Process Mining to Support Students' Collaborative Writing

Vilaythong Southavilay¹, Kalina Yacef¹ and Rafael A. Calvo² ¹{vstoto, kalina}@it.usyd.edu.au ²rafa@ee.usyd.edu.au ¹School of Information Technologies, University of Sydney ²School of Electrical and Information Engineering, University of Sydney

Abstract. Writing, particularly collaborative writing is a commonly needed skill. Investigating how ideas and concepts are developed during the process of writing can be used to improve not only the quality of the written documents but more importantly the writing skills of those involved. In this paper, process mining is used to analyze the process that groups of writers follow, and how the process correlates to the quality and semantic features of the final product. Particularly, we developed heuristics to extract the semantic nature of text changes during writing. These semantic changes were then used to identify writing activities in writing processes. We conducted a pilot study using documents collected from groups of undergraduate students writing collaboratively in order to evaluate the proposed heuristics and illustrate the applicability of process mining techniques in analyzing the process of writing.

1. Introduction

Nowadays, Computer-Supported Collaborative Learning, particularly Collaborative Writing (CW) is widely used in education. Students often use computers to take notes during lectures and write essays for their assignments. Thanks to the availability of the Internet, students increasingly write collaboratively by sharing their documents in a number of ways. This has led to increased research on how to support students' writing on computers.

Our research motivation is to investigate ways to support these collaborative writing activities by providing feedback to students during the collaborative writing process. We have built a prototype which, based on the text written by students, can automatically perform analysis on text changes in order to discover types of semantic changes and identify writing stages.

In order to develop a tool supporting CW, understanding how ideas and concepts are developed during writing process is essential. The process of writing consists of steps of writing activities. These steps of writing activities can be considered as the sequence patterns comprising of both time events and the semantics of changes made during those steps. Therefore, process mining, which focuses on extracting process-related knowledge from event logs recorded by an information system, can be used to extract sequence patterns of writing activities that lead to quality outcomes.

In this paper, a process mining technique is used to detect semantic changes in writing activities in order to gain insight on how student write their documents. Our work used a taxonomy of writing activities, proposed by Lowry et al. [4]. In addition, we used a model developed by Boiarsky [2] for analyzing semantic changes in the writing process.

Although process mining techniques have been successfully applied to extract processrelated knowledge from event logs recorded by information systems [3] for two decades, the techniques have only recently applied to educational data. For example, in Pecheniskiy et al [8], the authors utilized process mining tools to analyze data from online multiple choice examinations and demonstrated the use of process discovery and analysis techniques. They were interested in individual students' activities of answering online multiple-choice questions during assessment, not in activities where students write and edit texts collaboratively. In addition, there has been abundant research for improving the support of quality writing in education such as automatic scoring of essays [10], visualization [6], and document clustering [1]. However, these approaches, unlike ours, focus on the final product, not on the writing process itself.

The remainder of the paper is organized as follows. In Section 2, the framework for supporting CW is presented. The heuristics for extracting semantic changes and writing activities during writing process are proposed in Section 3. A pilot study mining student writing processes is discussed in Section 4. Finally, Section 5 concludes and outlines and our future work planned in this area.

2. WriteProc

The framework, WriteProc for exploring collaborative writing processes was introduced in [12]. It integrates a front-end writing tool, Google Docs, which not only supports collaborative writing activities, but also stores all revisions of documents created, shared, and written by groups of authors. WriteProc uses two process and text mining tools, ProM [9] and TML [13]. ProM provides a way to extract knowledge about writers' activities and collaboration. In [12], process mining was used to extract only patterns of students' interaction and collaboration for peer review. In this paper, the process analysis is used for identifying sequence patterns of writing activities that lead to a positive outcome and indicate patterns that may lead to problems. TML analyzes text changes between individual revisions of documents. This analysis can provide semantic meaning of changes in order to gain insight into how writers develop ideas and concepts during their writing process. Full details about WriteProc can be found in [12].

In our pilot study, WriteProc retrieves some revisions from Google Docs of all the documents written by the students, using Google Document Data API. Google Docs automatically saves documents every few seconds. Authors can also deliberately save their documents. In Google Docs, every saving command (either auto or committed) produces one revision for each document. WriteProc does not retrieve all revisions of the documents, as many of them do not contain any changes. For a particular document, it first obtains the revision history (log) containing information, such as revision number, timestamp of auto and committed saving, and author id. The revision history is then analyzed to identify writing sessions for individual writers. From this analysis, for each session WriteProc downloads a revision after each minute of writing. These downloaded revisions are then pre-processed to extract and index paragraphs and sentences for further analysis. Using predefined heuristics (defined in the next section), WriteProc performs a text-based analysis on the indexed revisions of each document which extracts semantic meaning of text changes and identifies writing activities. Sequences of writing activities

are then analyzed using process mining. In the next section, we propose the heuristics and explain how they will be used to detect writing activities.

Table 1. Heuristics for determining collaborative writing activities: Brainstorming (B), Outlining (O), Drafting (D), Revising (R), and Editing (E) based on text change operation (C1 – C8), text structure (S1 – S2), and functions (F1 – F3).

Abbreviation: An operation is allowed (Y) or not allowed (N), List (L), Structured List (SL), Sections and Paragraphs (Sec. & Par.), Number of Sentences (S), Number of Paragraphs (P), Changing/Fluctuating (F), Constant(C), Improving (I), and Not applicable (N/A).

Writing activities	Surface change	Reorganization of information	Consolidation of information	Distribution of information	Addition of information	Deletion of information	Alteration of form (Macro-structure change)	Micro-structure change	Structure	Number of Sentence vs Number of Paragraphs	Ratio of Number of Words	Topic Overlap	Cohesion Measure
	C1	C2	C3	C4	C5	C6	C7	C8	S1	S2	F1	F2	F3
B	Y	Y	Y	Y	Y	Y	Y	Y	L	$\mathbf{S} \approx \mathbf{P}$	F	N/A	N/A
0	Y	Y	Y	Y	Y	Y	Y	Y	SL	$S \approx P$	F	N/A	N/A
D	Ν	Ν	Ν	Ν	Y	Y	N**	Y	Sec. & Par.	S > P	F	F	F
R	Ν	Y	Y	Y	N*	N*	Y	N*	Sec. & Par	S > P	F	F	I/C
Ε	Y	Ν	Ν	Ν	Ν	Ν	Ν	Ν	Sec. & Par.	S > P	С	С	С

3. Heuristics for determining collaborative writing activities

Writing activities in collaborative writing can be categorized into 6 common activities: brainstorming, outlining, drafting, reviewing, revising, and editing. The definition of these activities is described in [4]. It is important to note that in general these six activities do not occur in a linear sequence. In a document writing process, we consider reviewing activities made not only by the writers (owners) of the document, but also by instructors or editors or peers who read and annotate the document for content, grammar, and style improvements. In this work, we concentrate on automatically identifying the five collaborative writing activities: brainstorming (B), outlining (O), drafting (D), revising (R), and editing (E).

In order to identify the five collaborative writing activities in the writing process of a particular document, basic heuristics are proposed. Particularly, our heuristics are based on text changes, text structures, topic changes and cohesion improvement in the document from one revision to another. The heuristics utilized in our analysis are presented in Table 1. Each writing activity can be identified using text change operations (C1 to C8), text structures (S1and S2), and functions (F1 to F3), which are explained below.

3.1 Text structures

The writing activities can be determined by the structure of the written texts (S1) and the number of sentences and paragraphs (S2). During brainstorming, authors normally write in bullet-point lists consisting of single words or phrasal words (compound nouns). Consequently, the number of paragraphs (the number of lines) is approximately equal to the number of sentences (the number of words or items). Although during an outlining phase the number of paragraphs and sentences are still the same, the text structure is more organized into sections and subsections. When writers start drafting their documents, number of sentences and paragraphs change dramatically. During this phase, the number of sentences is expected to be higher than the number of paragraphs. This is also truth for revising and editing phases.

3.2 Text change operations

Eight types of text change operations (C1 - C8) were used in our heuristics. These text change operations were based on the revision change functions proposed by Boiarsky [2]. Writers use the text change operations in their writing activities for different purposes in order to produce final pieces of writing. We developed basic assumptions for our pilot study as following:

- During brainstorming, writers can reorder, adding, or deleting items of lists of brainstorming ideas. They can also format the lists, alter the whole items of the lists, or change some items. Similarly, during outlining, writers can add, delete, reorder, format, and change some or the whole sections of their organized list.
- During drafting, revising and editing, text change operations become more complicated. Drafting activities start when the structure of the written text changes from bullet-point or structured lists to paragraphs. In other words, alteration of form (C7) usually indicates the start of drafting activities (as depicted by N** in the table). During drafting, information is added and removed all the time. Therefore, expansion of information (C5), deletion of information (C6), and micro-structure change (C8) imply drafting activities.
- However, if C5, C6 and C8 happen after reviewing activities, we consider them as revising activities (as noted by N* in the table). Particularly, peer review was incorporated in our pilot study. We assumed that after getting feedback from their peers, students may add, delete, and alter texts in their documents. Also students may completely erase the whole written text and rewrite the text from scratch after getting feedback from their peer review. This operation is C7.
- In addition, common revising activities are reordering (C2), consolidating (C3), and distributing texts (C4). These changes occur after writers start drafting and reoccur many times in the writing process. Our assumption is that during drafting writers may stop writing and revise their own written texts in order to improve the cohesion and effectively convey information and ideas to readers.
- For simplicity, all surface change operations (C1) including formatting, spelling and punctuation corrections are consider to be editing activities. Similar to C2 -

C4, editing activities are considered to be common and reoccur many times in the writing process.

3.3 Number of words

The ratios between the number of words of two consecutive revisions are computed (F1). The ratio was used in conjunction with topic overlap and cohesion measurement discussed below to determined writing activities.

3.4 Topic overlap

Our heuristics also used a topic overlap measurement (F2). We analyze the change in topics (concepts) for two consecutive revisions of documents. Our intuition is that when people write about something, they usually repeat the subject (topics) to keep readers' attention. By identifying concepts and comparing them between two consecutive revisions of pieces of writing, we gain information on how writers develop their idea and concept during writing process. Intuitively during drafting and revising, topics overlap (F2) changes dramatically. However, during editing F2 should be constant. The method for computing the topic overlap (F2) is described in Section 4.1.

3.5 Cohesion Changes

Another measurement used in our heuristics to detect writing activities is the cohesion of the text. We measure cohesion of each individual revision of documents. Particularly, we calculate the distance between consecutive sentences and paragraphs in the written texts in order to gain an insight on how paragraphs and the whole texts have been developed. Our assumption is that during a drafting phase, the cohesion of the written text fluctuates a lot. After authors revise the text, the cohesion of the text is usually improved. There should be no change in the cohesion of the written text during editing phase.

We use the Latent Semantic Analysis (LSA) technique to measure the cohesion of the text. In particular, for each revision of documents we compute average sentence and paragraph similarities using LSA for single documents as described in [14] and compare the result with the former revision of the same documents in order to determine if there is an improvement in cohesion for these two revisions.

4. Pilot study

As a way of evaluating the proposed heuristics and illustrating how process mining can be used to analyze writing activities, we conducted a pilot study to investigate writing processes of students in the course of E-business Analysis and Design, conducted during the first semester of 2009 at the University of Sydney. In this course, students were organized in groups of two and asked to write Project Specification Documents (PSD) of between 1,500 and 2,000 words (equivalent to 4-5 pages) for their e-business projects. They were required to write their PSD on Google Docs and share them with the course instructor. The course also used peer review in which each PSD was reviewed by other two students who were members of different groups. After getting feedback from their peers, students could revise and improve their documents if necessary before submitting the final version one week later.

In addition, the marks of the final submissions of the PSD (as presented in [12]) together with a very good understanding of the quality of each group through the semester was used to correlate behaviour patterns to quality outcomes. In particular, to be able to give insight into how students wrote their own documents, we performed a process diagnostic to give a broad overview of students' writing activities.

4.1 Pre-processing

In this section, we describe the document pre-processing method used in our study. We analyzed 21 documents in this study. As discussed in Subsection 3.4, LSA was used for measuring the changes in cohesion in the written text. The pre-processing step for LSA involved the extraction of the terms from all concerned revisions of the documents. First, each revision of documents was split into paragraphs using a simple matching to the newline character. Then, each paragraph was divided into sentences using Sun's standard Java text library. After that, each sentence, paragraph and the whole text were indexed using Apache's Lucene, which performed the tokenization, stemming, and stop word removal. We used Porter's stemmer algorithm (Snowball analyzer integrated in Lucene) for stemming words. Next for each revision, a term-document matrix was created. Term frequency (TF) and document frequency (DF) were selected as weight terms. We discarded terms that only appear once in each revision of documents. The space of term-document matrix was reduced using Singular Value Decomposition (SVD). We adopted the method of Villalon et al. [14] to reduce the dimension of the LSA space to 75% of the total number of sentences.

In order to compute the topic overlap discussed in Subsection 3.3, we first extracted topics from each revision of documents. Our approach in extracting topics from each revision of documents was based on Lingo clustering algorithm developed by Osinski et al. [7]. Especially, we extracted frequent phrases from each revision (we use the assumption and definition of frequent phrase discussed in detail in [7]). Next, by using the reduced term-document matrix calculated for LSA above, for each revision. After discovered any existing latent structure of diverse topics in a particular revision. After discovering topics of each revision of documents, we compared topics of two consecutive revisions to calculate the topic overlap between the two revisions. As a baseline measure, we selected a simplistic word overlap measures. This measure was used for sentence-level similarity such as in the work of [5].

The final step in our data preparation was to create a writing activity log of all documents in the format used by a process mining tool like ProM. First, for each revision (except the first revision) we compare it to the former revision and obtain the types of text change operations between the two using a file comparison utility like *diff* tool. Then, we applied the proposed heuristics using the obtained types of text changes, the results of LSA cohesion and topic overlap calculated above. In conjunction with timestamp and user identification obtained from the revision history discussed in Section 2, we can create an event log of writing activities, in which process analysis can be performed.

4.2 Writing process analysis

After preprocessing, the resulting event log consisted of 8,233 events in total. Each process case represented one document. The mean number of events per document was 392, with a minimum of 77 events per document and a maximum of 705 events per document. There were 6 different types of events corresponding to 6 types of writing activities. We performed process analysis in order to gain insight into how students wrote their documents.

The Dotted Chart Analysis utility of ProM [9] was used to analyze students' writing activities. The dotted chart was similar to a Gantt chart [11], showing the spread of events over time by plotting a dot for each event in the log. Figure 1 illustrated the output of the dotted chart analysis of students writing their PSD documents. All instances (one per document) were sorted by start time. In the figure, points represented writing activities occurring at certain time. There were six writing activities as mentioned above. They were represented with different colors and shapes: white circles denoted brainstorming, brown circles were outlining, black triangles represented drafting, black squares depicted reviewing, brown squares were revising, and white diamonds denoted editing activities.

4.3.2009 8.31:0 19	11.3.2009 8:31:0	18.3.2009 \$31:0	25.3.2009	1.4.2009 8:31:0	8.4.2009 7.310
g21 🔍 🛡	. .				
çı23	▼	** *	N N NY NY NY H		
g18		*			n sin sin
g 24		ं । ज रह		=	
çı 13	•	V 1	NY NY TANÀN 🖬 👘		
çı01			< ∎		
çı 16		0 00 0	MAIN CH B		•
çı07					
g14		iy wa			
çı 10		•			
çı27					
g17			SV KM E		C
çı02		X			
çı04					-
çı29			▼ ▼ ■		
çı 15			A AZ IR W		
çı22					
çı 2 5					
ç109			X9W		
g 26			W H		

Figure 1. Dotted chart of 21 groups of students writing collaboratively (from ProM tool). White circles denoted brainstorming, brown circles were outlining, black triangles were drafting, black squares were reviewing, brown squares were revising, and white diamonds were editing activities.

From the figure, we observed that most students started their writing approximately two weeks before the due date for the submission of peer review (27^{th} March 2009). Exceptionally, there existed 6 groups starting their writing activities quite early. Group 19 (received the final mark of 9/10) and 21(10)¹ started their outlining and brainstorming activities very early in the first week of the semester. Group 13(9), 18(9), 23(8), and 24(8) also commenced their writing tasks early in the second week of the semester.

¹ Group X(Y) denotes the group number X receives the final mark of Y out of 10.

Therefore, we noticed that students who performed brainstorming and outlining started their writing quite early and received quite good marks. In addition, as we expected, there were many writing activities during a week before the due date for peer review submission. Interestingly, there were 5 groups commencing their writing tasks few days before the due date. They were Group 9 (10), 15 (9), 22 (9), 25 (9), and 26 (9). These groups also received quite high marks for their writing, although they started writing quite late. Actually students of these groups did not use Google docs to perform outlining and brainstorming. They also started their writing using other word processing applications such as MS Word because all of them commenced their writing on Google Docs with substantial amount of texts (containing sections and paragraphs). During the one-week of peer review, we expected to have no writing activities since students were waiting for the feedback from peer review. However, Group 22, who just started their writing on Google Docs, performed some activities during this time. We checked the revisions of the document of Group 22 and found that there were substantial text changes performed by these activities. At the end they received a good mark of 9/10. Furthermore, after getting feedback from their peer review (3rd April 2009), students started revising and editing their documents before the final submission (10th April 2009). We observed that Group 16(9), 18(9), and 24(7) started working on their documents soon after getting feedback. They were among top groups in the class.

In addition, we were naturally interested in finding out more about writing activities of each group and the path each group was following in the process of writing. ProM provides a Performance Sequence Analysis (PSA) plug-in to find the most frequent paths in the event log [3]. Figure 2 illustrates a sequential diagram of students' writing activities in our pilot study. The patterns were ordered by the number of groups generating them. From the analysis above, we learned that not all groups of students performed their brainstorming and outlining before actually drafting their documents. The PSA also confirmed this. In addition, from the PSA we checked each individual group's activities in order to determine which groups did not conduct brainstorming and/or outlining for their writing tasks. From Figure 2, there were seven distinct patterns of activities. Pattern 0 and 5 indicated groups that started drafting without brainstorming and outlining. Pattern 0 was originated from 8 groups: 2, 9, 13, 14, 15, 22, 25, and 28. For Pattern 5, there was only one group: 26. Pattern 1, 3 and 4 involved all activities except brainstorming. There were 7 groups belonged to Pattern 1: 1, 4, 7, 10, 16, 17, and 27. For Pattern 3 and 4, each of them was generated by only one group namely 19 and 23, respectively. Clearly, more than half of the class did not conduct outlining and one third of the class did not bother performing brainstorming (at least on Google Docs) before drafting. We discussed this matter with the course instructor who was aware that most students performed their writing plans offline. Finally, there were 3 groups whose their writing activities generating Pattern 2 and 6: 18, 21 (for Pattern 2), and 27 (for Pattern 6). These groups planed their writing tasks with brainstorming and outlining. Consequently, all of them received high marks.

The process model of all 21 documents was discovered by using the Heuristic miner algorithm [15] with default threshold parameters (implemented in ProM). Figure 3 depicts a transition diagram of the model. The numbers in the boxes indicate the frequencies of the writing activities. The decimal numbers along the arcs show the probabilities of transitions between two activities and the natural numbers present the

number of times this order of activities occur. In addition, we also extracted the process model of each individual group in order to gain an insight on how the group conducting its writing.

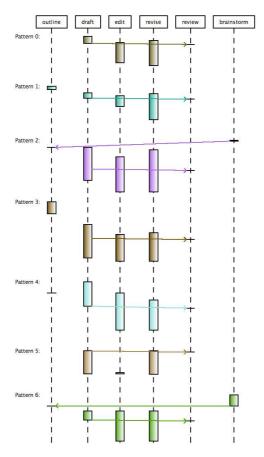


Figure 2. Sequence patterns of 21 groups of students writing collaboratively. The patterns are ordered by the number of groups generating them (from ProM tool).

5. Conclusion and future work

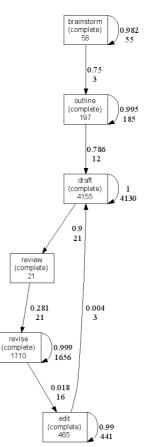


Figure 3. A transition diagram of the process model for 21 documents. The natural numbers refer to frequencies (of events and transitions), and the decimal numbers represent probability of the transitions.

The work presented in this paper is a work in progress. The pilot study described in the previous section provides fundamental work for us to develop basic heuristics to extract semantic meaning on text changes and determine writing activities. Based on the heuristics, we were able to analyze student's writing activities using a process mining and discover 7 patterns on writing activities of 21 groups of students. However, correlated with final assessment, we could not distinguish clearly the better from the weaker groups.

This preliminary work gives us direction for the next step of our work. In the future, the discovered patterns, process snapshots provided by performance sequence and dotted chart analysis can be used for providing feedback to students during their writing so that they are aware of their writing activities and can coordinate effectively.

One way to improve our understanding of what writing processes lead to better outcome is to improve our heuristics. In our current work, the surface change operation indicates only changes in spelling, number, punctuation, and format. We did not include grammatical correction in the current work yet. In addition, one of text change operations proposed by Boiarsky is the improvement in vocabulary [2]. We did not detect the improvement in vocabulary in our current analysis. These two text change operations will be cooperated in the heuristic in the future. In addition, we already measure the change in topics (concepts) which represent word repetition. Although word repetition is common in writing, good writers usually utilize synonymy and pronouns to avoid annoying repetition. This issue was not considered in this paper and will be cooperated in the future work.

Acknowledgements

This work has been funded by Australian Research Council DP0986873. We would like to thank the many people involving in the development of ProM and TML.

References

[1] Andrews, N. O. and Fox, E. A. *Recent Developments in Document Clustering*. Technical Report TR-07-35, Computer Science, Virginia Tech, 2007.

[2] Boiarsky, C. Model for Analyzing Revision. *Journal of Advanced Composition*, 1984, 5, p. 65-78.

[3] Bozkaya, M., Gabriel, J. and Werf, J. M. E. M. v. d. Process Diagnostics: A Method based on Process Mining. *International Conference on Information, Process, and Knowledge Management*, 2009.

[4] Lowry, P. B., Curtis, A. and Lowry, M. R. Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice. *Journal of Business Communication*, 2003, 41, p. 66-99.

[5] Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A. and Zobel, J. Similarity Measure for Tracking Information Flow. *Proceeding of the 14th ACM International Conference on Information and Knowledge Managment*, Bremen, Germany, 2005, p. 517 - 524.

[6] O'Rourke, S. and Calvo, R. A. Semantic Visualisations for Academic Writing Support. *14th Conference on Artificial Intelligence in Education*, Brighton, UK, 2009, p. 173-180.

[7] Osinski, S., Stefanowski, J. and Weiss, D. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference*, Zakopane, Poland 2004, p. 359--368.

[8] Pechenizkiy, M., Trcka, N., Vasilyeva, E., Aalst, W. M. P. v. d. and Bra, P. D. Process Mining Online Assessment Data. *Educational Data Mining*, Cordoba, Spain, 2009.

[9] ProM. http://prom.win.tue.nl/tools/prom, 2010.

[10] Shermis, M. D. and Burstein, J. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Volume 16, MIT Press 2003.

[11] Song, M. and Aalst, W. M. P. v. d. Supporting Process Mining by Showing Events at a Glance. *7th Annual Workshop on Information Technologies and Systems*, 2007, p. 139-145.
[12] Southavilay, V., Yacef, K. and Calvo, R. A. WriteProc: A Framework for Exploring Collaborative Writing Processes. *Australasian Document Computing Symposium*, Sydney, 2009.
[13] TML. *http://tml-java.sourceforge.net*, 2010.

[14] Villalon, J. and Calvo, R. A. Single Document Semantic Spaces. *The Australasian Data Mining conference*, Melbourne, 2009.

[15] Weijters, A. J. M. M., Aalst, W. M. P. v. d. and Medeiros, A. K. A. d. Process mining with the heuristics miner-algorithm. *BETA Working Paper Series WP 166*, Eindhoven University of Technology, NL, 2006.