Determining Review Coverage by Extracting Topic Sentences Using A Graph-based Clustering Approach

Lakshmi Ramachandran
North Carolina State University
lramach@ncsu.edu

Balaraman Ravindran
Indian Institute of Technology, Madras
ravi@cse.iitm.ac.in

Edward F. Gehringer
North Carolina State University
efg@ncsu.edu

ABSTRACT

Reviews of technical articles or documents must be thorough in discussing their content. At times a review may be based on just one section in a document, say the *Introduction*. Review coverage is the extent to which a review covers the "important topics" in a document. In this paper we present an approach to evaluate the coverage of a submission by a review. We use a novel agglomerative clustering technique to group the submission's sentences into topic clusters. We identify topic sentences from these clusters, and calculate review coverage in terms of the overlaps between the review and the submission's topic sentences. We evaluate our coverage identification approach on peer-review data from Expertiza, a collaborative, web-based learning application. Our approach produces a high correlation of 0.51 with human-provided coverage values.

Keywords

review quality, review coverage, topic identification, agglomerative clustering, lexico-semantic matching

1. INTRODUCTION

The past few years have witnessed a growth in Massive Open Online Courses (MOOCs) such as Coursera and Udacity, as a platform for web-based collaborative learning. MOOCs require a scalable means of assessment, and for material that cannot be assessed by multiple-choice tests, peer-review fills the bill. Feedback in the form of text-based reviews help authors identify mistakes in their work, and learn possible ways of improving them. Since reviews play a crucial role in helping authors, it is important to ensure that they are *complete*, and their content is *useful* to authors. At times reviews may cover just one section in the author's submission (the text under review), say the *Introduction*, and provide no feedback on any of the other sections in the document.

Kuhne et al. [1] found that authors are content with reviewers who have made an effort to read and understand the author's work. Reviews that cover the important sections of the author's work are likely to be more useful, since they are more complete than reviews discussing a single section. A complete review also reflects positively on a reviewer's understanding of the author's work.

Existing review assessment approaches use shallow text features such as word count to analyze their usefulness. Xiong et al. use a bag-of-words exact match approach to identify instances of problems (in the author's work) caught by peer-reviews [2]. Cho uses machine classification techniques such as naïve Bayes, SVM (support vector machines) and decision trees to classify feedback [3]. At present none of the automatic review analysis approaches look for the degree of coverage of a submission by a review. One of the chief contributions of this paper is the focus on the important but often ignored problem of identifying review coverage.

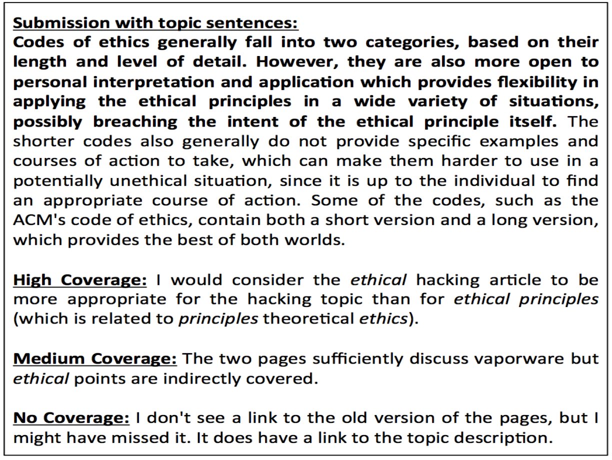
2. APPROACH

We employ an agglomerative clustering technique to group submission sentences into clusters or topics, and then identify the most representative sentences from across the different clusters. Cluster-based approaches have been widely applied to text and other knowledge mining applications. Steinbach et al. use bisecting k-means to cluster documents [7]. Qazvinian et al. [5] use a cluster-based approach to determine the sentences central to a document. They use a hierarchical agglomeration algorithm to cluster sentences. The ClusterRank algorithm, proposed by Garg et al. [6], applies a clustering technique to identify sentences belonging to the same topic.

Sentences discussing the same topic, but containing different terms may not be effectively grouped by a clustering approach relying purely on the frequency of words. Steinbach et al. found that agglomerative clustering with a word-frequency based matching made mistakes by grouping nearest documents belonging to different classes into the same cluster [7].

We employ a lexico-semantic matching technique, which captures context information. We use a word order graph to represent text, since it captures syntax or order of tokens in a text. Word order graphs are suited for identifying lexical and voice changes, which are common among paraphrased text. Similarity should capture the degree of relatedness between texts. Hence we use a WordNet-based metric [8]. Topic-representative sentences are selected from the most significant clusters. A review is compared with topic sentences to identify coverage.

Figure 1: Submission with its topic sentences, and three reviews with high, medium and no coverage of the topic sentences



Submission with topic sentences:
Codes of ethics generally fall into two categories, based on their length and level of detail. However, they are also more open to personal interpretation and application which provides flexibility in applying the ethical principles in a wide variety of situations, possibly breaching the intent of the ethical principle itself. The shorter codes also generally do not provide specific examples and courses of action to take, which can make them harder to use in a potentially unethical situation, since it is up to the individual to find an appropriate course of action. Some of the codes, such as the ACM's code of ethics, contain both a short version and a long version, which provides the best of both worlds.

High Coverage: I would consider the *ethical* hacking article to be more appropriate for the hacking topic than for *ethical principles* (which is related to *principles* theoretical *ethics*).

Medium Coverage: The two pages sufficiently discuss vaporware but *ethical* points are indirectly covered.

No Coverage: I don't see a link to the old version of the pages, but I might have missed it. It does have a link to the topic description.

Table 1: Correlation between system-generated and human-provided coverage values.

Approach	correlation	Avg. # words
Our system	**0.51**	108
MEAD summarizer	0.46	100

In order to illustrate our approach we use real-world submission and review data from assignments completed using Expertiza [9]. Expertiza is a collaborative, web-based learning application that helps students submit assignments and peer review each other's work. Figure 1 contains a sample submission with its topic-representative sentences highlighted in bold, and three sample reviews with high, medium and no coverage of the submission's topic sentences. The first review *covers* the submission because it mentions *ethical principles* and *ethics*. However, the review with *medium* coverage mentions just *ethics*, and the review with *no* coverage does not contain any relevant information.

3. EXPERIMENT

In this section we study the usefulness of our approach in determining a review's coverage. We compare our approach with MEAD, a centroid-based summarization approach [4]. Radev et al.'s approach uses the most common words in a document to identify the best sentences to be included in a summary. MEAD is an extractive summarization approach, and since in our approach too we extract the most representative sentences from a submission, we find MEAD to be an ideal system to compare our approach with.

We use peer-review data from computer science classes over a couple of semesters to evaluate our approach. A dataset containing 577 reviews and submissions was selected from Expertiza [9]. Review data is annotated on a scale of 0-5, where 0 indicates no coverage and 5 indicates maximum coverage.

We identify the Spearman correlation between system-generated and human-provided coverage values (Table 1). Our approach produces a correlation of 0.51, while MEAD's coverage produces a correlation of 0.46. A positive correlation of 0.51 indicates that the system has a good degree of agreement with human-provided coverage values.

Topic sentences generated by our approach have almost the same number of words as those generated by MEAD. However, our approach produces higher correlations with human ratings than the output from the MEAD summarizer. Thus, our approach is able to effectively identify topic-representative sentences from a document, and estimate a review's coverage of these topic sentences.

4. CONCLUSION

Assessment of reviews is an important problem in the fields of education, science and human resources, and so it is worthy of serious attention. Our aim is to utilize natural language processing techniques to determine review coverage. Since reviews are central to the process of assessment it is important to ensure that they cover the main points of a submission. In this paper we have explained our approach to solving the problem of automatically determining a review's coverage of a submission. Ours is a pioneering effort in applying clustering and topic identification techniques to calculate review coverage. We have also shown that our review coverage approach produces a good correlation of 0.51 with human-provided coverage values.

5. REFERENCES

[1] C. Kuhne, K. Bohm, and J. Z. Yue, "Reviewing the reviewers: A study of author perception on peer reviews in computer science," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on.* IEEE, 2010, pp. 1–8.

[2] W. Xiong, D. J. Litman, and C. D. Schunn, "Assessing reviewer's performance based on mining problem localization in peer-review data." in *EDM*, 2010, pp. 211–220.

[3] K. Cho, "Machine classification of peer comments in physics," in *EDM*, 2008, pp. 192–196.

[4] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, no. 6, pp. 919–938, Nov. 2004.

[5] V. Qazvinian and D. R. Radev, "Scientific paper summarization using citation summary networks," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING, 2008, pp. 689–696.

[6] N. Garg, B. Favre, K. Riedhammer, and D. Hakkani-Tür, "Clusterrank: a graph based method for meeting summarization," in *INTERSPEECH*, 2009, pp. 1499–1502.

[7] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, 2000, pp. 525–526.

[8] C. Fellbaum, "Wordnet: An electronic lexical database." *MIT Press*, p. 423, 1998.

[9] E. F. Gehringer, "Expertiza: Managing feedback in collaborative learning." in *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*, 2010, pp. 75–96.