Integration of a System for discussions study in Text-Based Computer-Mediated Communication

Kiriakos Patriarcheas and Michalis Xenos
Hellenic Open University, School of Sciences & Technology, Computer Science,
13-15 Tsamadou Street, GR 26222,
Patras, Greece
{k.patriac, xenos}@eap.gr

Abstract—This paper focuses on Integration of a System for discussions study in Text-Based Computer-Mediated Communication. In this paper the content category is presented as a unit of analysis for the decodification of messages at the asynchronous distance education fora which is incorporated in the modelling in a formal language and the development of a respective system created by Hellenic Open University for this purpose.

Keywords—E-Learning; Asynchronous distance education fora; Content analysis; Formal language; Modelling

I. INTRODUCTION

During the last years, an increasing number of educational institutions, as well as companies, apply asynchronous educational services via internet. One of the means utilised in distance education during the last decade is the electronic fora (fora hereinafter). Research efforts on distance education fora, at an international level, begun during the '90s. However, this is a dynamically formed field, requiring constant updating and redefinition. Given, also, that the fact of the practice of distance education during the last years has acquired new features, both in its methodology and in the tools which are utilised, the further exploration of this field becomes necessary.

The structure of this article is as follows: the theoretical framework section is a short description of the respective assignment on the content analysis technique of asynchronous discussions at distance education fora. The role of the fora of Hellenic Open University (HOU) concerning the educational procedure is described. The unit of analysis which was used is then presented, followed by the integration of the message content category as a unit of analysis in formal language. A presentation of the system of automatic text classification and the association with the message content category follows, including the description of the results of the experiments performed to control system operation. The paper closes with a discussion on the necessity of this system and the conclusions of this article.

II. THEORETICAL FRAMEWORK

Although the researchers seem to agree that collaboration may encourage the learning procedure [1], there is no clear theory available to guide research on computer mediated interaction [2], empirical markers which shall be the base of a codification tool as a standard against which to evaluate whether or not effective learning is occurring through the online discussions [3]. In the last few years, numerous efforts to approach this issue were made, stemming from different theoretical backgrounds.

As is shown from all the above, an important issue arising is the unit of analysis which shall be used for the content analysis. Fahy et al. [4] consider each single sentence as one unit of analysis, and Pena-Shaff and Nicholls [5] uses the sentence as unit of analysis, trying to approach it at a paragraph level. Others choose the definition thematic unit (or otherwise of a “theme” or an “idea”) to be their unit of analysis [6-8]. Another approach [9-11] is to consider the whole message that a student enters at a specific moment in the conversation as the unit of analysis. Jarvela and Hakkinen [12] choose a Complete discussion, while during the last years there has been an approach of multiple point both at a micro and at a macro level [13]. Further down, a comprehensive review is presented in a table form (Table I), referring to the unit of analysis used by this field’s researchers.

© IEEE 2012. This is the authors’-created version of the paper. It is posted here for your personal use, not for redistribution. The IEEE-published version is available at doi: 10.1109/PCi.2012.44

TABLE I. OVERVIEW OF THE CONTENT ANALYSIS SCHEMES

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Theoretical background</th>
<th>Unit of analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henri (1992)</td>
<td>Cognitive and metacognitive knowledge</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>Newman et al. (1995)</td>
<td>Critical thinking</td>
<td>Thematic unit</td>
</tr>
<tr>
<td>Zhu (1996)</td>
<td>Theories of cognitive and constructive</td>
<td>Message</td>
</tr>
<tr>
<td></td>
<td>learning – knowledge construction</td>
<td></td>
</tr>
<tr>
<td>Gunawardena et al. (1997)</td>
<td>Social constructivism – knowledge construction</td>
<td>Message</td>
</tr>
<tr>
<td>Fahy et al. (2000)</td>
<td>Social network theory – Interactional</td>
<td>Sentence</td>
</tr>
<tr>
<td></td>
<td>exchange patterns</td>
<td></td>
</tr>
<tr>
<td>Lockhorst et al. (2003)</td>
<td>Social constructivism – learning strategies</td>
<td>Thematic unit</td>
</tr>
<tr>
<td></td>
<td>construction</td>
<td></td>
</tr>
</tbody>
</table>

III. METHODOLOGICAL FRAMEWORK

A. The unit of analysis which was used

Given that the choice of a unit of analysis is dependent on the context and should be well-considered, because changes to the size of this unit will affect coding decisions and comparability of outcome between different models [14], as well as given the fact that Schrire [15] refers to a dynamic approach in which data is coded more than once and the grain size of the unit of analysis is set, depending on the purpose and the research question, it was decided not to take into consideration the discussion thread, not even the message as unit of analysis, nor the paragraph or the single sentence.

It was decided to use as unit of analysis, the category of the message’s content, as for the observation of the discussion threads, it was noticed that there are cases of messages which may comprise two (or/and more) content categories, e.g. a question about the next advisory meeting and a reply to a question concerning the study of the educational material [16-17]

According to the study of the messages of INF10 for academic years 2005-2012 the messages as for their content may concern (in brackets you see the respective symbols used in formal Language):

a) study of educational material (M),
b) questions/answers for exercises – assignments (X),
c) presentation of sample assignments by tutors (P),
d) instructions (I),
e) assignment comments, corrections (F),
f) student comments on assignments (D),
g) sending – receiving assignments (J),
h) sending - receiving grade marks (G),
i) notification of advisory meeting (V),
j) pointless message (L).

B. Integration of the category of the message context as a unit of analysis in formal language

According to the study of the messages on HOU’s fora during the academic years, a Language was developed, which is defined by mathematic terms and represents the messages using as unit of analysis the category of the message content. More specifically:

- There are two categories of communication’s carriers: a) Tutors, b) Students
- For brevity reasons, tutors shall be symbolized with T and students with E
- As for the type of message, they are discerned to questions and replies. Having the symbol q and a respectively.
- As for their content category, we have the symbols aforementioned in the previous section: M, X, P, I, F, D, J, G, V.
- Finally, the order in which the above symbols appear is: a) the message carrier, b) the type of message and c) the content category to which the message belongs.

Thus, the Language contains:

a) Terminal symbols alphabet VT, where VT = {T, E, q, a, n, M, X, P, I, F, D, J, G, V, L }

b) Non terminals alphabet VN, where VN = {u, r, y, c}, more specifically:
   
   r: represents the message carrier (where T for tutors and E for students)
\textit{u}: represents a pair \( yc \) i.e. a message type \( y \) (whether it is a question \( q \) or an answer \( a \)) followed by its content category.

c) The grammar \( P \)

A set of rules of the form \( \alpha \rightarrow \beta \), where \( \alpha \) and \( \beta \) sequences containing terminal and non-terminal symbols and \( \alpha \) is not an empty sequence, as follows:

1. \( S \rightarrow ruS \)
2. \( S \rightarrow e \)
3. \( u \rightarrow uyc \)
4. \( u \rightarrow e \)
5. \( r \rightarrow T \)
6. \( r \rightarrow E \)
7. \( y \rightarrow e \)
8. \( y \rightarrow q \)
9. \( y \rightarrow a \)
10. \( y \rightarrow \epsilon \)
11. \( c \rightarrow P \)
12. \( c \rightarrow I \)
13. \( c \rightarrow J \)
14. \( c \rightarrow F \)
15. \( c \rightarrow G \)
16. \( c \rightarrow D \)
17. \( c \rightarrow M \)
18. \( c \rightarrow X \)
19. \( c \rightarrow V \)
20. \( c \rightarrow L \)

Where \( \epsilon \) stands for an empty symbol

d) Symbol \( S \) where every sentence generated starts with this symbol.

IV. THE SYSTEM - ASSOCIATION OF THE TIME WITH THE CONTENT CATEGORY OF THE MESSAGE

According to this approach, a system of automatic classification was developed, which comprised the following:

a) Data filtering: where some web pages are considered as input accommodating the discussion threads of a distance education forum of HOU (which include much data having no essential information concerning the educational procedure e.g. titles, images etc.) and creates a temporary file with the “useful” part (User name, date, message’s content) which may become a source of information for educational conclusions.

b) Storage of roots files: which is a dynamic way according to which word or phrases or symbols roots are stored, as well as the respective terminal symbols \( q \) if it is a question or \( a \) if it is an answer. The same was also done for the storage of information necessary in the determination of content category of a message, i.e. if it is about study, assignment, comment etc. or combination of them (e.g. a message concerning both the study and an assignment). To wit, it takes as input couples of information of the type root of a word or phrase and terminal symbol of the content category (\( M, X, P, I, F, D, J, G, V, L \)). As it is obvious, the system provides the ability to add further content categories if necessary.

c) Strings’ production: receiving as input the temporary file with the “useful” information (User name, date, message’s content) and the files with the couples of roots words/ phrases/ symbols and terminal symbols and presents (and stores) the respective strings with the relative extensible file, so as the results to be kept for further exploitation.

It is worthy to note here that this specific system incorporates the sense of time along with its association with each of the nine (9) categories of message content chosen as unit of analysis. More specifically and given that within a message (as it is deduced both from literary review and from the observation of the fora of HOU) more than one contents may exist, the dates are recorded for each such case and not simply in each message.

In fact after each couple \( yc \) there is a date's record. Certainly, so as to effectuate the above procedure nine (9) stacks were used – as many at the message’s content categories, each one having as many figures as the number of appearance of terminal symbols (\( M, X, P, I, F, D, J, G, V, L \)) resulting from the non terminal symbol \( c \). Consequently, time differences may automatically exist (in days, if from each current date, by content category, it is deduced the previous one) and thus there may arise another nine (9) respective stacks with the above date references. Of course, the length of these stacks is equal to the length of dates minus one (-1), i.e. apart from the initial message, which is considered to be the point zero (0), where the numbering of the time differences begins. The contents of the stacks of time differences may constitute an important criterion, which may participate as such (in combination with other criteria) in case of evaluation of a forum's consequences to the educational process.

In this point, it is deemed necessary the results presentation of algorithm department by which is produced the results that are reported in: number of tutor and students interventions, time differences (between the messages) and number of message content category per student, as follows (Table II):

© IEEE 2012. This is the authors’-created version of the paper. It is posted here for your personal use, not for redistribution. The IEEE-published version is available at doi: 10.1109/PCI.2012.44

<table>
<thead>
<tr>
<th>error</th>
<th>i</th>
<th>Counter</th>
<th>j</th>
<th>Stuck</th>
<th>Students Interventions</th>
<th>Tutor Interventions</th>
<th>symbol</th>
<th>error</th>
<th>i</th>
<th>Counter</th>
<th>j</th>
<th>Stuck</th>
<th>Students Interventions</th>
<th>Tutor Interventions</th>
<th>symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>3/3/2007</td>
<td>T FALSE</td>
<td>X</td>
<td>1</td>
<td>1</td>
<td>204/2012</td>
<td>T FALSE</td>
<td>X</td>
<td>1</td>
<td>1</td>
<td>7/5/2012</td>
<td>T FALSE</td>
</tr>
<tr>
<td>TRUE</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>7/5/2012</td>
<td>T FALSE</td>
<td>X</td>
<td>1</td>
<td>1</td>
<td>255/2012</td>
<td>T FALSE</td>
<td>X</td>
<td>1</td>
<td>1</td>
<td>12/5/2012</td>
<td>T FALSE</td>
</tr>
</tbody>
</table>

**TABLE II. DATA TABLE FOR THE SEQUENCE TAXEgXTaXaIEgXTaXaLEaXTaXeQXTaPTaGFTaG**
Also, the final form that will have the 9 stacks with the dates they will be as follows (Table III):

<table>
<thead>
<tr>
<th>j</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15/5/2012</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5/5/2012</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>25/5/2012</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>25/5/2012</td>
<td>28/5/2012</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While the counter table (Table IV) has the following contents:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

As for the time differences stacks (Table V) has the following form:

| Content Category | Appearances number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------------|--------------------|---|---|---|---|---|---|---|---|---|---|
| M                | 0                  |   |   |   |   |   |   |   |   |   |
| X                | 10                 | 9 | 1 | 3 | 1 | 1 | 0 | 0 | 2 | 0 |
| P                | 1                  |   |   |   |   |   |   |   |   |   |
| I                | 1                  |   |   |   |   |   |   |   |   |   |
| F                | 1                  |   |   |   |   |   |   |   |   |   |
| D                | 0                  |   |   |   |   |   |   |   |   |   |
| J                | 0                  |   |   |   |   |   |   |   |   |   |
| G                | 2                  | 3 |   |   |   |   |   |   |   |   |
| V                | 0                  |   |   |   |   |   |   |   |   |   |

V. CONCLUSIONS

The practice of distance education during the last years has acquired new features, both in relation to methodology and in the tools it uses. It is also a fact that the subject of electronic fora in distance education is a dynamically formed field requiring constant updating and redefinition. A big part of the researches presented in the international literature concerning distance education’s fora, refer to the content analysis, which principally aims despite the fact that this research technique is frequently used, though there are still no standards established. There is a variety of approaches, varying both at detail’s level and at the type of categories of analysis they use. As it was deduced from the above presentation and study of the discussion threads of HOU, it was noticed that there are cases of messages which may comprise two (or/and more) content categories, e.g. a question about the next advisory meeting and a reply to a question concerning the study of the educational material. For this reason, this paper uses the content category as unit of analysis for the messages’ interpretation in Asynchronous distance education fora and for this purpose incorporates it in a modelling in a formal language. Furthermore, time indexes of participation were integrated in combination with the content categories of the message, in order to define the way these elements could improve the capacity of the tutor to evaluate the progress of a discussion thread in a distance education forum.

Among others, the prediction for future research actions are long-term studies concerning the main issue: what reinforces the participation at fora and how this contributes to the educational process effectiveness by investigating side questions, such as how much it affects the person who starts the thread (tutor or student), how it starts, the period when the thread starts, how important the time of response.
in threads, is the groups’ size etc. and their association with the elements concerning the students’ profiles and their performance in course modules of HOU, intending to reach educational conclusions.

REFERENCES


