International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 3 (2011) pp. 671-678

Qualitative Analysis of Discussion Forums

Breno Fabrício Terra Azevedo[1], Patricia Alejandra Behar[2] and Eliseo Berni Reategui[3]

[1] Federal Institute of Education, Science and Technology Fluminense
273, Rua Dr. Siqueira - Parque Dom Bosco - Campos dos Goytacazes, RJ - Brazil - ZIP Code 28030-130
bterra@iff.edu.br

[2] Federal University of Rio Grande do Sul
110, Av. Paulo Gama - prédio 12105 - 3º andar sala 332 - Porto Alegre, RS - Brazil - ZIP Code 90040-060
pbehar@terra.com.br

[3] Federal University of Rio Grande do Sul
110, Av. Paulo Gama - prédio 12105 - 3º andar sala 332 - Porto Alegre, RS - Brazil - ZIP Code 90040-060
eliseoreategui@gmail.com

***Abstract*: This paper proposes a qualitative analysis of discussion forums using a particular text mining technique which uses a graph formalism to represent relevant terms found in a text, as well as their relationships. The proposed approach is based on the comparison of graphs extracted from textual contributions made by students, and important terms provided by the teachers related to the topic being discussed in forums can be analyzed by applying text mining techniques. Five experiments were carried out to evaluate how the system developed is capable of producing outcomes which may show the relevance of the forum's posts. The results of the experiments showed that the proposed text mining approach is useful to assist teachers in identifying situations that need intervention. Such results can be used by teachers to follow student discussions, contributions and to better understand how to mediate forum's discussions as to support students learning.**

***Keywords*:** discussion forums, text mining, qualitative analysis.

I. Introduction

Discussion forums allow people with common interests to debate and share certain information, questions and opinions [1]. In virtual learning environments, forums are spaces of asynchronous interaction among users [2]. These tools present a theme to be debated by learners who contribute by expressing their opinions. Their messages express a particular view of concepts related to the theme, involving individual experiences, questions, suggestions, criticism. Textual contributions posted in forums are generated from the interchange of ideas among participants, and such interactions represent an opportunity for them to reflect on the contents studied [3].

This study is relevant due to the importance of doing a qualitative evaluation of the messages registered by students in online discussion forums, that is, meaningful contributions concerning the theme being debated. Quantitative analysis is simpler to be implemented, for instance, when verifying the participation rate of learners and the extent of the discussions. However, according to Mazzolini [4], it does not represent an adequate way of portraying contributions in a forum. For teachers, qualitative analysis is essential so that they carry out a close assistance of how learners are building their knowledge. Moreover, it allows teachers to verify the real contribution of each student, observe and interfere in the discussions.

In online forums, teachers suggest a discussion topic to students who, in turn, present their reports. Depending on the topic, on the learners profiles, on the interest raised by the theme, forums may contain hundreds or only a few contributions. Some of these messages may present favorable arguments or opposing ideas concerning the topic of discussion. While some of the texts do not present relevant ideas concerning the topic, other contributions are meaningful as they deal with important concepts by either citing or relating them.

Participation in discussion forums is an important learning activity in distance education, allowing the teacher to identify important information about their pupils. For teachers, forum analysis is essential as a means to provide a closer view on how learners build their knowledge. Moreover, it allows teachers to verify the real contribution of each student, observe and interfere in the discussions. However, in the case of very large groups, analyzing students' posts can be quite time consuming. Thus, a content analysis technique may be helpful to teachers as to enable them to focus more on finding out why some students have not discussed important concepts related to the main topic.

This study aims at verifying a viable way to help teachers check relevant contributions posted in discussion forums. In this approach, teachers can observe how learners interact in the discussion and, by doing so, they may identify less significant texts. This process fosters a type of intervention in which teachers try to mediate the debate by motivating and challenging those who posted relevant contributions, and inciting those not so active in the discussion to participate more.

The next section briefly introduces text mining techniques. Section 3 presents the methodology used in the experiment, and a discussion of the results. Concluding remarks about the

study and the next steps planned for the research are presented in Section 4.

II. Text Mining

According to Feldman [5], text mining can be defined as an intensive knowledge process in which a user interacts with a great amount of documents by using tools for analyzing them. The goal is to extract useful information from collections of documents. Such collections are identified in interesting patterns in non-structured textual data.

Text mining systems are based on pre-processing routines, algorithms for pattern discovery, and elements for presenting results. In such systems, pre-processing operations are based on the identification and extraction of representative characteristics of the documents in natural language. These operations are responsible for transforming non-structured data, stored in document collections, in a structure expressed in an intermediate format [5], [6].

Operations which make up the central part of the mining procedure, also called knowledge distillation processes, represent the core of a text mining system, and include: pattern discovery, trend analysis, and incremental algorithms for knowledge discovery. Commonly used patterns in knowledge discovery are distributions and proportions, a set of frequent concepts, and associations. These operations may also be related to comparisons, and to the identification of levels of interest with some of those patters. Advanced text mining systems, which are often domain oriented, may also improve operation quality from queries made on knowledge bases [5].

Elements which constitute the presentation of results represent the system interface, with navigation functionalities, and access to the language used in the queries [5], [7], [8].

Text mining techniques include:

a) Information extraction: identifies main terms and phrases in a given text, as well as their relationships. Extraction is made by searching pre-defined sequences, a process known as pattern matching. Extraction infers relationships between terms, people, places, dates, as to provide the use with meaningful identification [9], [10].
b) Topic tracking: stores user profiles and, based on documents visualized by the user, predict other documents that might be of interest. Tracking can be used to discover references on a research area [11], [12].
c) Concept linking: connect related documents from the identification of commonly-shared concepts. This technique helps users to find information that might not be found with traditional searching methods. Conceptual links facilitate information navigation instead of helping with the search [9], [11].
d) Information visualization: presents large amounts of documents in a visual hierarchy, or map, providing navigation options, as well as search tasks [5], [7], [8].
e) Question Answering: derived from natural language processing, it is a technique which deals with how to find the best answer to a certain question [11], [13].
f) Summarization: it is a useful technique when one is trying to discover whether or not a lengthy document fits the user's needs, so that he or she can evaluate if it is worth reading the text to obtain more information. The goal of this technique is to reduce the size and level of document detail, retaining main ideas and general meaning. The greatest difficulty for the software is to perform semantic analysis and to interpret meanings. One of the most used strategies in summarization is the extraction of important sentences from the statistical weight of sentences. Some additional heuristics may also be used, such as information on position. Summarization methods can be classified in two groups: shallow approaches, which are restricted to the syntatic level of representation; and deep approaches, which involve a level of semantic representation of the original text and use linguistic processing in some level [11], [14].
g) Categorization (classification): involves the identification of the main themes in a document, placing it in a pre-defined set of topics [11]. This process deals with the text as a set of words, instead of processing real information, as it occurs in the information extraction technique. Categorization counts the words in a document and, from this count it identifies the main topics of the text. This technique often uses a method of document classification, so as to rank those which are more content-related to a specific theme.
h) Clustering: a technique used to group similar documents. It differs from categorization since it does not use pre-defined topics, but clusters documents in real time. Another advantage of clustering is that documents may appear in several subtopics, ensuring that a useful document is not omitted from the search result. A basic clustering algorithm creates a topic vector for each text, and calculates the weights in order to identify in which group a document must be placed [11], [15].

A frequently used technique is the representation of the document characteristics by means of a vector space. In this technique, each term of the document becomes a dimensional feature. The value of each dimension may indicate the number of times a term appears in the text, or may indicate the valence weight of the term to be considered, as the amount of documents in which the term is used, for instance. However, this technique disregards relevant informant as, for example, the sequence in which the terms come up in the text, where they are used, and the proximity between them [16].

Graphs are important and effective mathematical constructions used to model relationships and structural information. Graphs are used in several kinds of problems, including sequencing, comprehension, traffic analysis, resource allocation, among others. As graphs retain more information than simple atomic vectors, they represent a valuable modeling resource which can be used in text representation [16].

Mining technique using graphs can discover words with greater occurrence within a text, and identify how close they are to one another. The graph obtained from the text mining procedure presents the most frequent words in its vertices. Thus, vertices associations in a graph indicate the proximity among words.

Different studies have been carried with the objective of helping analysis of discussion forums. Dringus and Ellis [17] conducted one of the first researches on the usage of data and text mining for analyzing discussion forums. Their study presented the integration of mining into forum query processes, and developed a series of tools to help visualizing temporal data, contribution rate and sequence of message interchanges.

Chen's work [18] presents some features that may be considered to measure high-quality topics in discussion forums. Based on these characteristics, a model to recognize them was proposed involving: extraction of time series signs with indicators of high-quality and low-quality topics, the diagnosis of these characteristics using Wavelet Packet Transform (WPT), and the recognition of high-quality topics using a back-propagation neural network.

Another research focusing on the analysis of discussion forums was carried by Lin [19]. His study uses text mining and proposes a system for classifying the different genres of contributions, such as: advertising, questioning, explanation, interpretation, conflicts, and statements. This system can be used to facilitate decoding processes when analyzing content in a forum.

Gerosa [20] has proposed a method of tracking messages by considering aspects related to discourse structure and message categorization. Structures allow the observation of hierarchical threads to check the level of engagement in the discussion; while message categorization helps identifying the aspects of each message genre.

Studies by Bassani and Behar [21] present a methodology to analyze discussion forums involving the mapping of interaction flow among participants.

Ravi and Kim [22] present an approach to automatically identify profiles of students' interactions in discussion forums. Act speech classifiers were developed to identify the roles of individual messages, such as: question, answer, elaboration, correction. The classifiers were used to search for messages that contain questions or answers. They used a set of rules for analysis the topics to find those who might have unanswered questions and need the attention of the teacher.

Kim et al. [23] present an intelligent agent implemented within a discussion forum to automatically provide answers to questions from students. The work shows how the topics of discussion were modeled using "speech acts". Each post was classified according to categories such as: question, answer, elaboration and correction. By classifying the contributions of the discussion, the authors were able to identify the roles of students and teachers in discussions. The authors developed a set of standards to analyze the interactions of learners in the discussions. Some of these patterns were used to find posts where students could have unanswered questions. The intelligent agent uses text mining techniques to extract words and their frequencies in the students' questions, course documents, and earlier discussions.

Another study to examine discussion forums with text mining techniques was presented by Lin [24]. The paper proposes a classification system of genres of textual contributions, such as: advertising, question, explanation, interpretation, conflict, contention, among others. This system can be used to assist the process of encoding the content analysis of a forum. Data were collected from a discussion forum to conduct the experiments. For researchers, the teacher can participate in a debate to contribute to the process of student learning. The major research question in the study was to validate the consistency of the results coded by an automatic classification of genres and the analysis performed by human evaluators. The article concludes that the waterfall model, embedded in the developed system, can facilitate the process of encoding the content analysis of forums.

Li and Huang [25] present a research to provide a more complete picture of the interactions among individuals in collaborative learning supported by computer. The authors propose a model for multidimensional analysis to investigate the interactions, based on techniques of content analysis, text mining and social network analysis. Content analysis was employed to investigate how students interact, discovering the patterns of process (the intention of speaking) within the conversation. The text mining was used to find the articles that appeared in the debates. The study also describes the design and implementation of a tool for intelligent analysis of content, called VINCA (Visual Intelligent Content Analyzer). An experiment was conducted with the tool to analyze a set of discussions, in order to unravel the interaction of students in terms of process patterns, space of topics, and social networking.

A discussion forum with advanced technological features was presented by Li [26]. The project used domain ontology and text mining techniques. At work, transcripts of forum discussions are automatically transformed into a structural modeling in three stages: recognition of the topic, identifying the type of each, and the semantic association between them. The first step brings together in a document the messages from the same set of discussions. The second step identifies six types of messages: questions, opinions, suggestions, recommendations, request and quote. The third step organizes the texts with semantic association. The forum offers three features to students: search for information relevant to their needs, thematic navigation through messages, and recommending other students who have interest in making communication and collaboration. An experimental study was conducted to show that the approach is effective to discover learning partners with the same interests and search for messages with navigation theme tab.

This study focused on the use of Sobek, software that uses text mining based on statistical methods [27]. This software enables the construction of a graph (using statistical text data) in which the vertices and edges contain information related to the absolute and relative number of term occurrences (nodes) and associations (arches) in a certain document. Graphs built from documents represent a network of the concepts studied in a document. Figures 1 and 2 shows, respectively, a text on cooperative learning entered in Sobek, and the graph generated by it. Each graph node presents a relevant text concept. The arches between nodes determine the association between the concepts. Macedo [28] uses this software to analyze student contributions in collective writing. The author focuses his work on teacher support by identifying, through graph visualization, textual problems involving, among others, cohesion and coherence.

This study used Sobek to analyze text contributions produced by students in discussion forums. The graph generated from each text contribution presents the relevant concepts and how they relate to each other.

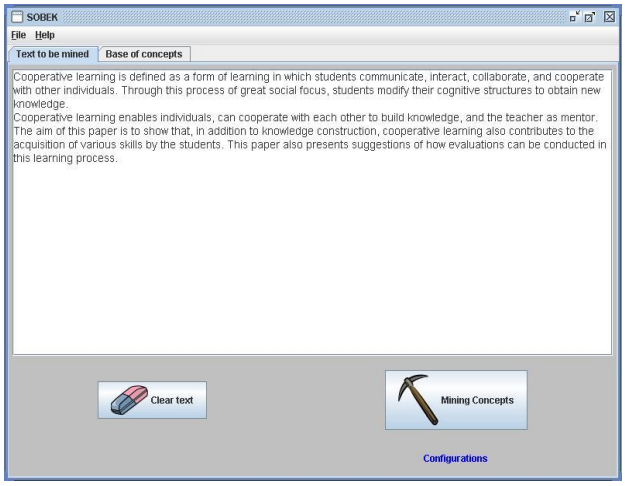

Figure 1. Text entered as example in Sobek

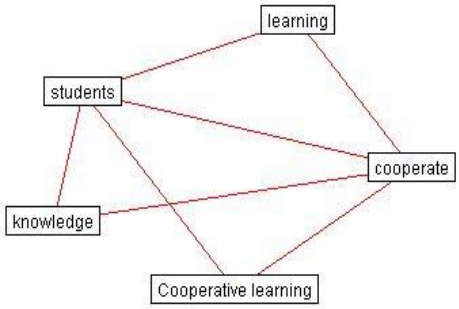

Figure 2. Graph generated from text in figure 1

III. Qualitative Analysis in Discussion Forums: a Pilot Study

According to Bakhtin [29, p. 91], in verbal communication, words should not be taken separately, neither as language units, nor as meanings only, but concrete and meaningful utterances. We are able not only to understand what words mean as language units but also respond to them, demonstrating our sympathy, agreement, disagreement, stimulus to action etc., a relation Bakhtin named "active responsive attitude". In Bakhtin's words: "Any concrete utterance is a link in the chain of speech communication of a particular sphere. The very boundaries of the utterance are determined by a change of speech subjects. Utterances are not indifferent to one another, and are not self-sufficient; they are aware of and mutually reflect one another... Every utterance must be regarded as primarily a response to preceding utterances of the given sphere (we understand the word 'response' here in the broadest sense). Each utterance refutes affirms, supplements, and relies upon the others, presupposes them to be known, and somehow takes them into account".

Thus, to understand discourse, it is of crucial importance to analyze context and not restrict the analysis to mere word identification. In other words, human communication is context-dependent and always organized as a response to someone or something.

Considering that text mining using graphs gather concepts and the associations between them within a certain text, one can conclude that the words representing such concepts are related and belong to a certain context. This idea justifies our choice to use text mining using graphs in this study, since the analysis does not consider words separately; it also focuses on how they relate to one another.

In this study, an analysis was made of the graphs generated in Sobek, in order to find out which textual contributions were relevant for the theme proposed.

To guide the activities for this article, we defined the Thematic Relevance Quotient (TRQ) of a text contribution, aiming at analyzing how relevant a text is within a certain discussion. The Thematic Relevance Quotient indicates the relevance degree of contributions in a discussion forum.

The Thematic Relevance Quotient can be calculated using the following formula:

$$TRQ = NC + NA \quad (1)$$

- NC: number of relevant concepts used in the text.
- NA: number of associations between relevant concepts used in the text.

To calculate NC, terms considered as semantically equivalent are also considered as relevant concepts. Prefixes, suffixes, and plural forms of significant concepts are converted into relevant concepts themselves. With TRQ, graphs with more important concepts for the theme, and more associations between concepts, have higher TRQs.

The methodology applied in this study involved:

a) The choice of a discussion forum.
b) The definition of important concepts within a discussion, as well as the definition of the associations between such concepts. These definitions were established by the teacher himself, not by the program.
c) The definition of possible terms that may be considered as semantically equivalent to the concepts involved.
d) The definition of the minimum value of the Thematic Relevance Quotient to be used in the analysis.
e) Gathering of text contributions produced by students in the forum.
f) Generation of a conceptual graph in Sobek for each contribution.
g) From generated graphs, calculation of the Thematic Relevance Quotient of contributions.
h) Organization of text contributions related to the graphs previously generated.
i) Analysis of the amount of relevant contributions made by each student in the forum based on the graphs.

To evaluate the methodology proposed in this study and analyze its results, five experiments were first carried out in different discussion forums.

The first experiment was made in one of the forums found in ROODA[1], and proposed for the subject "Special Topics Z1" for doctorate students in Computer Science and Education in the first semester of 2008. The forum theme was "Virtual Communities", and it presented 25 contributions.

According to the second step in the methodology above, to analyze this forum, the teacher defined and entered the concepts that would be considered as important. These concepts were: "communities", "virtual", "virtual communities", "learning", "virtual learning communities", "education", and "social relations". The relevant associations between these concepts were established in this way: the concept "virtual learning communities" is related to "communities", "virtual communities", "learning", and

[1] ROODA [2] is a virtual learning environment and one of the Distance Learning platforms used by Federal University of Rio Grande do Sul, RS, Brazil. Available at https://www.ead.ufrgs.br/rooda/

"education". The terms "VCs" e "LVCs" were defined as semantically equivalent to the phrases "virtual communities" and "learning virtual communities", respectively.

The minimum value 01 was defined for the Thematic Relevance Quotient (TRQ) to be considered in the analysis. This means that after the calculation of TRQ of each post, those with TRQ greater or equal to 01 were considered as relevant regarding the topic of discussion. The TRQ was calculated from the graph generated from each text contribution. The following message posted by student C serves as example of how the quotient was calculated.

Message: "I believe VLCs can be a valuable resource to education as they offer a tool for the collective construction of knowledge using the discussions made by members of the community."

The graph generated from this message is seen in Figure 3.

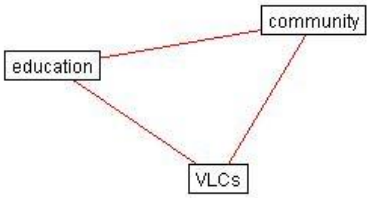

Figure 3. Graph generated from a forum message in experiment 1

For the graph in Figure 3, the TRQ may be calculated this way:

NC = 3, NA = 3

TRQ = NC + NA = 6

An analysis of the amount of relevant contributions made by each learner was made; its results are presented in Figure 4.

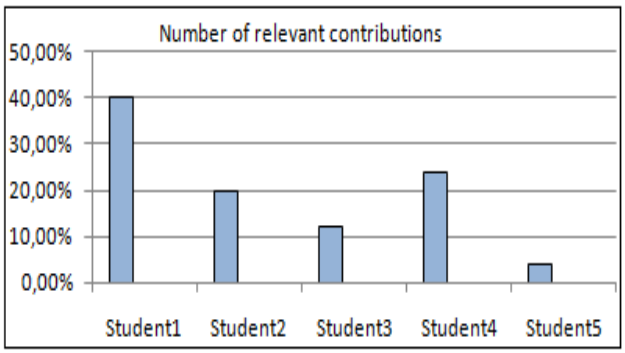

Figure 4. Number of relevant contributions in the first experiment

The graph in Figure 4 shows that the methodology allowed for the verification that most of the contributions were relevant. The second and third experiments were made in a discussion forum located in ROODA, for the subject "Integrative Seminar VII – B" for undergraduates in Education during the second semester of 2009. The theme of the forum was "Learning with others". The class was divided in two groups, each contributing in a different forum. The first forum had 27 participants who made 45 contributions; the second forum had 29 students with 67 contributions.

Following the second step of the methodology, the teacher typed concepts considered to be relevant. These concepts were: "to learn", "learning", "knowledge", "knowledge construction", "Piaget", "affect" "affection", "cognitive structures", "descentration", "affective relations", "affective experiences", "affective bonds", "cooperation", "collaboration". Relevant associations between the concepts were established as explained previously.

The minimum value 01 was defined for the TRQ considered for the analysis. The Thematic Relevance Quotient was calculated from the graph generated from individual contributions. An analysis of the amount of relevant contributions posted by each student in the first and second groups was carried, and the results are presented in Figures 5 and 6, respectively.

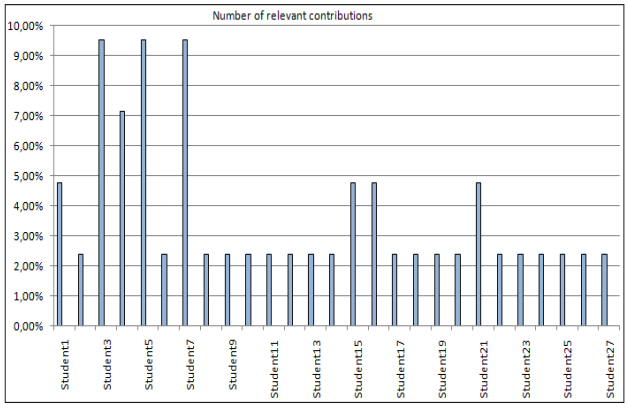

Figure 5. Number of relevant contributions from the first group

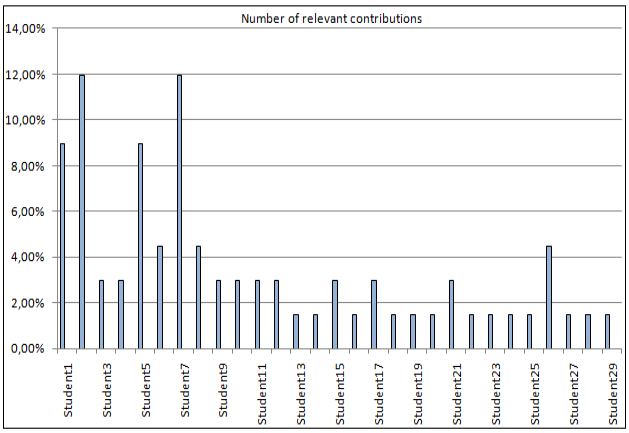

Figure 6. Number of relevant contributions in the second group

Observing the graphs in Figures 4, 5, and 6, one can see that the methodology supports the qualitative evaluation once it is able to verify checks whether the important concepts to be treated are effectively being used in the students' texts. In all graphs it is possible to see the amount of individual contributions that are relevant for the theme proposed in the forum.

Two more experiments were carried out in other discussion forums. Experiments E1 and E2 were carried out in forums available in the virtual learning environment ROODA, more specifically by the subject "Special Topics ZI" offered to doctoral students on the first semester of 2008. The discussion theme of the first forum (E1) was "Concept Maps", with 16 posts. The second forum (E2), with 13 posts, discussed "Information and Communication Technologies and Education". The minimum value 01 was defined for the TRQ considered for the analysis. Following the second step of the methodology, the teacher typed concepts relevant to the topic. An analysis of the amount of relevant contributions posted by

each student was done, and the results are presented in Figures 7 and 8, respectively.

In the study conducted for this paper, it was possible to observe that text mining using graphs is a viable alternative for the analysis of posts in discussion forums. With a graph generated from a message, one can assess whether it refers to the context, as well as its thematic relevance. In the analysis of a post, the more theme-related words are found, and the closer they are to one another, the greater the thematic relevance.

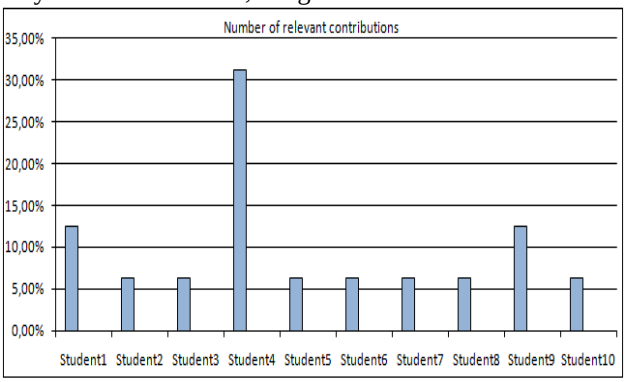

Figure 7. Number of relevant contributions in the experiment E1

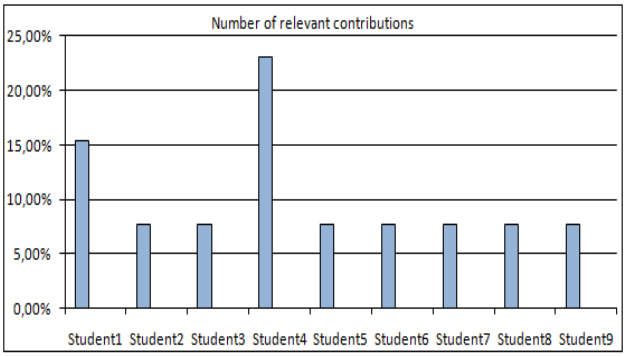

Figure 8. Number of relevant contributions in the experiment E2

We observed that the current version of Sobek used in the pilot experiments is accurate for both text mining and the generation of graphs representing text concepts. However, the other activities developed during the experiments had to be inputted by the teacher.

This study shows that to improve and allow Sobek to offer the necessary resources to make better qualitative analysis of text contributions in forums, it is necessary to refine it and create new functionalities such as:

a) Word stemming[2].
b) Reading of a group of semantically equivalent concepts.
c) Automatic reading of texts posted in discussion forums.
d) Definition of the minimum value of the Thematic Relevance Quotient to be considered in the analysis.
e) Execution of the Thematic Relevance Quotient for each graph generated from the texts produced by learners.
f) Analysis of the amount of relevant contributions posted by each learner in the forum using graphs.
g) Generation of a visual report for the teacher with information about the relevant contributions on the discussion theme. Such data should indicate the amount of both relevant and non-relevant contributions from each student.
h) Implementation of threads with relevant contributions based on generated graphs. With this ordination, the teacher will be able to visualize which messages are more relevant for the forum. Such information may also provide the teacher with an indicator to identify which students are dealing more effectively with concepts related to a certain discussion theme.

Another important feature that may help in qualitative analyses of text contributions and also be implemented as a new functionality in Sobek is the diagnosis of each text according to the following aspects:

a) Text consistency and cohesion.
b) Classification of text contributions as "questions', "answers", or "statements".

The qualitative analysis of messages helps the teacher to make a diagnosis on the students. In the experiments we observed that some students wrote several relevant messages. On the other hand, some students drafted only a few important messages.

Based on the results presented in this article, we verified that the methodology proposed here may help teachers to carry several activities related to forums, for instance:

- Analyzing which contributions need intervention.
- Visualizing students who placed few relevant contributions, and offer them support.
- Stimulating learners with many relevant contributions to interact with those with few messages.

Results obtained with the methodology presented here also allow teachers to direct their support and help to students who post few relevant messages in the forum. Teachers can also motivate or mediate collaborative learning processes by stimulating interaction and communication among learners with more relevant contributions and those with few or less relevant postages.

IV. Conclusion

The initial experiments made for this paper show that text mining using graph representation can be applied in the creation of relevance indicators for forum messages. With such analysis at hand, teachers have a tool that may help them follow relevant contributions made by learners in discussion forums.

It is important to observe that qualitative analysis of text contributions by students in discussion forums is an alternative resource to support teaching practices. Our analysis gives teachers a qualitative perspective of text samples produced by students involving concepts related to the discussion theme. Nevertheless, one can notice that some relevant texts may not be detected. Based on the criteria detailed in the study, these texts may not be considered in the analysis, which is justified by the fact that the study did not carry a more in depth linguistic text analysis.

The qualitative analysis presented in the initial experiments used some quantitative indicators drawn from the textual contributions produced by the students. The indicators were used to compose the calculation formula of Thematic Relevance Quotient, explained in section 3.

The use of software assists and accelerates the qualitative analysis of messages in a forum. However, it is important to

[2] Stemming is the reduction of a set of words to the same stem by removing prefixes and suffixes, i.e., keeping only the root form [30].

note that the time required for the software to run the process depends on the amount of posts.

Currently, a specific text mining tool for discussion forums is being developed by the authors. This software uses the resources of Sobek, and implements other new features. The software is being adapted to the virtual learning environment ROODA, and may be used to assist teachers in the analysis of posts made by students. In this software, the formula of Thematic Relevance Quotient was also improved. A new term was added to compute the weight of each vertex of the graph to be analyzed. The thematic relevance is being calculated, considering as reference, the concepts indicated by the teacher, or the concepts extracted from a text reference. Besides, the new formula tries to consider concepts discussed in students' posts which were not given by the teacher or extracted in a reference text.

References

[1] R. J. Dornelles. "A utilização de tecnologias de Internet na educação a distância: o caso de uma disciplina de graduação da Escola de Administração da Universidade Federal do Rio Grande do Sul". MSc Dissertation, UFRGS, Porto Alegre, Brazil, 2001.

[2] P. A. Behar, et al. "ROODA: desenvolvimento, implementação e validação de um AVA para UFRGS". In *Proceedings of the XII Taller Internacional de Software Educativo (TISE)*, pp. 321-338, v. 1, 2007.

[3] R. M. Palloff, K. Pratt. *O aluno virtual: um guia para trabalhar com estudantes on-line*, Artmed, Porto Alegre, 2004.

[4] M. Mazzolini. "When to jump in: The role of the instructor in online discussion forums", *Computers & Education*, 49(2), pp. 193-213, 2007.

[5] R. Feldman, J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, 2007.

[6] M. Hearst. "Untangling Text Mining". In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, pp. 3-10, 1999.

[7] A. A. Puretskiy, G. L. Shutt, M. W. Berry. "Survey of text visualization techniques", in *Text mining: applications and theory*, John Wiley & Sons Ltd, 2010.

[8] A. Tan. "Text Mining: The State of the Art and the Challenges". In *Proceedings of the Workshop on Knowledge Discovery from Advanced Databases (PKDAD'99)*, pp. 71-76, 1999.

[9] W. Fan, et al. "Tapping the power of text mining". *Communications of ACM*, 49(9), pp. 76-82, 2006.

[10] R. J. Mooney, U. Y. Nahm. "Text Mining with Information Extraction". In *Proceedings of International Midp Colloquium*, pp. 141-160, 2003.

[11] V. Gupta, G. S. Lehal. "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, 1(1), pp. 60-76, 2009.

[12] G. S. Mahalakshmi, S. Sendhilkumar. "Automatic Reference Tracking", in *Handbook of research on text and web mining technologies*, Hershey: Information Science Reference, 2009.

[13] G. Ramakrishnan, P. Bhattacharyya. "Question Answering Using Word Associations", in *Handbook of research on text and web mining technologies*, Hershey: Information Science Reference, 2009.

[14] M. Castellanos. "HotMiner: Discovering Hot Topics from Dirty Text", in *Survey of text mining: clustering, classification, and retrieval*, Springer-Verlag, New York, 2004.

[15] M. Kobayashi, M. Aono. "Vector Space Models for Search and Cluster Mining", in *Survey of text mining: clustering, classification, and retrieval*, Springer-Verlag, New York, 2004.

[16] A. Schenker. "Graph-Theoretic Techniques for Web Content Mining". PhD Thesis, University of South Florida, Florida, USA, 2003.

[17] L. P. Dringus, T. Ellis. "Using data mining as a strategy for assessing asynchronous discussion forums", *Computers & Education*, 45(1), pp. 141-160, 2005.

[18] Y. Chen, et al. "A Wavelet-Based Model to Recognize High-Quality Topics on Web Forum". In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, v.1, pp. 343-351, 2008.

[19] F. Lin, et al. "Discovering genres of online discussion threads via text mining", *Computers & Education*, 52(2), pp. 481-495, 2009.

[20] M. A. Gerosa, et al. "Coordenação de Fóruns Educacionais: Encadeamento e Categorização de Mensagens". In *Proceedings of the XIV Simpósio Brasileiro de Informática na Educação*, pp. 45-54, 2003.

[21] P. S. Bassani, P. A. Behar. "Análise das interações em ambientes virtuais de aprendizagem: uma possibilidade para avaliação da aprendizagem em EAD", *Revista Novas Tecnologias na Educação (RENOTE)*, 1(4), 2006.

[22] S. Ravi, J. Kim. "Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers". In *Proceedings of the 13th International Conference on Artificial Intelligence in Education (AI-ED 2007)*, pp. 357-364, 2007.

[23] J. Kim, et al. "An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions". In *Proceedings of the AAAI Spring Symposium On Interaction Challenges For Intelligent Assistants*, 2007.

[24] F. Lin, L. Hsieh, F. Chuang. "Discovering genres of online discussion threads via text mining", *Computers & Education*, 52(2), pp. 481-495, 2009.

[25] Y. Li, R. Huang. "Analyzing Peer Interactions in Computer-Supported Collaborative Learning: Model, Method and Tool", *Lecture Notes in Computer Science (LNCS)*, 5169, pp. 125-136, 2008.

[26] Y. Li, M. Dong, R. Huang. "Semantic Organization of Online Discussion Transcripts for Active Collaborative Learning". In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 756-760, 2008.

[27] A. Lorenzatti. "SOBEK: uma Ferramenta de Mineração de Textos". UCS, Caxias do Sul, Brazil, 2007.

[28] A. L. Macedo. "Um recurso de apoio para acompanhamento e avaliação do estudante na construção de um texto coletivo". UFRGS, Porto Alegre, Brazil, 2008.

[29] M. Bakhtin. *Estética da criação verbal*, Martins Fontes, São Paulo, 1986.

[30] K. Spark-Jones, P. Willet. *Readings in Information Retrieval*, Morgan Kaufmann Publishers, San Francisco, 1997.

Author Biographies

Breno Fabrício Terra Azevedo has MS in Computer Science from Universidade Federal do Espirito Santo. He is Ph.D. student in Information Technology in Education at the Federal University of Rio Grande do Sul. He is currently working as Professor in Federal Institute of Education, Science and Technology Fluminense. His current research interests include the area of Text Mining, and Discussion Forums.

Patricia Alejandra Behar has PhD in Computer Science. Master's and Ph.D. of Computer Science at the Federal University of Rio Grande do Sul. Currently is professor at the Education School and the Post Graduation Programs in Education and Computer Science in Education at the Federal University of Rio Grande do Sul. Coordinator of the Digital Technology Nucleous in Education (NUTED). Has experience on Distance Education, learning virtual environments, learners and lifelong learning, teaching and role of teachers in e-learning.
Professor and researcher from the Federal University of Rio Grande do Sul, Brazil
Head of the Digital Technology Nucleous in Education – NUTED (http://www.nuted.edu.ufrgs.br)
Curriculum: http://lattes.cnpq.br/7661737809414762

Eliseo Berni Reategui has a PhD degree in Computer Science from the University of London, England, and MS in Computer Science from Universidade Federal do Rio Grande do Sul, Brazil. After finishing his PhD and working in the industry for 5 years, Dr Reategui held a lecturer position at the University of Caxias do Sul, Brazil, for a few more years. Nowadays, he works as a lecturer and researcher at the Federal University of Rio Grande do Sul. His research interests are related to the use of computers in education, involving areas such as artificial intelligence and human-computer interaction.
Curriculum: http://lattes.cnpq.br/9140136724972740