An Interactive Analytic Tool for Peer-Review Exploration

Wenting Xiong[1,2]**, Diane Litman**[1,2]**, Jingtao Wang**[1,2] **and Christian Schunn**[2]
[1] Department of Computer Science & [2] Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA, 15260
wex12@cs.pitt.edu

Abstract

This paper presents an interactive analytic tool for educational peer-review analysis. It employs data visualization at multiple levels of granularity, and provides automated analytic support using clustering and natural language processing. This tool helps instructors discover interesting patterns in writing performance that are reflected through peer reviews.

1 Introduction

Peer review is a widely used educational approach for coaching writing in many domains (Topping, 1998; Topping, 2009). Because of the large number of review comments to examine, instructors giving peer review assignments find it difficult to examine peer comments. While there are web-based peer-review systems that help instructors set up peer-review assignments, no prior work has been done to support instructors' comprehension of the textual review comments.

To address this issue, we have designed and developed an interactive analytic interface (RevExplore) on top of SWoRD[1] (Cho and Schunn, 2007), a web-based peer-review reciprocal system that has been used by over 12,000 students over the last 8 years. In this paper, we show how RevExplore visualizes peer-review information in multiple dimensions and various granularity levels to support investigative exploration, and applies natural language processing (NLP) techniques to facilitate review comprehension and comparison.

[1] https://sites.google.com/site/swordlrdc/

2 Design Goals

Instructors face challenges when they try to make sense of the peer-review data collected by SWoRD for their assignments. Instructors we have interviewed have complained that peer reviews are time-consuming to read and almost "impossible" to interpret: 1) to understand the pros and cons of one student's paper, they need to synthesize all the peer reviews received by that student by reading them one by one; 2) furthermore, if instructors would like to discover general patterns regarding students' writing performance, they have to additionally compare peer reviews across multiple students which requires their simultaneously remembering various opinions for many students; 3) in the initial stage of peer review analysis, instructors have no clear idea of what potential patterns they should be looking for ("cold start").

These challenges motivate our design of RevExplore, a peer-review analytic tool that is a plugin to SWoRD. We set our design goals to address the challenges mentioned above, respectively: 1) create a simple and informative representation of peer-review data which automatically aggregates peer-reviews at the level of student; 2) provide intelligent support of text mining and semantic abstraction for the purpose of comparison; 3) enable an overview of key characteristics of peer reviews for initial exploration.

To fulfill our design goals, we design an interactive visualization system to ease the exploration process, following the pattern of overview plus detail (Card et al., 1999). In the overview, RevExplore

provides a high level of visualization of overall peer-review information at the student level for initial exploration. In the detail-view, RevExplore automatically abstracts the semantic information of peer reviews at the topic-word level, with the original texts visible on demand. In addition, we introduce clustering and NLP techniques to support automated analytics.

3 Related Work

One major goal of peer review studies in educational research is to understand how to better improve student learning, directly or indirectly. Empirical studies of textual review comments based on manual coding have discovered that certain review features (e.g., whether the solution to a problem is explicitly stated in a comment) can predict both whether the problem will be understand and the feedback implemented (Nelson and Schunn, 2009). Our previous studies used machine learning and NLP techniques to automatically identify the presence of such useful features in review comments (Xiong et al., 2010); similar techniques have also been used to determine review comment helpfulness (Xiong and Litman, 2011; Cho, 2008). With respect to paper analysis, Sándor and Vorndran (2009) used NLP to highlight key sentences, in order to focus reviewer attention on important paper aspects. Finally, Giannoukos et al. (2010) focused on peer matching based on students' profile information to maximize learning outcomes, while Crespo Garcia and Pardo (2010) explored the use of document clustering to adaptively guide the assignment of papers to peers. In contrast to the prior work above, the research presented here is primarily motivated by the needs of instructors, instead of the needs of students. In particular, the goal of RevExplore is to utilize the information in peer reviews and papers, to help instructors better understand student performance in the peer-review assignments for their courses.

Many computer tools have already been developed to support peer review activities in various types of classrooms, from programming courses (Hyyrynen et al., 2010) to courses involving writing in the disciplines (Nelson and Schunn, 2009; Yang, 2011). Within the writing domain, systems such as SWoRD (Cho and Schunn, 2007) mainly assist instructors by providing administrative management support and/or (optional) automatic grading services. While peer review systems especially designed for instructors do exist, their goal is typically to create a collaborative environment for instructors to improve their professional skills (Fu and Hawkes, 2010). In terms of artificial intelligence support, to our knowledge no current peer review system has the power to provide instructors with insights about the semantic content of peer reviews, due to the diversity and complexity of the textual review comments. For example, SWoRD currently provides teachers a numerical summary view that includes the number of reviews received for each paper, and the mean and standard deviation of numerical reviewing ratings for each paper. SWoRD also allows instructors to automatically compute a grade based on a students' writing and reviewing quality; the grading algorithm uses the numerical ratings but not the associated text comments. In this work, we attempted to address the lack of semantic insight both by having humans in the loop to identify points of interest for interactive data exploration, and by adapting existing natural language processing techniques to the peer review domain to support automated analytics.

4 RevExplore

As an example for illustration, we will use data collected in a college level history class (Nelson and Schunn, 2009): the instructor created the writing assignment through SWoRD and provided a peer-review rubric which required students to assess a history paper's quality on three dimensions (logic, flow and insight) separately, by giving a numeric rating on a scale of 1-7 in addition to textual comments. While reviewing dimensions and associated guidelines (see below) are typically created by an instructor for a particular assignment, instructors can also set up their rubric using a library provided by SWoRD.

For instance, the instructor created the following guidance for commenting on the "logic" dimension: "*Provide specific comments about the logic of the author's argument. If points were just made without support, describe which ones they were. If the support provided doesn't make logical sense, explain what that is. If some obvious counter-argument was*

not considered, explain what that counter-argument is. Then give potential fixes to these problems if you can think of any. This might involve suggesting that the author change their argument."

Instructor guidance for numerically rating the logical arguments of the paper based on the comments was also given. For this history assignment, the highest rating of 7 ("Excellent") was described as "*All arguments strongly supported and no logical flaws in the arguments.*" The lowest rating of 1 ("Disastrous") was described as "*No support presented for any arguments, or obvious flaws in all arguments.*"

24 students submitted their papers online through SWoRD and then reviewed 6 peers' papers assigned to them in a "double blind" manner (review examples are available in Figure 2). When peer review is finished, RevExplore loads all papers and peer reviews, both textual comments and numeric ratings, and then goes through several text processing steps to prepare for interactive analytics. This preprocessing includes computing the domain words, sentence simplification, domain-word masking, syntactic analysis, and key noun-phrase extraction.

4.1 Overview – Student Clustering

RevExplore starts with a student-centric visualization overview. It uses a visual node of a bar chart to represent each student, visualizing the average of the student's peer ratings in gray, as well as the rating histogram with gradient colors (from red to blue) that are mapped to the rating scale from 1 to 7 (denoted by the legend in Figure 1).

To investigate students' writing performance, instructors can manually group similar nodes together into one stacked bar chart, or use automatic grouping options that RevExplore supports to inform initial hypotheses about peer review patterns. In the auto-mode, RevExplore can group students regarding a certain property (e.g. rating average); it can also cluster students using standard clustering algorithms[2] based on either rating statistics or Bag-Of-Words extracted from the relevant peer reviews.

If a instructor is curious about the review content for certain students during exploration, the instructor can read the main issues, in the form of noun phrases (NPs) of a student's peer reviews in a tooltip by mouse hovering on the bar squares which the student corresponds to. For example, Figure 1 shows that the peer reviews received by this student are mainly focused on the argumentation and the introduction part of the paper.

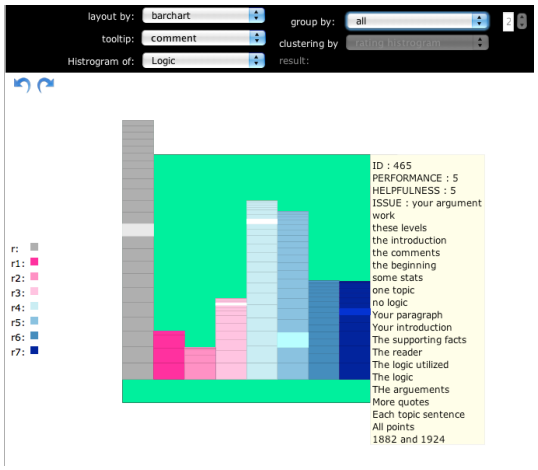

Figure 1: RevExplore overview. Stacked bar charts represent student groups. The tooltip shows the ID of the current student, writing *performance* (average peer ratings), review *helpfulness* (average helpfulness ratings), as well as the main *issues* in the descending order of their frequency, which are extracted from the peer reviews received by a highlighted student using NLP techniques.

To extract peer-review main issues, RevExplore syntactically simplifies each review sentence (Heilman and Smith, 2010), parses each simplified sentence using the Stanford dependency parser (de Marneffe et al., 2006), and then traverses each dependency tree to find the key NP in a rule-based manner.[3] Due to reviewers' frequent references to the relevant paper, most of the learned NPs are domain related facts used in the paper, rather than evaluative texts that suggest problems or suggestions. To avoid the interference of the domain content, we apply domain-word masking (explained in Section 4.2) to the simplified sentences before parsing, and eliminate any key NP that contains the mask.

4.2 Detail-View – Topic Comparison

When two groups of students are selected in the overview, their textual peer reviews can be further

[2]RevExplore implements both K-Means and a hierarchical clustering algorithm.

[3]Rules are constructed purely based on our intuition.

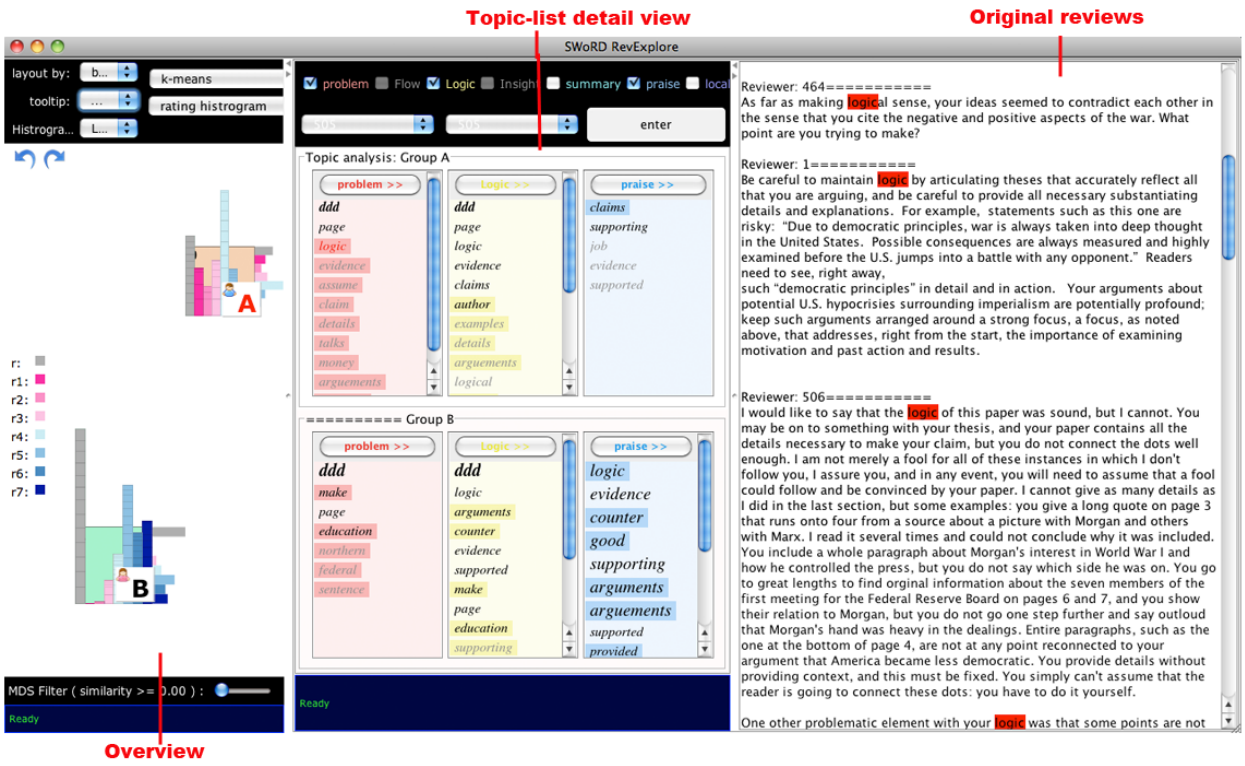

Figure 2: Peer-review exploration using RevExplore, for mining differences between strong and weak students.

compared with respect to specific reviewing dimensions using a list of topic words that are automatically computed in real-time.

Extracting topic words of peer reviews for comparison purposes is different from most traditional topic-word extraction tasks that are commonly involved in text summarization. In traditional text summarization, the informativeness measurement is designed to extract the common themes, while in our case of comparison, instructors are more concerned with the uniqueness of each target set of peer reviews compared to the others. Thus a topic-signature acquisition algorithm (Lin and Hovy, 2000), which extracts topic words through comparing the vocabulary distribution of a target corpus against that of a generic background corpus using a statistic metric, suits our application better than other approaches, such as probabilistic graphical models (e.g. LDA) and frequency based methods. Therefore, RevExplore considers topic signatures as the topic words for a group of reviews, using all peer reviews as the background corpus.[4] Again, to minimize the impact of the domain content of the relevant papers, we apply topic-masking which replaces all domain words[5] with "ddd" before computing the topic signatures.

As the software outputs topic signatures together with their associated weights which reflect signature importance, RevExplore uses this weight information to order the topic words as a list, and visualizes the weight as the font size and foreground color of the relevant topic word. These lists are placed in two rows regarding their group membership dimension by dimension. For each dimension, the corresponding lists of both rows are aligned vertically with the same background color to indicate that dimension (e.g. Topic-list detail view of Figure 2). To further facilitate the comparison within a dimension, RevExplore highlights the topic words that are unique to one group with a darker background color.

[4]We use TopicS (Nenkova and Louis, 2008) provided by Annie Louis.

[5]learned from all student papers against 5000 documents from the English Gigaword Corpus using TopicS.

If the user cannot interpret the topic that an extracted word might imply, the user can click on the word to read the relevant original reviews, with that word highlighted in red (e.g. Original reviews pane of Figure 2).

5 Analysis Example

Figure 2 shows how RevExplore is used to discover the difference between *strong* and *weak* students with respect to their writing performance on "logic" in the history peer-review assignment introduced in Section 4.

First we group students into strong versus weak regarding their writing performance on logic by selecting the K-Means algorithm to cluster students into two groups based on their rating histogram on logic. As shown in the Overview pane of Figure 2, we then label them as A and B for further topic comparison.

Next, in the topic-list detail view, we check "praise" and "problem"[6], and fire the "enter" button to start extracting topic words for group A and B on every selected dimension. Note that "logic" will be automatically selected since the focus has already been narrowed down to logic in the overview.

To first compare the difference in general logic issues between these two groups, we refer to the two lists on "logic" (in the middle of the topic-list detail view, Figure 2). As we can see, the weak students' reviews (Group A) are more about the logic of statements and the usage of facts (indicated by the unique words "examples" and "details"); the strong students' peer reviews (group B) focus more on argumentation (noted by "counter" and "supporting").

To further compare the two groups regarding different review sentiment, we look at the lists corresponding to "problem" and "praise" (left and right columns). For instance, we can see that strong students' suffer more from context specific problems, which is indicated by the bigger font size of the domain-word mask. Meanwhile, to understand what a topic word implies, say, "logic" in group A's topic list on "problem", we can click the word to bring out the relevant peer reviews, in which all occurrences of "logic" are colored in red (original reviews pane in Figure 2).

[6]Although "praise" and "problem" are manually annotated in this corpus (Nelson and Schunn, 2009), Xiong et al. (2010) have shown that they can be automatically learned in a data-driven fashion.

6 Ongoing Evaluation

We are currently evaluating our work along two dimensions. First, we are interested in examining the utility of RevExplore for instructors. After receiving positive feedback from several instructors at the University of Pittsburgh, as an informal pilot study, we deployed RevExplore for some of these instructors during the Spring 2012 semester and let them explore the peer reviews of their own ongoing classes. Instructors did observe interesting patterns using this tool after a short time of exploration (within two or three passes from the overview to the topic-word detail view). In addition, we are conducting a formal user study of 40 subjects to validate the topic-word extraction component for comparing reviews in groups. Our preliminary result shows that our use of topic signatures is significantly better than a frequency-based baseline.

7 Summary and Future work

RevExplore demonstrates the usage of data visualization in combination with NLP techniques to help instructors interactively make sense of peer review data, which was almost impracticable before. In the future we plan to further analyze the data collected in our formal user study, to validate the helpfulness of our proposed topic-word approach for making sense of large quantities of peer reviews. We also plan to incorporate NLP information beyond the word and NP level, to support additional types of review comparisons. In addition, we plan to summarize the interview data that we informally collected from several instructors, and will mine the log files of their interactions with RevExplore to understand how the tool would (and should) be used by instructors in general. Last but not least, we will continue revising our design of RevExplore based on instructor feedback, and plan to conduct a more formal evaluation with instructors.

Acknowledgments

Thanks to Melissa Patchan for providing the history peer-review corpus. We are also grateful to LRDC for financial support.

References

Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. San Francisco, CA, USA.

Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, 48(3):409–426.

Kwangsu Cho. 2008. Machine classification of peer comments in physics. In *Proceedings First International Conference on Educational Data Mining*, pages 192–196.

Raquel M Crespo Garcia and Abelardo Pardo. 2010. A supporting system for adaptive peer review based on learners' profiles. In *Proceedings of Computer Supported Peer Review in Education Workshop*, pages 22–31.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Hongxia Fu and Mark Hawkes. 2010. Technology-supported peer assessment as a means for teacher learning. In *Proceedings of the 2010 Workshop on Computer Supported Peer Review in Education*.

Ioannis Giannoukos, Ioanna Lykourentzou, Giorgos Mpardis, Vassilis Nikolopoulos, Vassilis Loumos, and Eleftherios Kayafas. 2010. An adaptive mechanism for author-reviewer matching in online peer assessment. In *Semantics in Adaptive and Personalized Services*, pages 109–126.

Michael Heilman and Noah A. Smith. 2010. Extracting simplified statements for factual question generation. In *Proceedings of the 3rd Workshop on Question Generation*.

Ville Hyyrynen, Harri Hämäläinen, Jouni Ikonen, and Jari Porras. 2010. Mypeerreview: an online peer-reviewing system for programming courses. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research*, pages 94–99.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings COLING*.

Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37:375–401.

Ani Nenkova and Annie Louis. 2008. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of Association for Computational Linguistics*.

Ágnes Sándor and Angela Vorndran. 2009. Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 36–44, Suntec City, Singapore, August. Association for Computational Linguistics.

Keith Topping. 1998. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276.

Keith J. Topping. 2009. Peer assessment. *Theory Into Practice*, 48(1):20–27.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 502–507.

Wenting Xiong, Diane J. Litman, and Christian D. Schunn. 2010. Assessing reviewers performance based on mining problem localization in peer-review data. In *Proceedings Third International Conference on Educational Data Mining*.

Yu-Fen Yang. 2011. A reciprocal peer review system to support college students' writing. *British Journal of Educational Technology*, 42(4):687–700.