. So, we believe, instead of being constructed once for all, the knowledge map should evolve in accordance with the progress of the domain. That is, the knowledge map is dynamically constructed with the contributions from all the participated learners, which reflects the up-to-date knowledge in the specific domain.

Topic detection and tracking (TDT) research (Allan et al., 2003)(Brants et al., 2003)(Kleinberg, 2002)(Zhang et al., 2002) mainly focuses on detecting and tracking events in streaming news data. TDT systems monitor continuously updated news stories and try to detect the first occurrence of a new story; i.e., an event significantly different from those news events seen before. Based on the approaches, the KM is initially constructed by the experts or advanced learners, and then periodically analyzing the messages to recognize the possible new topics. We combine the messages in a discussion thread into a document, and then compare a new document with the past documents and make a decision regarding the novelty of the message based on the content-based similarity values. If the new document is not similar to any past documents, it will be labeled particularly. When such documents amounts to a threshold, a new concept with additional describing terms that represents documents will be added to the KM whilst marked red to show its novelty. Likewise, if no documents are classified into concepts in KM for a period of time, the concepts in KM will be marked grey to denote its inactivity.

**Supporting Intelligent Facilities**

**Semantic Search with Relational Navigation Support**

In contrast to traditional search engines, the query-keywords input can be denoted with certain semantic information, for example a concrete concept in the knowledge map. With this semantic denotation, the query results can be
profundely-repacked with the inward semantic schema of the concept, including basic information and semantic relationships with other concepts. With the advantage of semantic inference with logic foundation, more implicit information can be extracted, and the search results can be more comprehensive and rational. Additionally, learners can specify the type of information he or she want to find, such as questions, recommendation, etc, which leads to the focused search and quick location of the useful information. If the learner does not limit the type of messages, the resulting messages are listed in the default type sequence. Regarding each type, we take into account several criteria to rank the messages, including the number of clicking, the number of replying, and the authority of the message authors.

![Figure 4. A schematic navigation view](image)

Instead of listing search algorithm (see (Li et al., 2006) for more details), we address the search process more clearly by a search example. When a learner proposes a query “ontology mapping” without type constraint, the system returns all the relevant messages. Firstly, process the query to search for the matching concept C in the KM as the anchor node. Secondly, semantic associated information must be extracted. There are two main semantic-associated concepts B and D for the anchor node via relationship of similarity and subtype. Additionally, by using the reasoning rules (e.g. transitivity) on the semantic relationship (e.g. Subtype, Sequential) between concepts, the prerequisite concept A is also selected. In this way, all the chosen concepts with their belonging messages are displayed to the learners in a well-structured and coherent manner. A schematic navigation view is depicted in Figure 4.

**Recommendation**

We use the following formula to compute the authority of a learner within a special interest group, and thus each learner can be informed of other learning peers with different authority on a common theme.

\[
A_i = \alpha \cdot \frac{\sum_{t=1}^{I_c^t} \left( \frac{I_c^t}{\max(I_c^t)} + \frac{I_r^t}{\max(I_r^t)} \right)}{2I_i} + (1 - \alpha) \cdot \frac{I_g^t}{I^g}, \alpha \in [0,1]
\]  

(6)

Where \( I_c^t \) denotes the times of clicking by others for the \( i \)th message initiated by the \( i \)th person, \( I_r^t \) denotes the times of replying by others for the \( i \)th message initiated by the \( i \)th person. \( I_g^t \) denotes the number of messages delivered by the \( i \)th person that are set as good messages. \( I_i \) denotes the total number of messages delivered by the \( i \)th person. \( \max(I_c) \) and \( \max(I_r) \) denotes the maximum clicking-times and replying-times for a message, \( I^g \) denotes the total number of good messages in this SIG.

Additionally, messages are organized and displayed in a multi-mode to cater for the different learning needs of learners:
- **Best-messages view**, which allows the learners to browse the messages that are set as good by the host.
- **Popular-messages view**, which enables the learners to browse the messages that draw a lot of attention from others by ranking the messages according to the number of replying and clicking.


• **Messages-type view**, which empowers the learners to quickly locate the messages with specific types, such as question, opinion, request, or recommendation.

• **Peer-centric message view**, which enables the learners to know about all the messages delivered by a peer. This mode is helpful for novice to learn from other peers with high authority on the specific theme.

**Experimental Study**

In order to evaluate the effectiveness of the approach presented in this paper, we performed an experimental study and compared the analysis feedback provided by assessors of a discussion forum with the analysis results obtained by the proposed approach. We used the W3CHINA discussion board (available at http://bbs.w3china.org/index.asp) as the data source. We randomly selected 763 discussion threads in a view of “Semantic Web and ontology”, with a total of 4512 messages from the source. We asked two postgraduate students whose research direction is semantic web to assess each discussion thread. They manually labeled the messages with the belonging concept in KM and the type of messages. We then compared these manual labeling results with the labels assigned by the proposed automated approach.

![Image: The defined partial Knowledge Map](Figure 5)

Figure 5 shows the partial knowledge map defined by the organizer of the forum, and the messages are automatically classified to be instances of the concepts. We use the following two formulas to compute the micro-average-precision and micro-average-recall for the classification approach, where $a$ denotes the number of auto-labeling belonging messages indeed belonging to the class, $b$ denotes the number of auto-labeling belonging messages not belonging to the class, and $c$ denotes the number of auto-labeling not belonging messages indeed belonging to the class. Table 3 illustrates the evaluation results for the automated classification.

$$
\text{Precision} = \frac{\sum_{c \in C} a}{\sum_{c \in C} a + \sum_{c \in C} b} \times 100\% \quad \text{Recall} = \frac{\sum_{c \in C} a}{\sum_{c \in C} a + \sum_{c \in C} c} \times 100\%
$$

(7)

<table>
<thead>
<tr>
<th></th>
<th>User-labeling belonging</th>
<th>User-labeling not belonging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-labeling belonging</td>
<td>$\sum_{c \in C} a = 691$</td>
<td>$\sum_{c \in C} b = 65$</td>
</tr>
<tr>
<td>Auto-labeling not belonging</td>
<td>$\sum_{c \in C} c = 72$</td>
<td>-</td>
</tr>
<tr>
<td>Micro-average evaluation</td>
<td>$\bar{r} = 90.5%$</td>
<td>$\bar{p} = 91.4%$</td>
</tr>
</tbody>
</table>

Table 3. Evaluation for the automated classification

On the other hand, we use formula 8 to compute the success-rate of message-type auto-identification approach, where $P_{\text{manual}}$ denotes the number of messages annotated by the assessors, and $P_{\text{auto}}$ denotes the number of messages annotated automatically. Table 4 shows the messages-type identification comparison for the manual-labeling and auto-labeling, from which we can see that the overall success-rate for identifying message types is relatively

80
satisfied. But there are 14 messages that cannot be classified into any class. After analyzing such messages, we found that the main reason is that the title and content of the initiating messages don’t contain any typical hint-keywords or match any templates.

Table 4. Messages-type identification comparison

<table>
<thead>
<tr>
<th>Message type</th>
<th>Approach</th>
<th>Manual labeling</th>
<th>Auto labeling</th>
<th>Common</th>
<th>Success-rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td></td>
<td>363</td>
<td>354</td>
<td>340</td>
<td>90.1%</td>
</tr>
<tr>
<td>Opinion</td>
<td></td>
<td>185</td>
<td>192</td>
<td>174</td>
<td>84.8%</td>
</tr>
<tr>
<td>Suggestion</td>
<td></td>
<td>30</td>
<td>33</td>
<td>28</td>
<td>90%</td>
</tr>
<tr>
<td>Recommendation</td>
<td></td>
<td>81</td>
<td>84</td>
<td>79</td>
<td>91.8%</td>
</tr>
<tr>
<td>Request</td>
<td></td>
<td>80</td>
<td>89</td>
<td>78</td>
<td>85.7%</td>
</tr>
<tr>
<td>Citing</td>
<td></td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>91.6%</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td>14</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

\[
\text{success - rate} = \frac{P_{\text{manual}} \cap P_{\text{auto}}}{P_{\text{manual}} \cup P_{\text{auto}}} \times 100\% \tag{8}
\]

We have implemented an integrated toolkit VINCA with comprehensive functionalities to support content analysis with intelligence and visualization features (Li et al., 2007). It can be installed stand-alone or support online downloading of the discussion transcripts from forum (currently support KF (http://kf.cite.hku.hk) and W3China forum) to conduct analysis. Figure 6 shows the interface to set parameter and conduct text analysis, which provides foundation for executing the message theme-classification and type-identification, and accordingly learners’ semantic relationship matrix can be obtained (see figure 7). Afterwards, in order to intuitively illustrate the discovered SIGs, we imported the obtained relationship matrix to the SNA analysis software UCINET to draw the networks for SIGs on different themes.

Figure 6. Setting interface of parameter and conduct text analysis

Figure 8 and figure 10 respectively show the original constructed relationship network on “DL” and “ontology”, in which numerous learners are associated as long as they delivered messages on the themes. Comparatively, figure 9 and figure 11 show the discovered SIGs on “DL” and “ontology” with learners’ energy filtering, respectively. As figure 9 and figure 11 illustrate, the number of learners decrease greatly and the organization structure of the SIGs become more clear-cut and visible. Taking the SIG on “DL” as an example, we can see that the learner 3771 (ID number) plays the most important role in this SIG, who draw a lot of attention from others. Learner1, 4440, 1, 11958, 9173, 22294 play a secondary role, and learner 1 has more close relationship with learner 3771. Especially, there is an isolate learner 16413 who has no any association with others.
Figure 7. The interface to export learners’ relation matrix

Figure 8. Original relationship network on “DL”

Figure 9. Discovered SIG on “DL”
Figure 10. Original relationship network on “ontology”

Figure 11. Discovered SIG on “ontology”

Figure 12. Structured search results augmented with related themes
The prototype of modules has been implemented towards the semantic forum. Figure 12 gives a snapshot of the search module interface within the discussion forum. It illustrates the search results for the query "ontology mapping". As the figure shows, a learner can select one item from the drop-down list to specify the type of information that he or she intends to search, or search for all types of information by skipping this selection. As a result, the matching messages are classified and displayed in a structured manner. Each message is described with type, title, author, focus (times of clicking and replying), date and replier. To augment the search results, related themes are also recommended at the bottom of the interface supporting learners’ extensive and further navigation. Thus, a learner can easily click the related themes to get more focused information. Figure 13 shows the recommendation interface for users to browse the messages in a multi-mode, such as best messages, popular messages, etc. Furthermore, more messages posted by learning peers are also recommended for extended learning.

Discussion

The popularity of online discussion forums enables learners to discuss problems concerning with the conception, development, implementation, etc. in some fields flexibly and freely. But because of the discrete discussion topics, response-freeness and uncertain partners, this kind of loosely organized mode hinders the learners to conduct in-depth communication and cooperation. If the learners could constitute some special interest groups with relatively steady and tight relationship, they will have more motivation and enthusiasm to involve in the deeper communication and interchange of ideas and knowledge.

On the other hand, online discussion forums usually only support simple navigation by listing messages in sequence of date, author or focus. But as the messages increasingly grow over time, learners feel difficult to navigate through the large amount of messages to find the useful information they really need. Currently, the search functions provided on forums are quite weak. Simple keyword-matching-based search leads to the large returning results and some related information missing, so learners still have to spend a lot of time browsing the messages for useful messages and they are easily to get lost during the navigation. Therefore, in order to accommodate the increasingly added messages within the forum, it is indispensable to improve the search precision and coverage so as to alleviate the navigation burden for the learners. Especially, along with semantic-associated information, search results are displayed in a well-structured and coherent manner. In this way, semantic forums offer great convenience and flexibility for learners to support their informal learning and problem-solving.

Conclusion

This paper presents a semantic forum for both active knowledge-rich interactions and peer-to-peer collaboration. Its kernel idea is to organize the discussion transcripts in a semantic link network rather than in a discrete and incoherent
structure, and to discover special interest groups with indicators as authority, participation, mutuality, and activity, allowing learners to closely collaborate with peers as they acquire and build the evolved knowledge repository. In addition, the proposed semantic search empowers the learners to conveniently and effectively access the useful information with relational navigation support. The experimental results show that the approach is feasible and efficient, enabling the effective discovering of interest groups and proper demand-driven navigational guidance.

Ongoing work is to enhance the approach performance by considering the context in a thread, to apply the forum in the practical learning settings and improve it according to the feedback from learners.

**Acknowledgment**

This research work is supported by National Natural Science Foundation of China (NSFC: 60705023). All data source has been kindly provided by the W3CHINA organization in China. We would like to thank Mr. Xu Han for his cooperative work on data analysis.

**References**


