2009 Ninth International Conference on Intelligent Systems Design and Applications

Time Analysis of Forum Evolution as Support Tool for E-Moderating

Carlo Alberto Bentivoglio
Dipartimento di Scienze dell'educazione e della formazione
Università di Macerata
Macerata, Italy
ca.bentivoglio@unimc.it

Abstract— **The web forum is a key tool in new knowledge building among students in Learning Management Systems. Unfortunately, the huge number of messages makes difficult, for tutors and teachers, to correctly evaluate the evolution of the forum and its efficacy in the learning process. In order to support the tutors in this effort, a solution, based on simple statistical indices inspired by the work in the text analysis and social network analysis field, is proposed. The obtained results show good performance with a minimum computational effort.**

Text-mining; statistical analysis; tutoring support system

I. INTRODUCTION

The web forum is a tool where opinions about some topics of common interest are shared between peers. In Learning Management System (LMS), according to a constructivist view of learning, as reported by Jonassen et al. in [8] and by Scardamalia and Bereiter in [14], the forum is the place where knowledge is restructured among the students' community.

The students' opinions, in fact, build up a relationship network which supports a collaborative reflection about the discussed topic. Such process is not spontaneous but must be catalyzed by the external intervention of the tutor that posts the seminal message, promotes valuable arguments, closes off unproductive discussion and so on (all these activities are well expressed by the Salmon's neologism "E-moderating" [13]).

In order to enhance the efficacy of the tutor's action the forum evolution must be monitored. Unfortunately, the huge number of messages and the tutor's subjectivity makes difficult to correctly evaluate such process. An automated support is needed as observed by Dringus and Ellis [6].

Supporting the tutor using data mining and text mining can be done in many different ways. At a very basic level, patterns of users' behavior can be analyzed in order to optimize resources allocation and plan the tutorial activities (moderated discussion, test, assessment,…) according to the users' time constraints (Burr and Spenneman [4]).

A step ahead, interaction between students can be observed (see, for example, Barros and Verdejo [2], de los Angeles Constantino-Gonzalez et al. [5] and Israel and Aiken [7]). Again, the focus of conversation in the discussion can be analyzed (Kim et al. [9],[10]).

On the other hand, Rosé et al. in [11] adopt a more linguistic oriented analysis in order to track the evolution of collaboration.

Our approach, instead, is based on metrics, inspired by the textual analysis and social network analysis, that describe the forum in terms of global vocabulary evolution and commitment of the users as introduced in Rossi et al. in [12].

In fact, at a more high level abstraction, the forum dynamic can be seen as a succession of different phases where the discussion productiveness changes. The very idea is to use quantitatively metrics in order to measure the necessary conditions characterizing each phase. In this way, the tutor, or a rules engine, checks the indices and estimates, probabilistically, the phase in which the forum is going to be; i.e. if the students are discussing together or have simply posted a message to answer to the tutor's seminal message. This awareness make possible, for instance, to optimize the time dimension since the unproductive phases are compressed.

The paper is organized as follows. In section 2 the forum model is proposed while in the next section the corresponding text analysis tools are discussed. In section 4 a case study is described. In the final section conclusions are discussed and future directions are proposed.

II. FORUM EVOLUTION MODELING

A discussion forum can be seen as a virtual place where a community of users interacts posting messages. The system evolves as the number of messages grows. It starts from an equilibrium state where no messages are posted. Then, when the first message is posted, a discussion thread begins and the system goes through a disequilibrium state where the community interacts answering to the initial message and to the following ones. When no more messages are appended to the discussion thread, since nobody has anything more to say, the system reaches a new equilibrium state.

More precisely, four main phases can be identified. First of all there's a starting phase where the user begin to answer to the initial message of the thread. In this phase, the number of messages is comparable with the number of new users that post for the first time in the thread. Then, the discussion begins and the number of messages and new relations abruptly rise up (discussion phase). Such phase can evolve in two really different ways. The first one is a consolidation

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.66

IEEE computer society

phase. In this case the discussion became more specialized and is carried on inside a clique of users. So it is characterized by a lot of new messages and few new relationship. The second one is a starving phase. It can happen because the forum has reached the final state or the community is not interested in discussing the proposed topic; anyway, the tutor should act.

In order to track the forum evolution two kinds of analysis are possible. The former is an equilibrium analysis that can be done a posteriori when the discussion is considered ended and can be used as an evaluation of the work done by the community. The latter is a dynamical analysis that must be done when the discussion is open and helps tutor in their action. The dynamic analysis can be done studying the same indexes of equilibrium analysis as they vary in time. Indexes computation can be done on a daily basis and scheduled in the less community activity periods. In this way the daily temporal dynamic can be seen as sequence of equilibrium states.

The forum evolution is studied according to a forum model that takes in account different aspects of its nature. According to the dynamic nature of such tool two kind of measure are used: cumulative (from day 0 to day t) and punctual (just in day t). In order to simplify notation, punctual measures are expressed in lower-case letters while cumulative in upper-case. For instance, said $M(t)$ the set of messages posted up to day t and $|M(t)|$ its cardinality, the set of new messages posted just in that day is $m(t)$. The generic message is said m. The following relationship holds:

$$|m(t)| = |M(t)| - |M(t-1)|. \quad (1)$$

Generally speaking, a web forum is a collection of messages M written by a set of users U and disposed in discussion threads T in a tree like disposition consisting in an initial message, the messages answering to it, the following messages and so on.

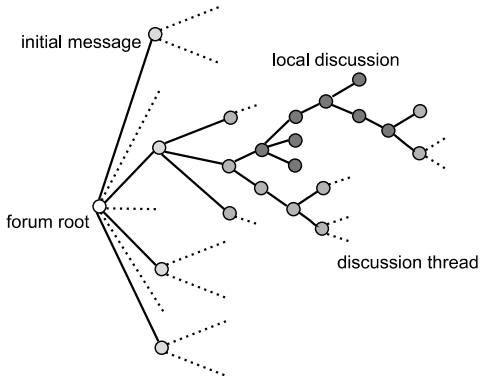

Figure 1. Forum structure and definitions.

A message is a ordered sequence of word-token. A word-token is a delimited string of character and its set is called N. A set of word-tokens, equal among them, are the instances of a word-type. The set of word-types is the forum vocabulary V. The vocabulary can be partioned in subsets V_n where n is the number of occurrences of the word-types belonging to it (e.g. V_4 is the set of word-types occurring only four times in the forum).

Particularly interesting is the set of halomorphemes (also known as hapax legomena or hapax for short) V_1, i.e. the word-types occurring only once.

Moreover, sequences of word-tokens delimited by periods are called sentences. The set of sentences of the generic message m is said s_m.

Because a corpus of documents and web forum are very different objects, some basic assumptions must be made to bridge the gap between them. The main assumption is the correspondence between the text in a corpus and the discussion thread in a forum. The next assumption is the partial correspondence between fragments in a text and messages in a thread. With this two assumptions is possible to extend the text analysis tool to the web forum analysis.

The connection with social network analysis is more straightforward. When a user i answers to the message posted by the user j a new relationship r_{ij} is considered born. More formally, the generic r_{ij} element of the matrix R is the number of messages that i answers to the message of j. The number of elements of R greater than zero is called R_+.

III. Description of Adopted Indexes

The tools used by the tutors to monitor the forum evolution rely on indices derived from textual analysis and social network analysis. The first group of indices are global lexical parameters that should help the instructors to have a global view of the vocabulary evolution of the forum. This strategy, in fact, is based on the hypothesis that the vocabulary evolution indirectly reflects more complex phenomena such users' commitment and the birth of a common conceptual space. The first index is related to the lexical extension L. It is defined as:

$$L(t) = \frac{|V(t)|}{|N(t)|} \cdot 100. \quad (2)$$

The second is the hapax percentage H. Its definition is the following:

$$H(t) = \frac{|V_1(t)|}{|V(t)|} \cdot 100. \quad (3)$$

As reported in Accorroni and Bentivoglio [1] such indexes are useful to detect the maturity level of discussion. More precisely are strictly correlated with the transition from the starting phase and the discussion phase.

A second group of indexes is related to the message characteristic in terms of structural complexity. The first index is the average number of words per sentences at day t. It is computed making the ratio between the overall number of word-tokens and the overall number of sentences in the messages posted at day t. Formally, it is defined as:

$$n_s(t) = \frac{|n(t)|}{\sum_{i=1}^{|m(t)|} |s_i(t)|}. \quad (4)$$

Because the number of word-tokens are usually a necessary condition for the syntactic complexity of a sentence the n_s index is designed to show the degree of such characteristic. In fact, if the value is high, the sentence generally shows a more complex and argumentative linguistic structure. The second index is the average number of words per message. It is simply defined as the ratio between the overall number of word-tokens and the messages at day t:

$$n_m(t) = \frac{|n(t)|}{|m(t)|}. \quad (5)$$

The last one is the average number of sentences per messages and it is defined as:

$$s_m(t) = \frac{\sum_{i=1}^{|m(t)|} |s_i(t)|}{|m(t)|}. \quad (6)$$

The last two indexes monitor the users interaction. It is supposed, in fact, that a high value of this ratios shows less attention to the listening and, consequently, a poor interaction. Because interaction is a key ingredients of the consolidation phase, the decreasing of such indices is a clue that the forum is approaching that phase.

Another possibility is to face directly the users' commitment in the forum. The first indexes are the users number and the messages number. Their cumulative distributions in time, when compared together, are a possible indicator of the current phase in which the forum could be. If the messages per user is growing probably the forum is far from the starting phase or the starving phase. Anyway, the analysis can be smarter. The previous indicator tell us that a discussion has begun but they said nothing about the quality of the discussion from a relational point of view.

In fact, fixed the number of messages that a user can post, a student can behave in two ways: answering to a lot of different students (making a lot of relationships) or concentrating the discussion with a limited number of peers. Moreover, this behavior can change over the time. In the first part of the discussion phase the number of relations grows, then, when the social relationships are well established, the messages begin to be exchanged only in restricted groups and, probably, the quality of the discussion get higher. The forum enters the consolidation phase.

In order to track this process and identify the prevalence of one behavior over the other the message-relationship ratio can be used. The index is the ratio between the number of messages posted to people that are yet in the clique and the number of users. More formally the $MR(t)$ index is defined in this way:

$$MR(t) = \frac{|M(t)| - R_+}{|U(t)|}. \quad (7)$$

Generally, in the initial phase of forum the *MR* index drops because students must build their relationship network instead making deeper discussion. Then, albeit R_+ continues to grow, *MR* rises on or, more frequently, becomes stable because the students begin to exchange messages with the same people. In this way, the level reached by the *MR* is a quantitative index of the relationship that a generic student has with its clique. More precisely, it is the average number of messages that a student exchanges with people that has already met in the forum.

From the tutor perspective the *MR* index is useful in two way. Firstly it gives a measure of the interaction level of the users. Hence, stated empirically a interaction level under which the discussion is considered poor, it can be used as a objective to reach. In this way, it can be also used as a one of the feedback parameter to evaluate the results of tutor's effort to make the discussion level higher. In second instance, it can be used to distinguish the phase in which the forum is moving due to the presence of transition in the *MR* plot.

IV. Case Study

In the first phase 26 forums, belonging to the LMS adopted by four faculties of the University of Macerata (1200 online students), were analyzed. Two groups, according to the number of users, emerge from the analysis: the first one, 18 forums, with a mean of 7 users and 58 messages. The second one, 8 forums, with a mean number of 152 message and 48 users. Every forum lasts on average 16 to 35 days.

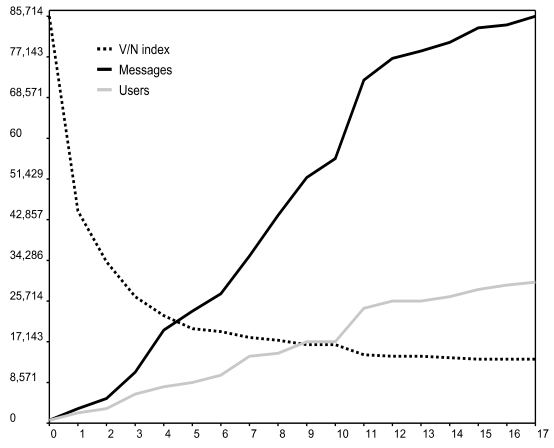

Figure 2. Lexical extension evolution

The first result was the regularity of lexical extension index in its asymptotic behavior, 30.82% (5,25 standard

deviation) in the first group and 15.25% (2,47 standard deviation) in the second one. The observed behavior of the lexical extension suggests two possible use. In fact, when the index get closer to the equilibrium value, it has been observed that the discussion phase begins to start. This is quite useful because it can be used by tutors as a feedback parameter of their facilitation activity.

On the other hand, if the index has reached an equilibrium state and the discussion is not productive it can be the sign that the discussion is going to end but it was quite useless so a new action must be strongly taken by the tutors.

A regularity was found also in the hapax percentage. In the first group the value was 57% (5 standard deviation) while in the second was 49,55% (3.1 standard deviation). This a interesting result because it suggests to us to study how the use of hapax is distributed among the user. Evaluating a normalized entropy measure of hapax per user vocabulary can be useful to detect disequilibria. A low value of this parameter, i.e. few people use a lot of hapax in their messages, could be seen as an evident sign of scarce attention of the community to the topics proposed by a part of it.

Another interesting behavior is the the s_m index tendency to decrease when the analyzed forums reaches the consolidation phase. Such index, in fact, identifies the level of interaction in the discussion. More precisely, albeit the index time plot tends to have a quite erratic behavior, its envelope tends to decrease in time as the forum approaches a consolidation phase.

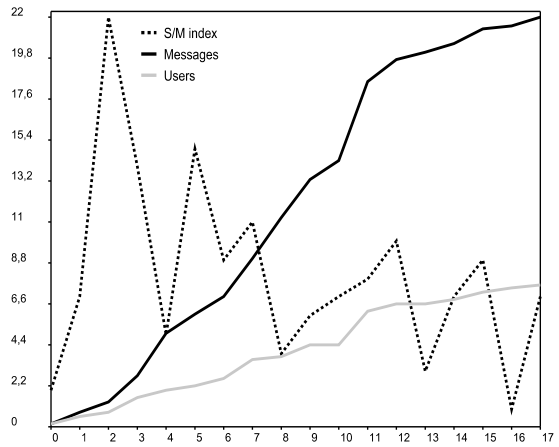

Figure 3. Sentences per message evolution

Moreover, such phenomenon starts in the discussion phase and enters in a well identified oscillation band whose range is between 3 and 10 sentences per messages. This index is very useful for the tutors. In fact, if the index is out of range and, at the same time, the other indexes show that the forum has entered the discussion phase, the tutors can identify the people, looking at their message length, supposed to pay less attention to the other message. Then, they can help them to change their behavior.

Another way to observe the interaction between student is the *MR* index. It has been noted that the index follows a general behavior consisting in two phases. At the beginning the index reaches a peak then, as the discussion phase begins, it drops dramatically. In the second phases its evolution can be of two kind: it reaches a local maximum and enters a band where it maintains a nearly constant value or, in the second case, it continues to rise until the forum ends.

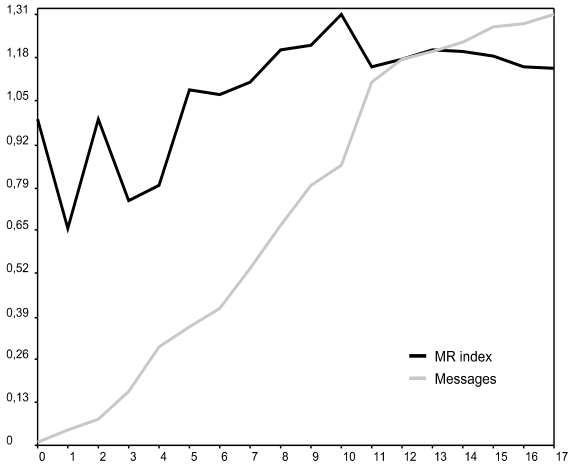

Figure 4. MR index evolution

Taking in account such behavior, the central idea is to study the peak reached by the index when the discussion phase has reached maturity. This helps to evaluate the commitment of the student with their clique. From an experimental point of view, in the first group the mean was 3.5 (1.75 standard deviation). In the second group of forum the mean was 1.4 (0.8 standard deviation).

The reported mean value is a clue that is quite difficult to maintain a consistent level of interaction with the others. In fact, in the second group, only 1.4 messages are exchanged with known people in the forum. In order to get deeper in this phenomenon, further studies are planned such, for instance, the definition of an entropy index of the intensity of relationship per student. In this way, it will be possible to identify students who have an intense relationship with only few person or, on the other hand, people that maintains no significant relationship.

V. Discussions and Further Work

The proposed metrics are used in a the wider context of a monitoring system based on the JADE multi agent framework, presented by Bellifemine et al. in [3]. The elaboration is scheduled daily and the obtained data are shown to the tutor on a personalized panel. Albeit the visual inspection of plotted data was a quite tiring experience for the tutors, the feedback about the new tools was fairly good and, along with such subjective opinions, the main result was that the discussion phase, thanks to the awareness on forum dynamic gained by the tutor, was slightly anticipated respect the previous forums (generally two or three days). This result could be regarded as the clue that tutors, thanks to a quantitative feedback, can know when and how to stimulate the discussion in a more effective way.

The next step is to use the data to feed a rule based engine that alerts the instructor when a certain event happens

in the forum. More precisely, the forum can be depicted as a system that goes through a limited number of phases of different activity level. The indexes should be used to identify such phases. In this way, instead of numerical data only, a natural language report about the current state of the forum could be displayed to the tutor.

Moreover, the acquired data can be used to feed a personal agent who interacts with the student as a chat-bot. The agent uses the data about the student's behavior and the current phase of the forum. With this two sources of information it can identify divergence of behavior between the user and the community.

This is an enhancement in tutoring because the approach of the human tutor is mainly direct towards the community as a whole while a software agent is more oriented to the individual and doesn't take in account the context where the user operates. With this approach the intervention of the agent is more contextualized. For instance, if the student's activity, according to his commitment in the forum, is low while the community is already in the discussion phase, the agent can help him to take awareness of this condition since the other students are more involved in the discussion.

Using a contextualized communication between agent and student a richer interaction can be reached so the final goal of stimulating a more active commitment in the forum can be better pursued.

REFERENCES

[1] M. Accoroni and C.A. Bentivoglio, "Supporting Tutoring by Improved Statistical Analysis of Discussion Forum", *Proceedings of the 2nd IEEE International Conference on Human System Interaction*, 2009, p. 463-466.

[2] B. Barros and M. F. Verdejo, "Analysing student's interaction process for improving collaboration: the degree approach", *International Journal of Artificial Intelligence in Education*, 2000, 11, p. 221-241.

[3] F. Bellifemine, A. Poggi and G. Rimassa, " Developing multi-agent systems with a FIPA-compliant agent framework". *Software: Practice and Experience*, 2000, 31(2), p. 103-128.

[4] L. Burr and D. HR Spenneman, "Patterns of User Behavior in University Online Forums". *International Journal of Instructional Technology and Distance Learning*, 2004, 1(10), p. 11-28.

[5] M. de los Angeles Constantino-Gonzalez, D. D. Suthers and J. G. Escamilla de los Santos, "Coaching Web-based Collaborative Learning based on Problem Solution Differences and Participation", *International Journal of Artificial Intelligence in Education*, 2003, 13(2-3), p. 263-299.

[6] L.P. Dringus and T. Ellis, "Using data mining as a strategy for assessing asynchronous discussion forum", *Computers & Education*, 2005, 45, p. 141-160.

[7] J. Israel and R. Aiken, "Supporting Collaborative Learning With An Intelligent Web-Based System", *International Journal of Artificial Intelligence in Education*, 2007, 17(1), p. 3-40.

[8] D.H. Jonassen, K.L. Peck and B.G. Wilson, *Learning with technology. A constructivist Perspective*, Prentice Hall, Upper Saddle River, 1999.

[9] J. Kim, G. Chern, D.H. Feng, E. Shaw and E. Hovy "Mining and assessing discussions on the web through speech act analysis", Proceedings of the ISWC'06 Workshop on Web Content Mining with Human Language Technologies, 2006

[10] J. Kim, E. Shaw, D.H. Feng, C. Beal and E. Hovy, "Modeling and Assessing Student Activities in On-Line Discussions", *Proceedings of the AAAI-2006 Workshop on Educational Data Mining*, 2006. p. 67-74.

[11] C. Rosé, Y.C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger and F. Fischer, "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning", *International Journal of Computer-Supported Collaborative Learning*, 2008, 3(3), p. 237-271.

[12] P.G. Rossi, A. Zuczkowski, G. Alessandri, L. Giannandrea and P. Magnoler, "Linguistic indicators for the knowledge building analysis and the interaction in online learning processes" in B.M. Varisco (Ed.) *Psychological, pedagogical and sociological models for learning and assessment in virtual communities*, Polimetrica Publisher, Milan, Italy, 2008.

[13] G. Salmon, *E-moderating: The Key to Teaching and Learning Online*, Kogan Page, London, U.K. ,2000.

[14] M. Scardamalia and C. Bereiter, "Computer support for knowledge-building communities", *The Journal of the Learning Sciences*, 1994, 3, p. 265-283.