Automatic analysis of messages in discussion forums

Breno Fabrício Terra Azevedo
Department of Information Technology
IFF
Campos dos Goytacazes-RJ, Brazil
bterra@iff.edu.br

Patricia Alejandra Behar, Eliseo Berni Reategui
Graduate Program in Information Technology in Education
UFRGS
Porto Alegre-RS, Brazil
pbehar@terra.com.br, eliseoreategui@gmail.com

Abstract—This paper presents a study aimed at investigating on whether the text mining technique using graphs can be used to analyze the relevance of the messages in online forums. Experiments were carried out with a program that calculates the thematic relevance of text contributions. In addition, the paper presents results found in the automatic analysis obtained by the application of the software.

Keywords-discussion forums; text mining; automatic analysis

I. INTRODUCTION

Asynchronous discussion should be stimulated by teachers as it is the best way to establish and keep interaction with students. Student participation in discussions provides opportunity to think about contents seen in the course; thus discussion forums are favorable spaces for this kind of activity. Students’ answers in these debates indicate that they are possibly reflecting over the contents [1].

Involvement in discussion forums is an important activity in online education, as it allows the teacher to examine information about learners. However, if the teacher has a great number of students, he/she will need a significant amount of time to analyze all discussions.

This paper presents a study that aims at investigating if the text mining technique using graphs can be applied to analyze messages posted by students in discussion forums. The software MineraFórum, which calculates the thematic relevance of posts, was used in the experiments.

Analysis of thematic relevance allows teachers to identify learners that are (as well as those who are not) discussing concepts proposed in the debate. By doing this, teachers may devote more time to find out why some learners are not debating the discussion topic. In case such identification is successful, the teacher may offer help.

Next section presents a brief introduction to text mining. Section 3 presents works that have applied this technique to discussion forums. Section 4 and 5 describes the methodology and the experiments made for this study. Section 6 presents considerations on the research, followed by the bibliographical references used in this paper.

II. TEXT MINING


knowledge intensive process in which the user interacts with a great amount of documents by using tools to analyze them. The goal is to extract useful elements from a collection of documents. Information is identified in interesting patterns found in unstructured text data.

Text mining systems are based on preprocessing routines, pattern discovery algorithms, and elements for presenting results. The steps that make up the architecture of such systems are: preprocessing operations, generation of processed documents, mining, and presentation of results. The system user interacts in the preprocessing stage, with the mining nucleus, and in the presentation of results.

Preprocessing operations are based on the identification and extraction of distinguishing features of documents written in natural language. These operations are responsible for turning unstructured data, stored in collections of documents, into a structure expressed in an intermediary model [2], [3].

Operations which are in the mining nucleus, also named knowledge distillation processes, represent the core of a text mining system, and include: pattern discovery, trend analysis, and incremental algorithms for knowledge discovery. The most frequent patterns used in knowledge discovery are distributions and proportions, sets of frequent results, and associations. Advanced systems for text mining, or domain-oriented ones, can also improve the quality of operations by consulting knowledge bases [3].

Elements that are part of the presentation of results make up the system interface, with navigation functionalities, and access to the language used in the consulting process [2], [3], [4].

Some techniques used in text mining include: information extraction, topic tracking, summary production, text categorization, text clustering, conceptual links, information visualization, analysis of questions and answers [5], [6].

Information extraction identifies main sentences, and relations within a given text. This is achieved by searching predefined sequences in the text, a process known as pattern matching. This procedure infers relationships between all identified people, places and data, in order to provide meaningful information to the user [5], [7].

Topic tracking store user profiles and, based on documents seen by the user, can predict others that might be of his/her interest. This process can be applied to discover references in a research area [6], [8].

978-1-4577-1746-8/11/$26.00 ©2011 IEEE
14th International Conference on Interactive Collaborative Learning (ICL2011) ̶ 11th International Conference Virtual University (vu'11)
21–23 September 2011, Piešťany, Slovakia
Page 76
Conceptual links connect related documents by identifying commonly shared concepts. Such technique helps users to find information that might not be obtained by traditional search methods [5], [6].

Information visualization presents great amounts of documents in a visual hierarchy, or in a map; thus providing navigation alternatives, and search activities [2], [3], [4].

Analysis of questions and answers originates from the natural language processing area, a field which deals with finding the best answer for a given question [6], [9].

Production of summaries (summarization) is useful when the aim is finding out whether or not a long document fits the needs of the user, so that he/she can evaluate, for instance, if it is worthy reading for more information. The goal of summarization is to reduce the size and level of detail in a document, by keeping its main ideas and general meaning. The greatest difficulty for the software is to do semantic analysis and interpretation of meanings. One of the most used strategies in summarization is the extraction of important sentences from a text, by considering the statistical weight of sentences. Some additional heuristics, such as information on positioning, are also used in this process. Summarization methods can be classified in two large groups: (i) superficial approaches, which are restricted to the syntactic level of representation; (ii) deep approaches, which involve some level of semantic representation of the original text, and use some level of linguistic processing [6], [10].

Categorization (classification) involves the identification of the main themes in a document, by placing it in a predefined collection of topics [6]. The categorization process includes counting words in a given document and, based on this counting it can identify the main topics in the document. This technique frequently uses a method of document classification to sequence those with more content about a specific theme.

Text grouping (clustering) is a technique used to group similar documents. It differs from categorization, because instead of using predefined topics, documents are clustered in real time. Another advantage of clustering is that documents can be displayed in various subtopics, thus ensuring that a useful document is not omitted from search results. A basic clustering algorithm creates a topic vector for each text, and calculates the weights in order to identify in which group the document should be included [6], [11].

In text mining, a common technique used to represent a document is the identification of its features by means of a vectorial space model. In such process, each term of the document becomes a dimensional feature. The value of each one may indicate the number of times a term occurs in the text; or it may indicate the weight of the term under consideration, such as the amount of documents in which the term occurs, for instance. However, this procedure discards important information such as the order in which terms come up, where they occur, and the proximity between them, for example [12].

Graphs are important and effective mathematical constructions for modeling relationships and structural information. Graphs are used in several kinds of problems, including sequencing, compression, traffic analysis, resource location, among others. As graphs retain more information than vectors of simple atomic features, they represent an interesting modeling alternative, and can be used in text representation [12].

Text mining with graphs may discover words with greater number of occurrences in a text, and identify if they are close to one another. Graphs obtained in mining show the most frequent words in their nodes. Associations among the nodes indicate proximity between words.

By considering the graph generated by a post in a forum, one can analyze if it refers to the context, and evaluate its thematic relevance. The investigation of a message may show that the more topic-related words are used, and the closer they are to one another, the greater will be the thematic relevance.

III. ANALYSIS OF DISCUSSION FORUMS WITH TEXT MINING

Dringus and Ellis [13] have conducted a study to evaluate the use of data and text mining to analyze discussion forums. In their paper, the authors discuss the problems that contribute to render difficult to the evaluation of those tools. They suggest participation indicators that the teacher can extract from discussions and use as a resource to evaluate development and performance of students. The authors recommend that text mining be used to find the distribution and frequency of contributions made by the student throughout the duration of the forum. Text mining can also be used to identify students' initiatives or answers, as well as discovering how long it took the student to answer the initial post. This study recommends that text mining can be applied to evaluate if contributions are related to the discussion topic, and to identify if students are sharing resources, such as Web references or literature quoting. According to the authors, text and data mining techniques can be part of the solution in helping teachers analyze discussion forums, and obtaining student feedback.

Kim et al. [14] present an intelligent agent implemented in a discussion forum to provide automatic answers to questions posted by students. Their work show how discussion topics were modeled with the application of “speech acts”. Each post was classified according to “speech acts” categories, such as: question, answer, elaboration, and correction. The authors developed patterns to analyze discussions among learners. Some of these patterns were used to discover messages in which students could have non-answered questions. The intelligent agent uses text mining techniques to extract words and their frequency in the students’ questions, in the course documents, and in previous discussions. The similarity measure “cosine” was used to find out which of the documents and/or previous discussions had similarities with the text used in the questions posted by the student.

Li and Huang [15] present a study that aimed at providing a more comprehensive picture of interactions in computer-supported collaborative learning. The authors propose a model of multidimensional analysis to investigate interactions based on techniques applied in content analysis, text mining, and social networking. Content analysis is used to investigate how students interact, thus discovering possible process patterns (the discourse intention) in the conversation. Text mining is used to
discover topics that come up in the debates. Messages in a group of discussions are clustered in a document represented by a vector of determined terms. The cosine method is used to calculate the similarity between the document vector and the theme vector as determined by the teacher. The study also describes the project and implementation of a tool for intelligent content analysis named VINCA (Visual Intelligent Content Analyzer). The tool was used in an experiment to analyze a number of discussions, and aiming at understanding students’ interactions in terms of process patterns, topic space, and social networking.

Another study conducted to analyze discussion threads is presented by Lin et al. [16]. Their work uses text mining, and proposes a system for classifying textual contribution genres, such as: announcements, questions, explanations, interpretations, conflicts, statements, among others. This system can be used to facilitate coding in content analysis of a forum. For the experiment, they collected data from a discussion forum in a Moodle environment. Their main research objective was the validation of coherence in the results coded by an automatic system of genre classification, and those found by human judgment. The authors’ conclusion is that the cascade model, embedded in the system, may facilitate coding in content analysis of forums.

IV. METHODOLOGY

To investigate discussion forum, this study defined the following methodology:

a) Choice of a technique to analyze content of posts.

b) Selection of the software to carry the automatic analysis of the messages.

c) Definition of discussion forums containing the corpus of analysis.

d) Application of the software in text contributions made in the selected forums.

e) Analysis of results presented by the software.

Regarding the first item, several techniques can be used in the analysis of message content. Some can be found in [13], [14], [15], [16], [17]. This study chose text mining using graphs due to the promising results presented by Azevedo et al. [18].

Analysis of forum messages can be made manually. However, a computer program can be used to aid and speed up this process. This study selected MineráFórum [19] as it presents useful features to carry the experiment. The software calculates the thematic relevance of posts in a discussion forum. Fig. 1 presents its main interface.

Some resources offered by MineráFórum are listed next:

- It allows the user to inform a reference text about the debate topic. Fig. 1 presents the interface in which the user can indicate the text.

- It allows the user to type concepts considered relevant to the topic of discussion and establish associations among them. Fig. 2 presents the interface for this resource.

- It uses a synonyms dictionary in the mining process. This dictionary can be informed by the user, or one can use what was previously determined in the software.

- It allows the user to inform words with semantic equivalence.

- It uses text mining with graphs. The process is carried with the reference text indicated by the user, and with each message posted in the forum.

- It calculates the thematic relevance of each post. The minimum relevance value to be considered can be informed by the user.

- It displays a report with information on the analysis of posts. This information includes: the total number of posts by each author, the amount of individual relevant and non-relevant contributions, and important concepts used in the contributions considered relevant in the discussion.
To calculate the thematic relevance of messages, MineraFórum carries out the following actions:

a) Building a graph using the reference text for the debate (indicated by the user). In this process, stopwords (words that can be removed in the mining process, such as: adverbs, articles, and prepositions) are removed, and words with greater occurrence in the text are identified. These most frequent words represent the most relevant concepts in the text that was mined, and correspond to the vertices in the graph. The edges between the concepts are generated according to the proximity among words in the text. In case the user decides to type important concepts related to the discussion, instead of indicating the reference text, MineraFórum builds a graph using these concepts.

b) Creation of graph from each message posted in the forum.

c) To calculate the thematic relevance of a message, MineraFórum must analyze the correspondence between the graph generated from the reference text and the graph built from the message. The first stage of analysis comprises the identification of which vertices in the first graph are equivalent to others in the second graph. In the context of this study, two vertices are equivalent if they present similar content, that is, if they have the same words, or if the words can be reduced to the same stem, or if they have synonyms, or if they present words with semantic equivalent. The second stage of analysis involves the use of a formula that takes into consideration three aspects of the equivalent vertices: the amount of these vertices in the two graphs, the distance between them within the same graph, and their weight in its corresponding graph. The formula allows for the final sum of the three mentioned factors. The value of such result corresponds to the thematic relevance of the message.

Aiming at verifying the methodology proposed in this paper, five discussion forums were selected. The text contributions in these forums were analyzed by MineraFórum.

Next section describes the selected forums. The result of the thematic analysis of posts in each forum is shown in the bar graphics. Each bar graphic presents the amount of individual contributions considered relevant in the debate.

V. EXPERIMENTS

In order to analyze the thematic relevance of students’ posts, five experiments were carried using different discussion forums.

For all experiments, the minimum value 01 was defined for the thematic relevance (TR) in the texts. This means that after calculating the thematic relevance of each message, those in which TR was greater or equal to 01 were considered relevant for the discussion theme.

The first three experiments were carried in discussion forums found in the ROODA environment, and offered in the discipline “Special Topics Z1” for students in the doctoral...
program Computer Science and Education, during the first semester of 2008. The discussion theme in the first forum was “Concept Maps”, with 16 text contributions. The second forum had “ICTs and Education” as main theme, and 13 text contributions. The discussion topic of the third forum was “Virtual Communities” with 26 posts.

The fourth and fifth experiments were also carried in the ROODA environment, for the subject “Integrative Seminar VII – B”, for the graduate students in Pedagogy. The debate theme was “Learning with Others”. The class was divided in two groups, each one contributing in separate forums. The first forum had 27 participants who posted 42 messages. The second forum had 28 students who contributed with 67 posts. In all experiments, the teacher inserted a reference text in MineraFórum about the discussion theme.

Results of the experiments are presented in Fig. 4 to 8. Graphics indicate the value of the thematic relevance of the messages posted by each student.

Fig. 4 shows that Student 7’s message presents high thematic relevance. On the other hand, Student 4 contributed with five messages, but his final post has low relevance value.

Results in Fig. 5 show that messages written by Student 1, Student 5, and Student 7 present the highest thematic relevances. All messages posted by students were relevant for the discussion theme.

As to Fig. 6, Student 1 had a significant participation in the forum, with several relevant messages. In contrast, Student 5 wrote only one post.

Fig. 7 shows that Student 3, Student 4, Student 5, and Student 7 present higher participation, and relevant messages. However, Student 9, Student 12, Student 17, and Student 26 posted only one message with relevance value ranging from 3.0 to 5.0.

In Fig. 8, results show that Student 1, Student 2, Student 5 and Student 7 present the highest number of participation and of relevant messages; whereas Student 13, Student 23, Student 25, and Student 28 posted one message only, with relevance value varying from 3.0 to 4.0.
Results presented above confirm that the methodology proposed in this paper allowed for the analysis of thematic relevance of messages in discussion forums. It was possible to observe which students posted more text contributions related to the topic of discussion, and who wrote few relevant messages. In addition, results show which students contributed with messages containing great relevance in relation to the theme being debated.

VI. CONCLUSION

In the study conducted in this paper, it was possible to observe that 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.

The major contribution of this paper was the presentation of a study aimed at analyzing messages posted in discussion forums. The study shows that the methodology proposed in section three can be applied to qualitative analysis of forums. Experiments were made to observe results obtained by applying this methodology.

The software used in the analysis of text contributions made in discussion forums is capable of calculating the thematic relevance of messages. With this value, it was possible to identify whether or not the posts referred to the discussion topic.

It is important to point out that the use of a computer program may help and speed up the thematic analysis of forum messages. Nevertheless, it must be said that the length of time the software needs to execute the process depends on the amount of posts. In the experiments discussed here, it took MineraFórum only a few seconds to analyze the forums, each one with 16, 13, 26, 42, 67 posts, respectively.

By making use of a program able to check if messages posted by students refer to the desired context, the teacher can direct his support to learners. With the analysis of the thematic relevance of texts, the teacher has information that may help him/her in the observation of contributions made by students.

Thematic analysis of forum messages can also help teachers to make diagnostics about students. In the experiments, it was observed that some students wrote posts with mean thematic relevance. On the other hand, some made only one or two posts with high relevance.

From the results presented in this study, teachers can make their analysis, and verify which text contributions need intervention. It is possible to identify which learners contribute less and, then, offer them greater support. Using this information, teachers can also motivate students who post more relevant contributions to interact with those who write few messages.

REFERENCES


