Using Text-Mining to Support the Evaluation of Texts Produced Collaboratively

Alexandra Lorandi Macedo[1], Eliseo Reategui[1], Alexandre Lorenzatti[2], and Patricia Behar[1]

[1] PPGIE, UFRGS, Av. Paulo Gama, 110 - Porto Alegre/RS - CEP: 90040-060, Brazil
{alorandimacedo,eliseoreategui}@gmail.com, pbehar@terra.com.br
[2] PPGCC, UFRGS, Av. Bento Gonçalves, 9500 – 91501-970, Porto Alegre, Brazil
alorenza@gmail.com

Abstract. This paper presents a collaborative writing system which has been conceived to be used by teachers as a collaborative learning tool in distance learning courses. Besides enabling students to communicate with each other and elaborate a text in a collaborative way, the system has an embedded text mining tool to enable teachers to extract graphs from student's writings. The graphs give teachers a concise view of the students' works by showing important concepts that appear in the texts. An extension course was organized in order to provide an initial validation for the collaborative writing tool. The experiments carried out during the course demonstrated the potential of text mining for the analysis of students' work. The experiments carried out as well as their results are presented here, followed by conclusions and suggestions for future work.

Keywords: Collaborative writing, text mining, distance learning.

1 Introduction

In the last few years the number of collaborative writing tools has proliferated, especially with all the services and interactive features made possible by the Web 2.0. At the same time, educators have realized the potential of such tools in learning activities. Among other advantages, the use of collaborative writing tools may increase group awareness, making group members more informed about other members' writings and more conscious about being engaged in a cooperative team work [1].

From a teacher's perspective, the possibility of getting students to work collaboratively through the use of computational tools is both attractive, from a learning perspective, and convenient: each student's progress may be monitored through historical records without too much difficulty.

However, although computational tools may store the steps taken by each student in the creation of a document produced collaboratively, the actual monitoring of each student's work is a very demanding task [2].

This paper presents ETC, a collaborative writing system which has an embedded text mining tool to enable teachers to extract graphs from student's writings. The graphs give teachers a concise view of the students' works by showing concepts and

A. Tatnall and A. Jones (Eds.): WCCE 2009, IFIP AICT 302, pp. 368–377, 2009.

relationships that seem to be relevant. The tool has been evaluated in an extension course in which 9 students participated. The results achieved are presented and discussed in the last sections of the paper. The next section gives a brief overview of the collaborative writing tool ETC; section 3 presents the embedded text mining tool called Sobek, which is capable of extracting graphs from students' writings; section 4 presents the experiment carried out with the 9 students who used the collaborative writing tool during a whole month; section 5 discusses results, presents conclusions and directions for future work.

2 ETC: A Web-Based System for Collaborative Writing

The appeal of collaborative writing in learning activities is particularly interesting as the act of producing a text in a collaborative way can motivate writers to work in a recurring process of critique and re-elaboration of their work in the pursuit of better results [3]. The web-based tool ETC, designed and developed at the NUTED center, Federal University of Rio Grande do Sul, has been conceived specifically to be used by teachers as a collaborative learning tool in distance learning courses. ETC's main features are listed below:

- administration control to allow only registered users to access each text;
- simultaneous access to enable several users edit the same text at the same time;
- possibility to "lock" parts of a text in order to prevent other users from editing the "locked" portions;
- text mining feature enabling graphs to be extracted from students' writings;
- conventional text formatting functions.

Most of these features can be found in the majority of collaborative writing systems, such as the historical tracking of text changes, or formatting functions. But some of them are not so common, such as the possibility to block a portion of a text in order to prevent other users to change it while one is working on it. Such a feature is interesting specially when a text is being edited by several hands concurrently, and a user needs to work on a given part of the text without the intromission of others.

But the truly innovative feature of ETC is its capacity to extract graphs from the users' writings, giving teachers a brief view of the students' work. The next section presents Sobek, the text mining tool embedded in ETC, detailing its main features as well as its mining algorithm.

3 The Text Mining Tool: Sobek

Text mining can be defined as a knowledge-intensive process in which a user employs different tools in order to look for useful information from data sources through the identification and exploration of interesting patters [4]. While in the area of data mining these patterns are sought in formalized database records, in text mining the data sources are unstructured document collections.

Our text mining tool has been called Sobek, which comes from the Egyptian mythology where it represents a god of discernment and patience. Although Sobek can be used for the analysis of documents in different formats ("txt", "pdf" and

"doc"), its development has been inspired by an actual need of university professors who work with distant education and who have to review a large number of texts produced by students. By presenting a concise view of a text, Sobek intends to provide clues about problems, or about the quality of a text, that can be recognized promptly. Sobek can be used in different ways. The analysis of plain text is Sobek's simplest operation. The text to be analysed can be copied and pasted in the tool or it can be loaded from a file. If the text is in a PDF or DOC format, it is automatically converted to the text format. The main goal of the text analysis is to extract concepts from the text and to visualize the graphical representation of those concepts and their relationships in a graph. Figure 1 shows a graph extracted from the five initial paragraphs of a Wikipedia text about global warming [5]. In the graph one may find important concepts that were extracted from the text, such as *global, warming, climate, change, surface temperature and greenhouse.*

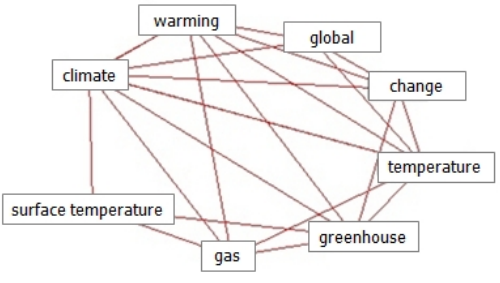

Fig. 1. Graph elicited from part of a Wikipedia text about global warming

Although the graphs cannot be used to reconstruct the original text, they may give a good notion of the main ideas and concepts considered. The use of graphs to represent relationships between objects and/or concepts can be justified by the fact that they are a form of abstraction that is widely applied and is easy to understand [6]. The next subsection details the text mining process.

3.1 The Text Mining Process

In our project, a particular text mining technique based on statistical analysis has been used to generate a graphical representation of the concepts extracted from texts. The information extracted from the texts is represented in a modified graph, based on a graph model proposed in Sckenker's PhD thesis [7], whose goal was to extract information from internet pages. He also proposed six different graph models to represent the information extracted from texts. One of these models, the *n-simple distance model*, was modified and used in our work to represent the texts. The *n-simple distance graphs* are based on the idea that each statistically relevant word of the text is going to be connected to the N subsequent relevant words. In Schenker's model, each node of the graph contains one single word. In the modified version created here, a node can have more than one word, so that it can express a more complex idea. For instance, figure1 showed a graph mined from a global warming text. Notice that there were nodes with one term (*e.g. climate, global, change,...*) and a node with two terms (*surface temperature*).

While other text mining approaches rely on the analysis of relevant morpho-syntactic patterns (such as Noun Noun, Noun Preposition Noun, Adjective Noun, etc.) in order to generate compound terms for the mining process [8], here we used a simpler method which was based on the frequency with which these compound terms appeared in the text. Our method relies on a parameter N to extract the compound concepts with more than one word. According to this parameter we create a combination of the current word with the N subsequent words. What we try to do is to create a wide combination of words to find the most frequent group of words that appear in the text. For instance, considering N=3, the analysis of the sequence of terms AA BB CC DD EE FF GG HH would lead us to the following combinations AA, AA BB, AA BB CC, BB, BB CC, BB CC DD, and so on. In order to avoid sequences starting with prepositions or articles, specific filters are used. After identifying the most frequent combinations of words, which we will call concepts, the mining process selects the most relevant ones based on their frequency in the text.

The next step is to compute the similarity between concepts. Consider two concepts a = AA DD BB and b = BB CC DD EE FF AA. The similarity coefficient is calculated with the scale product, in the same fashion used in Vector Space Models [9]. The similarity coefficient, represented by SC, computes the quantity of words present in both concepts represented by QB, and the quantity of words of the largest concept represented by BC. Therefore we have:

$$SC=QB/BC$$

In the example above SC=0,5 as the concepts have three words in common, words AA, BB and DD. Concept b, being the biggest, has six terms. After computing the value of SC, the relevancy coefficient RC is computed for each concept. The size of the concept (number of words) (NW) and the absolute frequency (AF) are introduced in the computation process. To calculate the RC for each concept, the following formula is employed:

$$RC=SC*NW+AF$$

The concept with the biggest value for RC is kept on the base, and at the end of the process, it is included in the graph. In the example above, let us consider that concept ***a*** has NW=3 and AF=3, and concept ***b*** has NW=6 and AF=2. We can conclude that concept ***b*** is going to remain in the base to be part of the graph, even if its AF value is smaller then that of concept v ***a*** In summary, when Sobek receives a text for processing, it breaks it down word by word and after that, it tries to single out the concepts that will compose the graph. After completing the analysis and before building the graph, a list of stopwords is used to remove articles, prepositions and terms with no meaning from the base of concepts.

4 Experimentation

An initial experiment was carried out in order to evaluate ETC and its text mining tool Sobek, focusing on their capacity to provide clues about the texts written collaboratively by students. An extension course about collaborative writing was organized by NUTED/UFRGS, as part of the research on the ETC project. Nine

students participated in the course during a whole month. After learning the importance and the main features of collaborative writing tools, the students learned how to use ETC to produce collaborative texts themselves. However, they did not have access to the tool's text mining features. At the end of this period, the students were asked to produce a text on the topic "authorship". The texts produced by the students were analysed by an experienced teacher in collaborative writing, using the text mining feature, in an attempt to verify whether Sobek could really provide interesting clues about the texts written. The graphs below were extracted from the final texts produced by the students.

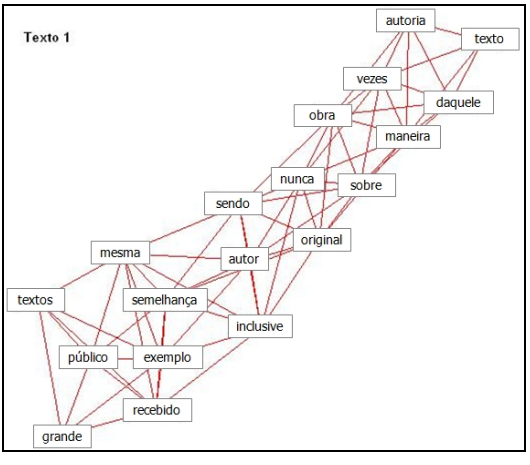

Fig. 2. Graph extracted from students' text – group 1

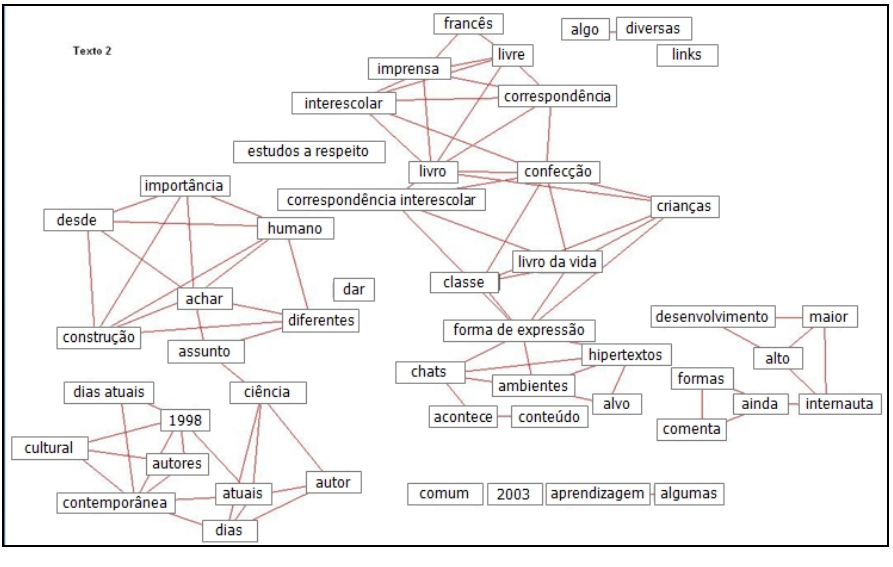

Fig. 3. Graph extracted from students' text – group 2

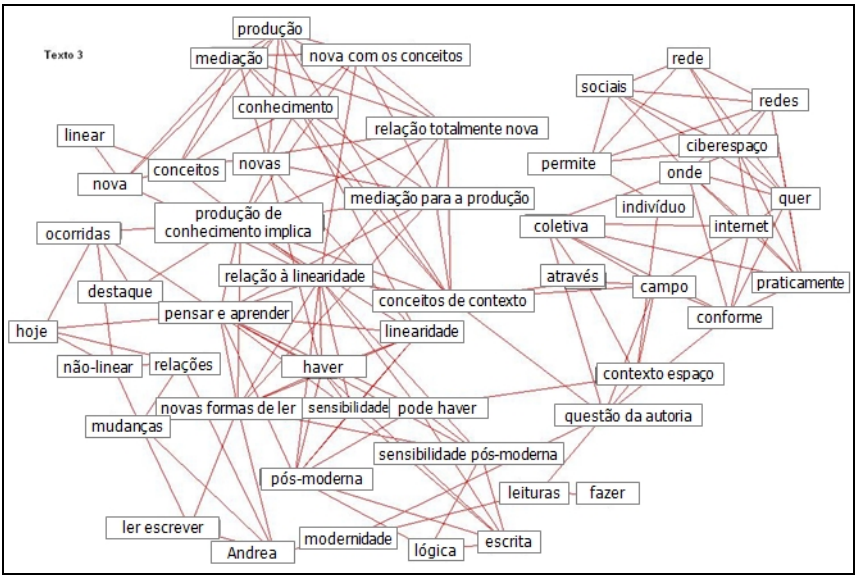

Fig. 4. Graph extracted from students' text – group 3

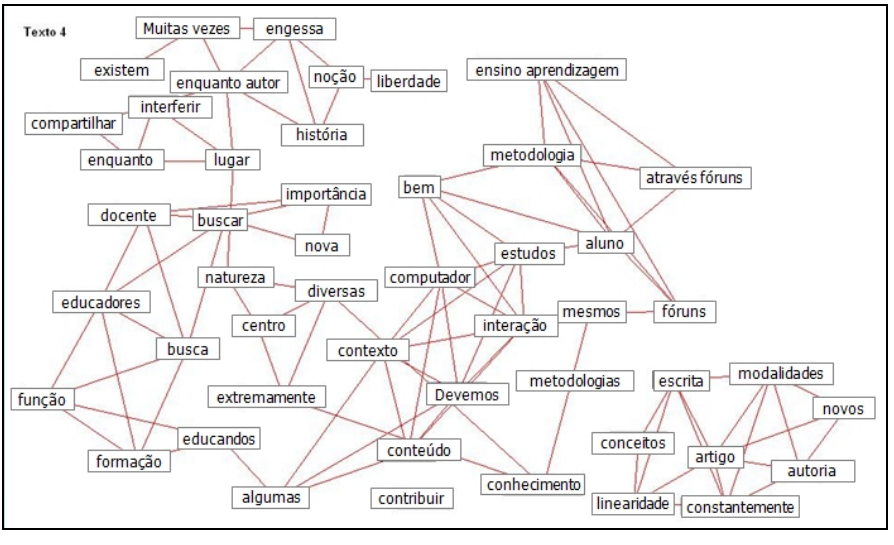

Fig. 5. Graph extracted from students' text – group 4

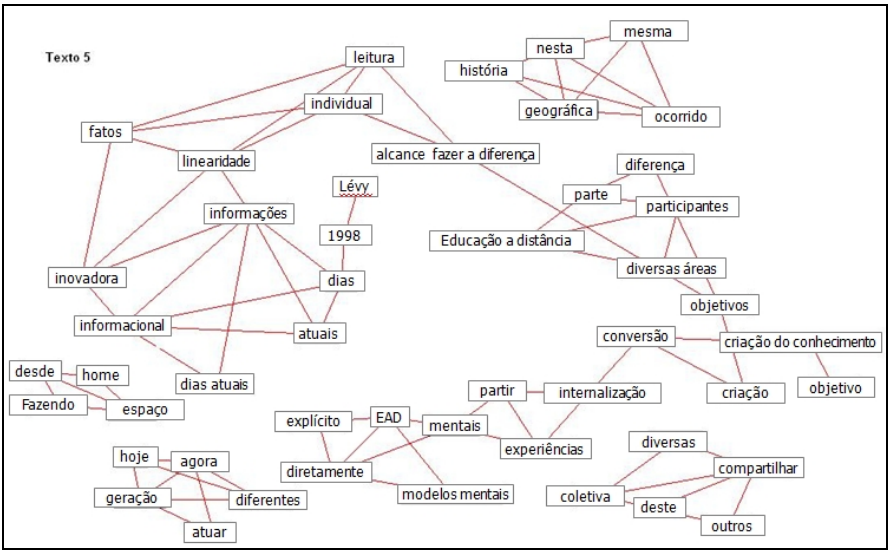

Fig. 6. Graph extracted from students' text – group 5

By looking at the graphs in figures 2 to 6, it was possible to say that they contained important concepts related to the central topic of the assignment, such as[1].

- Text 1: authorship, text, original
- Text 2: learning, author, construction, hypertexts, forms of expression
- Text 3: authorship, mediation, linear/non-linear, collective, individual, cyberspace, social, network.
- Text 4: sharing, educators, student, authorship, linearity, writing, forums, teaching-learning
- Text 5: authorship, individual, linear, innovation, distance education, knowledge creation, collective, mental models

It was noticeable in the list of terms above that the number of important concepts appearing in the graph of text 1 was much smaller than in the other graphs, which may signal out that text 1 did not discuss extensively other relevant topics related to the main theme. This hypothesis was later confirmed by the evaluation of the original text.

A further analysis of the graphs may show other characteristics of the texts from which they originated. For instance, graphs that were composed of smaller isolated terms and sub-graphs matched their corresponding texts where concepts were also treated in an isolated fashion. It was noticeable that these texts were created as a juxtaposition of paragraphs, and not as a fluid exposition of ideas and relationships between terms related to the central theme. In the examples presented, these smaller sub-graphs occurred more in texts number 2 and 5, where the actual reading of the documents confirmed that the connection between paragraphs in the texts were not

[1] The terms appearing in the graph, originally in Portuguese, have been translated here to make it easier for the reader to understand this section of the paper.

fluid. Text number 3, on the other hand, had a different and better writing style ,where the main concepts were considered and related throughout the text. The same can be said about text number 4, even if number 3 was the most consistent of all.

Considering this same premise, the graph extracted from Text 1 did not have isolated concepts as in the graph extracted from texts 2 and 5, but it also did not present significant terms related to the central theme proposed. A brief look at Text 1 was sufficient to demonstrate that the authors did not treat any subject in depth. The text spoke about the general theme proposed, and followed by presenting the interpretation and re-writing of the same subject by each collaborator, without bringing new information that related to the central topic.

In this sense, the text mining tool may provide positive and/or negative clues about a text, enabling the identification of problems such as: the need for further exploration of a given topic; the need to produce a text that is more fluid, and not only a juxtaposition of paragraphs that are not well connected.

5 Discussion and Final Considerations

The main contribution of this work has been to propose the use of a text mining tool embedded in a collaborative writing system, and to show how it could support the qualitative evaluation of written material produced by students. The results obtained from a preliminary evaluation of the system showed that the graphs elicited from the students' writings may show intrinsic characteristics of the texts that can lead teachers to further evaluate the students' work regarding certain problems, such as the need for additional development of a given topic, or the need to produce a more fluid text.

Another contribution of this work has been to propose an improvement in a known text mining process based on the use of graphs, as to produce more knowledgeable outcomes. While the original method generated graphs with one single term represented in each node, in our approach several terms could be placed in a graph node. It could be argued that by connecting nodes with words that appear together frequently in the text, one could represent concepts just the same way we do by placing them together in a single node. However, for the user who has to interpret the graph, it is more difficult to grasp the meaning of a compound term that is dispersed in different nodes, than if all of them were represented in a single node. A possible future development could be the comparison of other text mining techniques with the chosen technique based on Schanker's graph extraction. The idea of building a new text mining tool instead of using an existing application has been mainly because we wanted to develop some features that did not exist in other software, such as the capability of building a base of concepts from a set of papers, and getting the tool to consider only those concepts in the generation of a graph from students' writings. Besides, as we needed to integrate the mining tool to ETC, and to adjust many of its functions to our educational application, we understood that the best way to do it would be to build the application from scratch.

Natural Language Processing (NLP) is another approach that deals with textual data. Although it is easy for a human being to understand a document written in natural language, developing algorithms that can understand and extract the meaning of a document is a big challenge. Therefore, in practice NLP is frequently combined

with statistical analysis in order to build more accurate systems for the understanding and the interpretation of textual data [10]. In our case, Sobek's text mining approach is based exclusively on statistical analysis, which has the down side of sometimes eliciting from documents terms that would not really be relevant. A possible solution is to work with a database of concepts previously formatted and to use a mechanism such as WordNet to take into account synonyms, as one may find nowadays in different application such as in text categorization [11].

Other known text mining methods group together terms in order to make more accurate concept extraction from texts, as in [8] where relevant morpho-syntactic patterns are searched for in order to create meaningful tokens. While such procedure relies on the some level of linguistic processing, our approach is much simpler in that it is based mainly on a statistical analysis of the frequency with which the complete tokens appear in the texts.

As in [12], it has been observed that the simple application of statistical analysis on small texts can, in many cases, produce undesirable results, and that's inevitable. In order to deal with this problem, a complimentary process of using a database of concepts before mining students' contributions is also being considered in the next version of Sobek.

The use of Sobek by students, instead of teachers, is another research that is starting in our group, which aims at verifying how students could benefit from automatically seeing summaries of their writings.

References

1. Tammaro, S.G., Mosier, J.N., Goodwin, N.C., Mosier, J.N.: Collaborative Writing Is Hard to Support: A Field Study of Collaborative Writing. Computer Supported Cooperative Work: The Journal of Collaborative Computing 6, 19–51 (1997)
2. Juan, A.A., Daradoumis, T., Faulin, J., Xhafa, F.: Developing an Information System for Monitoring Student's Activity in Online Collaborative Learning. In: Proceedings of the international Conference on Complex, intelligent and Software intensive Systems, CISIS, March 04 - 07, pp. 270–275. IEEE Computer Society, Washington (2008)
3. Hodges, G.C.: Learning through Collaborative Writing. Literacy and Language 36(1), 4–10 (2002)
4. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
5. Wikipedia. Global Warming, `http://en.wikipedia.org/wiki/Global_warming` (accessed in December 2008)
6. Berry, M.J.A., Linoff, G.: Data Mining Technniques For Marketing, Sales, and Customer Support. Wiley, Chichester (1997)
7. Schenker, A.: Graph-Theoretic Techniques for Web Content Mining. PhD thesis, University of South Florida (2003)
8. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., Zamir, O.: Text Mining at the Term Level. In: PKDD 1999. LNCS, pp. 65–73. Springer, Heidelberg (1998)

9. Greengrass, E.: Information retrieval, a Survey (2001), http://clgiles.ist.psu.edu/IST441/materials/texts/IR.report.120600.book.pdf (accessed in December 2008)
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
11. Gómez Hidalgo, J.M., Cortizo Pérez, J.C., Puertas Sanz, E., Ruíz Leyva, M.: Concept Indexing for Automated Text Categorization. In: Natural Language Processing and Information Systems: 9th International Conference on Applications of Natural Language to Information Systems, pp. 195–206 (2004)
12. Leite, D.S., Rino, L.H.M., Pardo, T.A.S., Nunes, M.G.V.N.: Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In: TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp. 17–24. Association for Computational Linguistics, Rochester-NY (2007)