Computer-Supported Content Analysis for Collaborative Knowledge Building in CSCL

Jian Liao, Yanyan Li, Ying Zhou, Ronghuai Huang, Jingjing Wang

Knowledge Science & Engineering Institute, Beijing Normal University, 100875, Beijing, China
huangrh@bnu.edu.cn

Abstract: Interaction analysis plays an important role in computer-supported collaborative learning. This paper proposes an interaction analysis model CSCAC composed of three dimensions: group building level, member contribution, member support, which lays the foundation for interaction content analysis. Based on the model and domain ontology, this paper explores the text mining and semantic analysis technologies to automatically analyze the interaction content for collaborative knowledge building. An experiment is conducted to compare the auto-analysis result with that of manual analysis, which shows the the proposed approach is feasible and effective mostly to automatically analyze the interaction content in CSCL.

Keywords: CSCL, Collaborative Knowledge Building, Content analysis, Domain ontology, Concept Map, Keyword extraction, Hownet

Introduction

With the development of e-learning, Computer-Supported Collaborative Learning (CSCL) has drawn more and more attention of researchers. In the process of collaborative learning, the validity of evaluation to group collaboration is an important but difficult task.

Henri (1991) is the first person who put forward an analysis model of understanding interaction contents. [1] Newman(1995) et al. design a content analysis model based on Henri, including the following ten dimensions: Relevance, Importance, Novelty, Bringing outside knowledge/experience to bear on problem, Sticking to prejudice or assumptions, Ambiguities: clarified or confused, Linking ideas: Interpretation, Justification, Critical assessment, Practical utility (grounding), Width of understanding... And they use the formula $X^+ - X^-$ to calculate the degree of cognitive process by each group member or the whole group. [2] Besides, Liu (2005), Zhu(1996) , Pena-Shaff(2001) , N.M.Avouris and other scholars analyze and research on the interaction contents in collaborative learning from different dimensions.
1 The Model of Computer-Supported Content Analysis for Collaborative Knowledge Building (CSCAC)

On the topic of how CSCL analyzes contents, the important representative scholar in CSCL Koschmann (1996) considers research on the meanings and the building processes of meanings in common activities as one of the main research contents of CSCL. Many researchers use “Collaborative Knowledge Building” to explain the building processes of meanings in common activities. Therefore, the analysis of collaborative knowledge building in the collaborative groups is the important stand for setting up the whole content analysis model.

The first problem for judging collaborative knowledge building is the knowledge representation. Many researchers regard Concept Map as an effective means of representing and measuring students’ knowledge structures. So we can set up the domain ontology of collaborative learning contents by drawing concept maps related to the topic of group discussion.

Keywords in interactive texts can be used to represent the individual knowledge of the group members. The relatively important keywords, which are regarded as the representation of group member’s individual knowledge, can be extracted by searching the interactive speeches of members.

The group knowledge contain in each group members’, but not the simple sum of all individual knowledge. Some knowledge may be implied by all members, and some may only be in the possession of very few members. Therefore, the knowledge of collaborative groups needs to be analyzed in the holistic perspective.

When comparing member or group knowledge with domain ontology, as domain ontology is described by concept and member or group knowledge is described by keyword and the relations between words and concepts are many to many relations. Therefore, we need to carry out semantic analysis of the keywords in order to obtain correct mapping between concepts and words.

![Figure 1. CSCAC Model.](image)

Based on the above knowledge representation and directed by the activity theory, we set up the following model of analyzing collaborative knowledge building, as shown in figure 1. In figure 1, the three circles around the small triangle in the middle represent the three core components of activity theory respectively, which are subject, object and group knowledge. The three sides of the small triangle represent the inter comparison methods.
between different sorts of knowledge. The comparison between member knowledge and group knowledge or between member knowledge and other group members’ knowledge is carried out by analyzing the contents of their speeches, such as keyword matching method, etc. While the comparison between member knowledge and domain ontology or group knowledge and domain ontology is carried out by semantic mapping method. The three rectangles contained in the large triangle represent the three perspectives for analyzing collaborative knowledge building. The analysis of collaborative knowledge building is stated from three perspectives as follow.

2 Multi-dimensional Analysis

2.1 Level of group collaborative knowledge building (GBL)

At present, many researches analyze the level of collaborative knowledge building from the perspectives of “contention degree” and “building degree”. Fisher(1993)、Mercer(1995)、Coelho(1994)和 so on consider effective interaction as “exploratory talk”, whose characteristics is that the participants communicate in a critical but constructive way. [6]

Therefore, we need to examine both the critical and constructive characteristics of the speech contents. We assume that the more decentralized the speech contents of the group members are, the more critical it will be; and also the closer the speech contents of all group members is compared to the domain ontology constructed by teachers and experts, the more constructive it will be.

Definition 1: Level of Group Collaborative Knowledge Building \[ \text{GBL} = R \times U \]

Where R represents the Relevance of group speeches to the topic. U represents the inconsistence of the speeches among group members.

Definition 2: Relevance \[ R = \frac{M}{L} \]

Relevance indicates the relevance of group speeches to the discussion topic, where M represents the number of concepts in the speeches of the whole group which cover the domain ontology. L represents the number of all concepts in the domain ontology.

Definition 3: Inconsistence \[ U = \mu \sum_{i=1}^{S} \chi_i^2 = \mu \sum_{i=1}^{S} \sum_{j=1}^{N} \frac{(F_{ij} - F_j')^2}{F_j} \]

Inconsistence means the difference of the speeches among group members, where N represents the sum of keywords in the whole discussion speeches. S represents the number of the group members. \( \mu \) represents an adjustment factor, whose value is \( \frac{1}{S \times N} \). \( \chi_i^2 \) represents the decentralization degree of the \( i \)th member’s speeches compared with the speeches of the whole group. The more the value is, the more inconsistent it will be. \( F_{ij} \) represents the frequency of the \( j \)th keyword in the \( i \)th member’s speeches. \( F_j' \) represents
the average number of the $j$th keyword mentioned by each member in the group speeches.

2.2 Member contributions to collaborative knowledge building

We can estimate the contribution of group members to the task by comparing the knowledge of group members and the domain ontology structure presumed by teachers and experts and analyzing the similarity of the two sorts of knowledge.

Furthermore, Newman(1995) and Liu(2005) both consider the novelty of members’ speeches and the extension of the topic to be discussed as the important criteria of evaluating the quality of interactive texts.

**Definition 4:** Member’s Contribution  
$$MC_i = \alpha R_i + \beta V_i + \gamma E_i$$

Where $R_i$ represents the Relevance of group member $i$’s speeches to the topic, whose algorithm is shown as Definition 2. $V_i$ represents the novelty of group member $i$’s speeches. $E_i$ represents the extension of group member $i$’s discussion contents. $\alpha, \beta, \gamma$ are adjustment factors, which represent the impact degrees of Relevance, novelty and extension respectively on member’s contribution. And their values are usually as $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ respectively.

**Definition 5:** Novelty  
$$V_i = \frac{P_i}{N}$$

This indicates the novelty of member’s speeches, where $F_i$ represents the number of keywords mentioned for the first time by group member $i$. $N$ represents the total number of keywords. For example, in a discussion, the total number of keywords is 370, among which member A first mentions 118, so we reach the novelty of member A’s speeches as $118/370 = 0.32$.

**Definition 6:** Extension  
$$E_i = \frac{N_i}{N}$$

This indicates the extension of member’s speeches, where $N_i$ represents the number of keywords in each member’s speeches. $N$ represents the total number of keywords in the whole discussion speeches. For instance, the number of keywords in member A’s speeches is 207 and the total number of keywords mentioned by all group members is 370, so the extension value is the ratio of 0.56.

It can be proved that the values of $MC_i$, $R_i$, $V_i$, $E_i$ are all between [0, 1].

2.3 Group member’s support and assistance to the other group members

The evaluation of the group member’s support to or the assistance from the other group members is done by the comparison of knowledge structure between individual
group members. We believe that the more similar of two members’ speeches, the mutual support will be greater. The algorithm of similarity is the same as the vector cosine method which is frequently used in calculating the document similarity in text classification and clustering.

\[ S_{ij} = \frac{F_i \cdot F_j}{|F_i| \times |F_j|} = \frac{\sum_{k=1}^{N} F_{ik} \times F_{jk}}{\sqrt{(\sum_{k=1}^{N} F_{ik}^2)(\sum_{k=1}^{N} F_{jk}^2)}} \]

Where \( F_i \) represents the vector of word frequency in user I’s speeches in the vector space model (VSM). \( F_j \) represents the vector of word frequency of keywords in user j’s speeches. \( F_{ik} \) represents the word frequency of the kth keyword in the speeches of the ith user. \( F_{jk} \) represents the word frequency of the kth keyword in the speeches of the jth user. \( N \) represents the total number of keywords. Therefore, we can obtain the matrix of the degree of member’s support \( S \), which represents the degree of mutual support among group members. And it is also proved that the value of \( S_{ij} \) is between \([0,1]\).

3 Crucial techniques in CSCAC Model

3.1 Building the domain ontology

Domain ontology is to describe domain ontology, which is one of the bases for CSCAC model in analysis. It is realized by the concept software—EasyThinking Cognitive Assistant, which is developed by the Knowledge Science and Engineering Institute at Beijing Normal University.

In this software, a concept is shown by the node in concept map and described in the word that can best represent the concept. The relation between concepts is shown by the arrow line connecting two nodes. The relations between concepts usually include hypernym, hyponym, attribute, part-of, agent, dative, time, site, etc.

3.2 Semantic mapping between keywords and concepts

In order to compare the keywords in members’ speeches with the concepts in the concept map, we need to calculate the semantic similarity of words and those that represent the concepts to be compared with in the concept map. If the similarity is more than a threshold, which is usually taken as 0.8, we will consider the compared words are able to represent the corresponding concepts in the concept map.

The semantic similarity can be calculated by How Net, which is knowledge database that mainly digs out the relations between concepts and the characteristics of each concept based on the sememes both in English and Chinese.[9] The semantic similarity between two words can be calculated by How Net and the algorithm can refer to article [10].

3.3 Keywords Extraction
Keyword extraction is the most basic part of processing natural language. As we need to use keywords to represent the knowledge of groups and members, we consider all nouns, verbs, adjectives except some stop words in the text as the keywords. By this definition, the precision of extraction can be considered the precision of the words splitting as. According to the report of National 973 Evaluative Team in machine translation project, the precision in Chinese character has been beyond 97% now, so the extraction is valid. The main factors that affect the experimental results are the correct percentage of word categorization and the degree of word base perfection. So as the experimental shows, we can believe that the keywords extracted by the keyword extraction process are able to represent the knowledge of groups and members. With this algorithm, we develop the module of keyword extraction.

4 Experimental Study

In order to prove the validity of CSCAC model in collaborative learning, we design the following experiment. First of all, we develop the tool with C# to support the computation about the Collaborative Knowledge Building. Then we designs a collaborative task on the topic of “Collaborative Composition”, whose content is that each group discuss sufficiently in the chat room on the topic of “Who is the excellent teacher mostly needed in China” according to the assigned materials and compose a short essay of 300 words collaboratively within 1 hour. After this, the teacher designs the domain ontology of the topic of this task, as shown in the concept map of figure 2.

![Figure 2. Domain Ontology of Collaborative Task.](image)

Then the teacher logs in the platform to distribute the task to 5 groups. Each group have 4 students, which are selected randomly from a class of 75 students, and organizes the groups to carry out collaborative learning on the task. After the task accomplish, we have a survey in each team about the member contribution. In this survey, everyone will assess the rank of his/her team member’s contribution according to their impact in this task. We take the means of the result of all members in a team as the member contribution value ranked by manual. In addition, we evaluate manually the interactive text of whole group according to their relevance and inconsistence. After this, we run our tool to statistic all value. The detail is shown in Table 1.

Thus, we can compare the rank of Group Building Level and Member Contribution between the manual and auto evaluation result in SPSS. We take Kendall’s W Test to compute the consistence of result between manual and auto evaluation to confirm the
reliability of our model. The consistence in evaluation of Group is Kendall's W(a)=0.950 Asymp. Sig.=0.107, And the consistence in evaluation of Member: Kendall's W(a)=0.840 Asymp. Sig.=0.032. The result proves that the evaluation result in Group Building Level and Member Contribution by CSCAC model is related to the manual evaluation highly.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>5(0.60)</td>
<td>0.76</td>
<td>0.80</td>
<td>A 2</td>
<td>1(0.50)</td>
<td>0.70</td>
<td>0.28</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B 3</td>
<td>2(0.50)</td>
<td>0.69</td>
<td>0.28</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C 4</td>
<td>3(0.47)</td>
<td>0.72</td>
<td>0.20</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D 1</td>
<td>4(0.47)</td>
<td>0.67</td>
<td>0.24</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2(0.74)</td>
<td>0.79</td>
<td>0.94</td>
<td>A 1</td>
<td>1(0.57)</td>
<td>0.77</td>
<td>0.36</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B 2</td>
<td>2(0.52)</td>
<td>0.77</td>
<td>0.23</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C 3</td>
<td>3(0.50)</td>
<td>0.71</td>
<td>0.24</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D 4</td>
<td>4(0.47)</td>
<td>0.73</td>
<td>0.17</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1(0.79)</td>
<td>0.78</td>
<td>1.01</td>
<td>A 1</td>
<td>1(0.61)</td>
<td>0.72</td>
<td>0.45</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B 3</td>
<td>3(0.50)</td>
<td>0.75</td>
<td>0.22</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C 4</td>
<td>4(0.30)</td>
<td>0.53</td>
<td>0.05</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D 2</td>
<td>2(0.53)</td>
<td>0.76</td>
<td>0.28</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3(0.73)</td>
<td>0.77</td>
<td>0.95</td>
<td>A 3</td>
<td>2(0.51)</td>
<td>0.74</td>
<td>0.25</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B 2</td>
<td>3(0.50)</td>
<td>0.74</td>
<td>0.22</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C 4</td>
<td>4(0.44)</td>
<td>0.71</td>
<td>0.14</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D 1</td>
<td>1(0.59)</td>
<td>0.75</td>
<td>0.40</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4(0.65)</td>
<td>0.79</td>
<td>0.83</td>
<td>A 4</td>
<td>4(0.42)</td>
<td>0.68</td>
<td>0.14</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B 2</td>
<td>1(0.61)</td>
<td>0.76</td>
<td>0.42</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C 1</td>
<td>2(0.52)</td>
<td>0.77</td>
<td>0.24</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D 3</td>
<td>3(0.47)</td>
<td>0.70</td>
<td>0.20</td>
<td>0.32</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Experimental Result.

At last, we visualize the result above in bar chart and visualize the Member Support by network graph. These chart can be presented to the members timely. They are shown in figure 3. In the figure 3(C), we choose a team’s result randomly as a sample. As the figure illustrates, the edge describes the support among the members. The higher the member support mutually, the wider the edge will be. According to this graph, we can observe the
support among the members clearly.

5 Conclusion

This paper analyzes interactive text contents automatically by constructing an analysis framework of collaborative knowledge building. This method is able to give timely feedbacks on the collaborative situation of each group in collaborative learning to teachers and students, which will advance further collaborative learning effectively. And the experiment proves it to be practical.

The research for the next step is to collect more data to prove effectiveness of each dimension, especially about member support, to promote the automatic analysis results from aspects of perfecting analysis model, elevating the precision of keyword extraction, taking more semantic method such as analyzing the relations between concepts in domain ontology and so on.

References

[8] Giouvanakis Thanasis; Fragidis Garyfallo; Paschaloudis Dimitrios; Tarabanis Konstantinos; Guoqing Zhao; Embedding a Vocabulary-based Application for Concept Mapping into a Learning Management System Open Education Research Vol.11, No.5, P38-43, 2005.