Analysis of Collaborative Writing Processes Using Hidden Markov Models and Semantic Heuristics

Vilaythong Southavilay, Kalina Yacef
School of Information Technologies
University of Sydney
Sydney, Austalia
e-mail: {vstoto,kalina}@it.usyd.edu.au

Rafael A. Calvo
School of Electrical and Information Engineering
University of Sydney
Sydney, Australia
e-mail: rafa@ee.usyd.edu.au

***Abstract*— In this paper we are interested in discovering collaborative writing patterns in student data collected from a system we designed to support student collaborative writing, and which has been used by over 1,000 students in the past year. A particular functionality that we are investigating is the extraction and display to learners and teachers of the process followed during the course of the writing. We used a heuristic to derive semantic interpretation of specific sequences of raw data and Markov models (MM) to derive the processes. We propose two models, a Heuristic MM and a Hidden MM for analysing student's writing behavior. We also refined the semantic preprocessing by adding the notion of pauses between activities. We illustrate our approach and compare these models using real data from two groups of high and low performance level and highlight the different information they each provide.**

Keywords-text mining; process mining; collaborative writing; hidden Markov model;

I. INTRODUCTION

The availability of the Internet has made collaborative writing very easy to implement in schools and at work. This leads to new forms of writing, such as blogging and wiki writing. In addition, the emerging of "cloud computing" tools and Web 2.0 applications such as Google Docs has led to the availability of almost desktop-quality online writing environments. However, despite the widespread of such tools supporting collaborative writing, there is very little support to improve the quality of the *process* of collaborative writing. Our overall aim is to creating adequate support for these writing activities for improving not only the quality of documents, but also the collaborative writing skills of the authors involved.

There has been abundant research applying data and text mining techniques in order to extract semantic aspects of text and documents. However, this research focuses on the final product of writing (i.e. documents), rather than the process of writing (i.e. writing activities). We are interested in the latter.

In the process of writing, authors usually start writing by adding some text and continue on modifying and editing the text to form ideas (concepts) in the document and elaborate them. The authors may add or delete text to introduce new ideas or remove irrelevant ones, or make surface changes to the document. We call these *text change operations*.

The process of writing can be decomposed into a sequence of *writing activities*, such as brainstorming, outlining, drafting and so on. In practical terms, each writing activity transfers the document from one revision to the next. In addition, there are also revisions that do not contain significant changes. Therefore, a particular document consists of many revisions over time.

In order to match the revisions with writing activities, some semantic interpretation of the text changes is needed. We use text mining techniques, along with heuristics to identify the writing activities from the text. For a particular document we compare the text between consecutive revisions to identify the nature of the text change operation generating the revision. We also measure the change of topics and cohesiveness from one revision to another one using topic extraction algorithms and Latent Semantic Analysis (LSA) techniques as proposed in [11]. Finally, using heuristics, we derive the most likely writing activity given a set of these measures [11]. Using these kinds of computed information, we can automatically infer the writing activities during the process of writing a specific document. Based on the writing activities, we can then investigate writing behavior patterns to gain an insight on the process that the authors follow while writing their document.

To discover writing behavior patterns, hidden Markov Models (HMMs) [9] are good candidates. Based on a sequence of text change operations, which can be measured and observed, the HMMs can extract writing states, which can not be directly measured as well as the transitioning probabilities among these states. However, Markov Models do not model the timing of activities and of transitions. We are exploring ways to include this important information in the model we propose. Indeed, in the process of writing, the duration that authors spend to write from the beginning to the end, the time they pause, and the number of revisions they make during the process of their writing are all potentially important.

In this paper, we first describe the framework to support our analysis of collaborative writing process in Section II. Section III presents a set of heuristics used to extract the semantic meaning of writing activities. Section IV presents our approach used to discover 2 models: Heuristic Markov model and hidden Markov model of the collaborative writing process. We then illustrate our approach and analysis with a

pilot experiment in Section V, before providing a discussion and conclusion in Section VI.

II. A FRAMEWORK TO SUPPORT COLLABORATIVE WRITING

We have developed a system to support students' collaborative writing [2]. The system has been used by over 1000 students in the past year. In this section, we briefly describe the framework and how the data used in our analysis was collected.

Our framework uses Google Docs [3] as a front-end writing tool. A document is created and assigned to groups of students by administrators or instructors. Students share their documents with their team members. They work on the documents by writing and editing them, and finally submit their final versions. Along the way they have access to feedback from the system if they wish to [10]. The aspect that is of interest in this paper is the fact that Google Docs stores all revisions (changes) of texts that students make. In addition, for each document, Google Docs records a revision history consisting of timestamps and author IDs for each revision of the document. Our system then downloads all revisions and revisions histories of particular documents for analysis.

III. EXTRACTING SEMANTIC MEANING OF WRITING ACTIVITIES

A way to extract the states of writing processes was proposed in [11], using a taxonomy developed by Lowry et al. [7]. The taxonomy consisted of five main stages: *brainstorming, outlining, drafting, revising and editing*. Using a revision model proposed by Boiarsky [1], we specified 8 types of text change operations, which are listed in Table 1. We then created a set of heuristics which map the combination of these text change operations, topic changes, and a topic cohesion measure (using Latent Semantic Analysis (LSA)) to stages of writing activities.

TABLE 1. SEMANTIC HEURISTICS FOR EXTRACTING WRITING ACTIVITIES

Writing Activity	**Text Change Operation**	
	Code	***Description***
Editing	C1	Surface Change
Revising	C2	Reorganization of Information
	C3	Consolidation of Information
	C4	Distribution of Information
Drafting	C5	Addition of Information
	C6	Deletion of Information
Drafting	C7	Alteration of Form (Macro-Structure change)
Drafting	C8	Micro-Structure Change
Pause	p	No change

Table 1 gives a simplified version of the heuristics (the complete heuristics, except for the Pause information, can be found in [11]). For example, when authors add (C5) or delete (C6) paragraphs, we consider these activities as drafting, whereas when they reorganize (C2) paragraphs, we interpret this activity as revising.

We added in this work a "Pause" activity (corresponding to "p" type of text change operation), which represents the event where authors did not make any change to their text. This (in)activity indicates a *pause* time in the writing, possibly because authors have stopped for thinking about and reflecting on the texts before starting to write again, or gone researching material for their writing, or any other reason. We believe that pauses in the writing, and time taken to complete an activity, have a potential impact on the interpretation of the process. An activity sequence of writing processes can include many consecutive long pauses. In our process mining, we considered these accumulated pauses as a delay (waiting) time of activities or events.

IV. EXTRACTING HMMS

We used both writing activities and text change operation sequences as inputs to the HMM generating algorithm (HMM constructor) to create two models of writing processes, which we then compared.

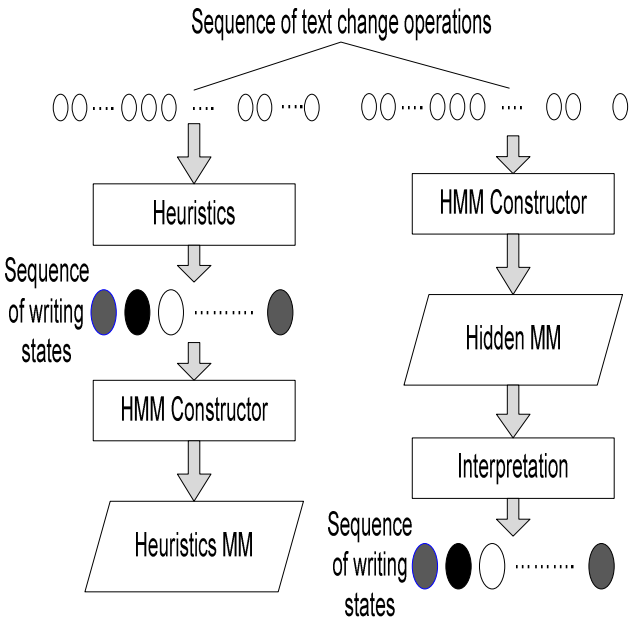

Figure 1, HMM model created with semantic heuristics on the left (Heuristic MM) and without the heuristics on the right (Hidden MM)

The first model, which we call a Heuristics MM, is depicted on the left of Figure 1. A heuristics MM is a Markov model created from a sequence of writing activities, derived from text change operations by applying the semantic heuristics explained in the previous section. The full construction is as follows: A sequence of text change

operations made in each revision forms the input to our set of heuristics. The result is a corresponding sequence of writing activities, which will become the states of the MM. Using the HMM constructor described in Subsection A below, a Heuristics Markov model of the collaborative writing process for the corresponding document is discovered from the input sequence of writing activities. Fig. 1 (left) depicts the process of extracting the Heuristics MM.

The second model, which we call a Hidden MM, is depicted on the right hand part of Figure 1. Unlike a Heuristics MM, a Hidden MM is built directly from the sequence of text change operations. A Hidden MM is a model with writing states as unobserved states. Using the sequence of text change operations, the HMM constructor is able to discover the structure of Hidden MM (i.e. a set of states and the output probability associated with each state) and other parameters (e.g. transition probabilities from one state to others or itself). The Hidden MM is then analyzed to interpret and identify writing states. Sequences of writing patterns can be discovered after the interpretation. The process of extracting the Hidden MM is shown in Fig. 1 (right).

So the difference between these two models is that in one case, writing states are derived prior to constructing the model whereas in the other case the model is built first, and the writing states are derived after.

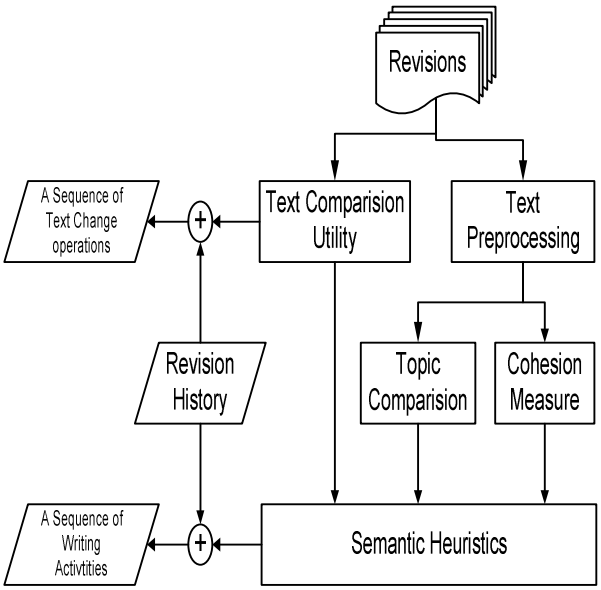

Figure 2, preprocessing steps

A. *Preprocessing*

First, preprocessing steps need to be performed in order to use student trace data for generating our two models Hidden MM and Heuristics MM. The preprocessing involves two main steps: (i) identifying the text change operation and, for the Heuristics MM, the corresponding writing activity producing each revision; (ii) creating sequences of text change operations and, for the Heuristics MM, writing activities. We describe these steps below.

As mentioned earlier, we use a method outlined in [11] to extract semantic meaning of text changes during the writing process. As illustrated in Fig. 2, this method does the following.

For each document, we have access to two sets of data. The first one is the revision history, which contains timestamps and author IDs of all its revisions. The second one contains the actual text of each revision. Text mining preprocessing techniques are used to perform the tokenization, stemming, and stop word removal for all the revision texts. Then, a text change operation type is identified for each revision by comparing it to its former version and applying our text comparison utility [11]. The result consists of a sequence of text change operations for that document. In addition, a topic change and a cohesion measure are also calculated using Lingo algorithm [8] and LSA, respectively. Using the heuristics proposed in [11], we can then associate a writing activity to each revision. Therefore, we obtain a sequence of writing activities for each document.

As mentioned in the previous section, inactive events (*Pause* activities and *p* text change operations) represent pauses in the writing process. In the Heuristic MM, these are replaced with the previous writing activity. For example, a sequence composed of a Drafting activity followed by 3 pauses and a Revising activity will become 4 Drafting activities followed by a Revising activity. In the Hidden Markov Model, pause events are replaced with the previous text change operation. For example, adding information, followed by a pause and a reorganization of information will become 2 adding information events followed by a reorganization of information.

The numbers of these inactive events are also used in calculating stationary probabilities [4]. The idea is to investigate whether the proportion of time that students spent in each of the writing activities has any importance. We used the notion of stationary probability as the relative proportion of activities that belongs to a certain state. In other words, the stationary probability of a state A is the proportion of occurrences of state A among all states that occur in a sequence of length n iterations generated by the model. n is normally the average number of activities in the input sequences.

The generated text change and activity sequences can each be used to derive a hidden Markov model for a document. We applied the HMM generating algorithms described in [4]. We use the Viterbi algorithm for sequential encoding in the segmentation step and the segmental K-Means [5] algorithm in conjunction with Li and Biswas's algorithm [6] for the optimization steps. The optimization steps find the optimal model parameters including initial probability vector, transition probability matrix, output

probability matrix and the number of states in the models. The models were used in our analysis below.

V. PILOT STUDY

In our pilot study, we analyzed student writing behaviors in an engineering course taught at our university in the first semester 2010. 53 students were enrolled. These students divided into 27 groups and had to write a collaborative document as part of an assignment. In this section, we will illustrate the techniques used to extract student's writing behavior and discuss the results obtained for 2 documents written by a high performing group and a low performing one (the notion of performance is based on the final mark the group obtained for the assignment).

Table II lists the number of activities and the average scores for these two documents.

TABLE II. THE NUMBER OF WRITING ACTIVITIES, PAUSES AND DURATION OF WRITING PROCESS

Document	Total Number of Writing Activities	Number of Pauses	Duration of Writing
High Performing	2398	1338	31.06 days
Low Performing	1451	743	31.56 days

A. Constructing the HMMs

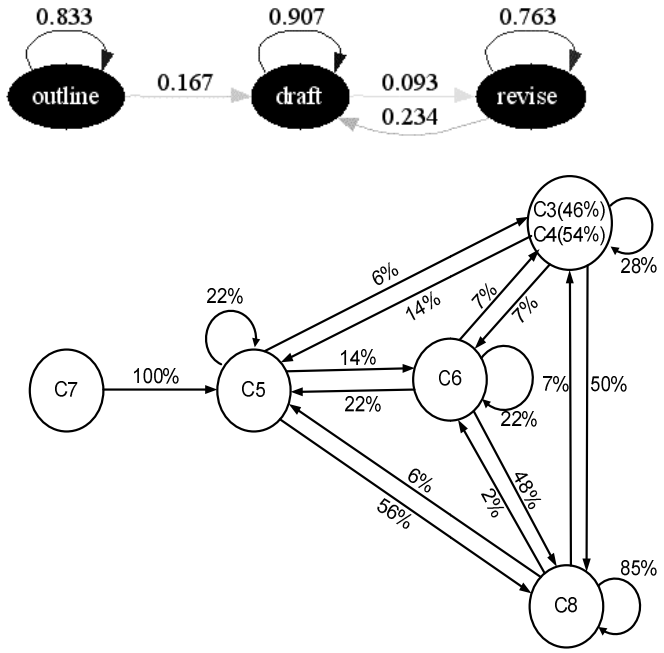

Figure 3, MMs of the document of a High Performing group (Heuristic MM and Hidden MM respectively)

The generated text change and activity sequences were used to derive two Markov models for each document.

The writing process models for two documents: High (H) and Low (L) performing groups creating using our HMM algorithms mentioned above are shown in Fig. 3 and Fig. 4. In each figure, the models at the top, with black states, represent the Heuristic MM and the ones at the bottom, with white states, represent the Hidden MM. The Hidden MM models are made up of a set of states, the text change operations' patterns (the output probability) associated with each state, and the transition probabilities between states. For example, the model discovers that authors of document H in the C3(46%)C4(54%) state performed combining paragraphs 46% of the time and distributing paragraphs 45% of the time. The transition probability associated with a link between two states indicates the likelihood of the authors transitioning from the current state to the indicated state. For instance, the model of H document predicts that in the C3(46%)C4(54%) state: after either consolidating and distributed text authors would add new text with a likelihood of 14%, delete existing text with a likelihood of 7%, change the text with a likelihood of 50% or remain in the same state (continue consolidating and distributing the existing text) with a likelihood of 28%. Likelihood less than 2% were not represented in the figure, explaining why the probability numbers do not exactly sum up to 100%. Similar to the Hidden MM, the Heuristic MM consists of a set of states and the transition probabilities. Since each states of Heuristic MM represent an entire writing activity, there is no output probabilities associated with each state.

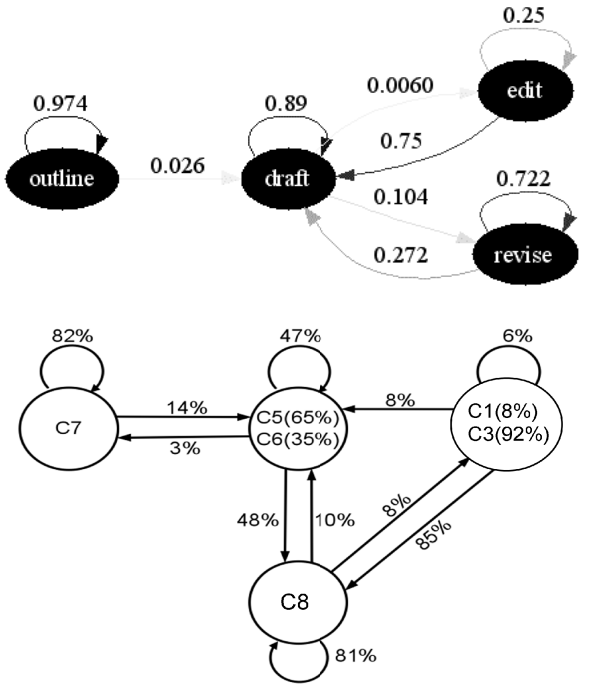

Figure 4, MMs of the document of a Low Performing group (Heuristic MM and Hidden MM respectively)

B. Analysis of the MMs

We investigated these MMs to gain insight on how students wrote their documents. The Heuristic MMs provide a good overview on how students wrote their documents and are very similar for the two groups. In contrast the HMMs (with white states) contain some noticeable differences and allow better to distinguish the high from the low performing group based on finer activities such as text change operations.

We further analyzed the HMMs to gain more insight on how students edited and modified their texts. The model of the high performing group (group H) has 5 states (Fig. 3), whereas the model of the low performing one (group L) has only 4 states (Fig. 4). For group H, we noted that the alteration of form operation (C7) happened only one time. We interpreted this as when the students of group H started drafting after completing outlining or brainstorming. They began their drafting activities by adding new paragraphs (C5). After that, they were likely to change the existing paragraphs (C8) because the C5 state has 56% likelihood of transitioning to C8 state, comparing to 14%, 6% and 22% likelihood of transitioning to C6, C3/C4 states and itself, respectively. Students most likely continued on changing the existing paragraphs to improve the cohesion of the text. This was confirmed by the fact that C8 state had the highest reiterating probability (80%). After changing the text until satisfaction, the students were either likely to combine/distribute (C3/C4) the existing paragraphs to make the text more cohesive or started describing new topics in the text by adding new paragraphs (C5). This happened because C8 state had the 7% and 6% probabilities of transitioning to C3(46%)C4(54%) and C5 states, respectively. After combining and dividing paragraphs in C3(46%)C4(54%) state, the students were likely to come back and change the existing paragraphs because this state had the 50% probability of transitioning to C8 state. In addition, there were strong relations between C5 and C8 states (56% of transitioning probability) and C6 and C8 (48%). This indicated that the students were likely to modify text after adding and deleting text. These interpreted patterns were common writing activities of students of group H.

Comparing the Hidden MM of students of group L to that of group H revealed that students of group L, we could see some similar activity and some differences. The state transition behaviors between the two models were quite similar, although the models had different structures i.e. number of states. For the structure difference, the model of group L included surface change or editing (C1) activities, which never occurred in the model of group H. However, there are stronger differences that distinguish the two groups. First, there were more C7 activities in the writing process of L than in that of H. This suggests that the students of this group altered the whole text completely several times. Particularly, this occurred when the students changed the topics of the text. In addition, there was not obvious transition from the C5(65%)C6(35%) state to the C1(8%)C3(92%), unlike the model of group H which had transitions from both C5 and C6 states to C3 state. Importantly, there were very strong relationship between C1(8%)C3(92%) state and C8 state (85% transitioning probability) in group L. This indicates that after editing and formatting texts, students continued on changing the text very often.

C. Analysis of stationary probabilities

We incorporated the inactive activities (Pause), in our calculation of stationary probabilities. For example, if an activity A followed by 5 Pause consecutive activities, we would consider activity A to occur 6 times for this interval. The computed stationary probabilities are summarized in Table 3.

TABLE III, STATIONARY PROBABILITIES

Document	C1	C3	C4	C5	C6	C7	C8
H	-	4	5	9	5	0	77
L	1	5	-	10	8	5	71

From the table, it was obvious that group H and L's models were almost identical. Especially, the two groups spent lot of time changing their paragraphs. The main difference in term of the proportion of time between H and L groups was that students of group L spent 5% of their time changing topics, whereas students of group H had defined theirs early on. The models mentioned in previous section also discovered this.

VI. Discussion and Future Work

We presented our techniques for extracting the semantic meaning of writing activities and analyzing the writing processing. In particular, we derived the heuristic and hidden Markov models of the documents written by two groups of students who achieved a high and a low performance, respectively. The models represent the writing behaviors of these students. The heuristic MM offers a concise model giving a good bird-eye view of the overall writing process, where each state of the model represents a defined writing activity (such brainstorming, editing and so on). However the Hidden MM, by discovering the states automatically, offers a finer grained analysis by showing the sequences of text operations and transitional probabilities. Therefore, we were able to see why this high performance group was able to produce better quality outcome. In the future, we would like to analyze all the groups' writing processes in the class and extract the most common patterns in high performing groups, constrasting them with the ones common in low performing ones. We plan to provide these behavior models of writing activities as feedback, in the form of visualizations, to students and instructors while the students are writing their documents.

The algorithm that we used to construct the HMMs has a local maxima problem. In this study, we followed the work developed by [4] and executed the algorithm one hundred times with random initializations (by sampling the initial parameter values from uniform distributions). All of these executions converged to the same configuration. A better solution is needed to execute this algorithm in real time and provide feedback to students as they write.

We plan to work on incorporating the notion of time directly into the Markov model. The analysis in Subsection V.C indicates that the durations the two groups spent on individual writing activities reflect their writing quality outcomes. However, the analysis of stationary probabilities cannot truly model the pause time during writing. Therefore, it would be interesting to capture the duration of those activities in our Markov models. As mentioned in Section IV, during the construction of our models we coded inactive events (*Pause* activities) in the sequences of writing events as an extension of the previous event. This is an initial attempt to capture the duration of activities but is probably not very accurate in all cases. Another representation would have been to consider the pause as an activity or a text operation in itself, but that would also be inappropriate. Indeed, if we consider the *Pause* activity as one type of text change operations, we would end up with constructed models with one *Pause* state. These would have all other states having out-going arcs to the *Pause* state with some transition probabilities because all writing activities have some pauses in between and before transferring to other activities. Obviously, this does not convey useful information and adds more complexity to the analysis of the models. We believe that the inactive events (*Pauses*) should not be considered as a state, but as a "speed" of transition in the models of writing process. Because we want to investigate the total time an activity takes to complete, a new representation of HMM like [12] should be adapted and used in order to model writing processes. Our next stage is to work on this.

ACKNOWLEDGEMENT

This project has been funded by Australian Research Council DP0986873.

REFERENCE

[1] C. Boiarsky, "Model for analyzing revision," Journal of Advanced Composition, vol. 5, pp. 65-78, 1984.

[2] R. A. Calvo, S. T. O'Rourke, J. Jones, K. Yacef, and P. Reimann, "Collaborative writing support tools on the cloud," IEEE Transactions on Learning Technologies, 2010, in press.

[3] Google Docs, *http://docs.google.com*, 2010.

[4] H. Jeong, G. Biswas, J. Johnson, and L. Howard, "Analysis of productive learning behaviors in a structured inquiry cycle using hidden markov models," in International Conference on Educational Data Mining, Pittsburgh, PA, USA, 2010, pp. 81-89.

[5] B. Juang and L. Rabiner, "The segmental k-mean algorithm for estimating the parameters of hidden markov models," IEEE Trans Acoustics, Speech and Signal Processing, vol. 38, 1990, pp 1639-1641.

[6] C. Li and G. Biswas, "A bayesian approach for learning hidden markov models from data," Special issue on Markov Chain and Hidden Markov Models, Scientific Programming, vol. 10, pp. 201-219, 2002.

[7] P. B. Lowry, A. Curtis, and M. R. Lowry, "Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice," Journal of Business Communication, vol. 41, pp. 66-99, 2003.

[8] S. Osinski, J. Stefanowski, and D. Weiss, "Lingo: Search results clustering algorithm based on singular value decomposition," in Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference, Poland, 2004, pp. 359--368.

[9] L. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," Proceedings of the IEEE, vol. 77, pp. 257–286, 1989.

[10] V. Southavilay, K. Yacef, and R.A. Calvo, "WriteProc: A Framework for Exploring Collaborative Writing Processes," Australasian Document Computing Symposium, Sydney, Australia, 2009.

[11] V. Southavilay, K. Yacef, and R.A. Calvo, "Process Mining to Support Students' Collaborative Writing," International Conference on Educational Data Mining, Pittsburgh, PA, USA, 2010, pp. 257-266.

[12] S.-Z. Yu, "Hidden semi-markov models," Artificial Intelligence, vol. 174, pp. 215–243, 2010.