Exploring Document Clustering Techniques for Personalized Peer Assessment in Exploratory Courses

Raquel M. Crespo García[1],

[1] Universidad Carlos III de Madrid, Avda. de la Universidad 30, E-28911 Leganés (Madrid),
Spain
raquel.crespo@uc3m.es

Abstract. Peer review has been proposed as a complement to project-based learning in courses covering a wide and heterogeneous syllabus. By reviewing peers' projects, students can explore other subjects thoroughly apart from their own project topic. This objective relies however in a proper distribution of the works to review, which is a complex and time-consuming task. Beyond simple topic selection, students may report different types of works, which influence their peers' assessment; for example, works focused on a project development approach versus in-depth literature researches. Introducing detailed metadata is time-consuming (thus users are typically reluctant) and, even more important, prone to error. In this paper we explore the potential of text mining and natural language processing technologies for automatic classification of texts, in order to facilitate the adaptation and diversification of the works assigned to the students for review, in the context of a course on Artificial Intelligence.

Keywords: adaptive peer review, document clustering, text mining.

1 Introduction and background

Project-based learning is a typical approach in exploratory learning, providing in depth learning of the project topic. Peer assessment, defined in the educational context as "*an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status*" [1], has been extensively applied in a variety of educational settings too. Numerous systems, mostly web-based, have been developed for supporting the peer review process, such as CPR [2], PG [3, 4] and its evolution Expertiza [5], OPAS [6] or OASIS [7].

Despite its undeniable advantages, project-based learning poses a major drawback in courses with a wide and heterogeneous syllabus, as students focus on one topic but only have a superficial look at the rest of the topics of the course. Intelligence in Computer Networks [8], an introductory course on Artificial Intelligence (AI) for senior Telecommunication Engineering students, is one of such courses.

Peer review has been proposed as a complement for project-based learning in order to allow students to explore in depth other topics besides the one developed in their own projects [9]. This approach however requires selecting the reviewers for a given submission depending on the project topics of both the submission and the reviewers. I.e. adapting the peer reviewer process according to student's profile [10], where the student's profile models information on his/her project topic(s).

Normally, topic classification is done relying on metadata provided by the student. Errors may however appear in this manual categorization. Even if using title or keywords, or even abstract, errors may happen, as students reports sometimes include thorough information on fundamentals or related topics rather than the actual objective promised. Using a content-based approach should achieve more accurate results.

In this paper, document clustering is explored for automatically characterizing students' submissions. A sample of 27 documents has been analyzed, corresponding to the final reports submitted by the students during the 2008-2009 course (see **Table 1** for project titles, originally in Spanish). Documents are expected to be grouped in clusters according to their topics. Adaptive matching of reviewers could then be applied based on such clustering-based profiles.

Table 1. Project titles

1	Optimal engineering degree choose
2	Data Mining applications for premature Cancer diganosis
3	Google Labs: Flu Trends & Google Trends
4	Semantic Web
5	Knowledge representation: Semantic Web
6	Learning algorithms: KNN & KMeans
7	Expert system: animal classification according to their taxonomy
8	Data mining applied to social networks
9	Optical Character Recognition (OCR)
10	Fuzzy Logic
11	Data mining application for criminal pattern exploration and detection
12	Artificial Intelligence techniques applied to educational settings based on games
13	Prediction of El Niño cycles
14	Data mining applied to breast cáncer detection
15	Pattern search using genetic algorithms
16	Implementation of the Chinchón game
17	Text recognition
18	Implementation of the puzzle-n game using two algorithms: Backtracking and A*
19	Artificial vision in intelligent houses
20	Prediction of niche words using Weka
21	SuperManager ACB: Constraints satisfaction using Backtracking
23	Data mining applications to the telecommunications industry
24	The intelligent home
25	Choose your own adventure
26	Expressive audio and artificial intelligence: SaxEx and JIG
27	The Escoba game with a player controlled by the computer
28	Expert system: KingsHelper

As explained in [11] "during the last decade text mining has become a widely used discipline utilizing statistical and machine learning methods. […] Classical applications in text mining come from the data mining community, like document clustering and document classification. For both the idea is to transform the text into a structured format based on term frequencies and subsequently apply standard data mining techniques".

In this paper, the *tm* framework [11,12] is used for clustering the students' reports. tm is a text mining package for R [13], an open source statistical computing environment. Additional R packages have also been used for preprocessing the text, such as Rstem [14] and Snowball [15] for stemming the words in the document.

2 Results

The document clustering process applied to the reports submitted by the students is schematized in **Fig. 1**.

Fig. 1. Process

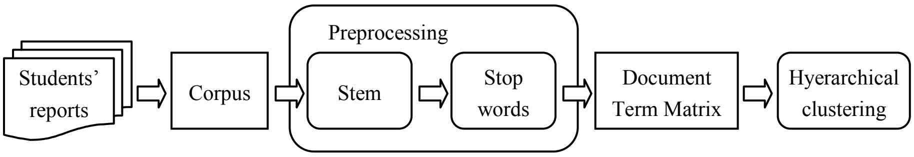
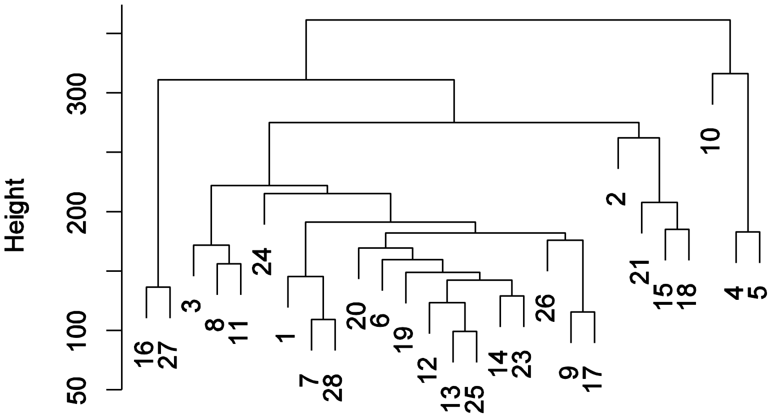


Fig. 2 shows the results of the hierarchical clustering. Correspondence between document codes and titles is reported in **Table 1**.

Fig. 2. Cluster Dendrogram



The cluster dendrogram shows some interesting and promising results. The only projects about Semantic Web (4 and 5) are closely grouped in the same basic cluster. Projects on expert systems (1, 7 and 28) also form their own cluster. Projects implementing an intelligent player for a game (16, 27) are also clustered together. Text recognition projects (9 & 17) are grouped together, and with the other about speech recognition (26), which is obviously related. Reports related to data mining applications appear in the central cluster. On the other hand, the last document integrated into the clusters structure (10) is the only one about Fuzzy Logic, which is indeed loosely related to the rest of the projects topics (except semantic web projects where it is also mentioned).

Although further analysis is required (in particular, improving the preprocessing steps, mainly word filtering or stemming), document clustering provides promising results for content-based automatic categorization of students' reports.

The categorization information can then be used for building the student profile and adapting the peer review process accordingly. Different strategies may apply depending on the educational objectives pursued.

In this particular course, one core objective of the peer review process is providing the students with the opportunity to explore complementary topics, different from the one their own work is focused on, in order to widen their view of the subject. In order to accomplish this goal, the reviewers should be assigned works classified into clusters different than the one their own submission belongs to.

Other educational settings may nevertheless demand different approaches. For example, in the experience discussed in [16], the usefulness of the peers' feedback is prioritized. In order to maximize quality and interest of such comments, the more familiar the reviewer is with the assessed topic, the better. Thus, reviewers should be assigned submissions belonging to the same cluster than their own in this case, in order to guarantee their expertise on the reviewed matter.

Any of these strategies can be easily applied, once built the submissions and learners' profiles with the help of the report clustering, using the Adaptive Peer Review methodology and supporting system discussed in [17].

3 Conclusions

In this paper, document clustering has been explored for automatic classification of student reports in a course on artificial intelligence with promising results. Previous work had reported the convenience of applying adaptive peer review, taking into account project topics, in exploratory courses like this. Topic diversification introduced by adaptive peer review gives the students the opportunity to widen their view of the subject, previously restricted to their own project topic. Automatic clustering of the projects can facilitate building the students and submissions profiles on which such adaptation is based on.

Future work will require a more rigorous evaluation as well as widening the documental corpus. Students' reports from previous years are available and could be used for reviewing and improving these results. As explained before, improving the preprocessing steps should lead to more accurate results; alternative clustering techniques are also to be explored.

Acknowledgments This work was partially funded by Best Practice Network ICOPER (Grant No. ECP-2007-EDU-417007), Learn3 project, "Plan Nacional de I+D+I" TIN2008-05163/TSI, and eMadrid network, S2009/TIC-1650, "Investigación y Desarrollo de tecnologías para el e-learning en la Comunidad de Madrid".

References

1. Topping, K.: Peer assessment between students in colleges and universities. *Review of Educational Research* 68 (1998) 249–276
2. CPR: Calibrated peer review. [online] cpr molsci.ucla.edu (2004)
3. Gehringer, E.F.: Strategies and mechanisms for electronic peer review. *In: Frontiers in Education Conference,* ASEE/IEEE (2000)
4. Gehringer, E.F.: Electronic peer review and peer grading in computer-science courses. *In: Proc. of the Technical Symposium on Computer Science Education,* SIGCSE (2001) 139–143
5. Edward Gehringer, Luke Ehresman, Susan G. Conger, and Prasad Wagle: Reusable learning objects through peer review: The Expertiza approach, *Innovate: Journal of Online Education* 3:5, June/July 2007
6. Trahasch, S.: From peer assessment towards collaborative learning. *In: 34th ASEE/IEEE Frontiers in Education Conference*, Savannah, GA (2004)
7. Ward, A., Sitthiworachart, J., Joy, M.: Aspects of web-based peer assessment systems for teaching and learning computer programming. *In: IASTED International Conference on Web-based Education.* (2004) 292–297
8. Villena Román, J.: *"Inteligencia en Redes de Ordenadores"* (Web page of the course). Available: www.it.uc3m.es/jvillena/irc/indice html
9. Raquel M. Crespo García, Julio Villena Román, and Abelardo Pardo: Peer review to improve artificial intelligence teaching. *In: Frontiers in Education Conference.* ASEE/IEEE (October 2006).
10. Crespo García, R.M., Pardo, A., Delgado Kloos, C.: An adaptive strategy for peer review. *In: Frontiers in Education Conference*, ASEE/IEEE (2004)
11. Ingo Feinerer, Kurt Hornik and David Meyer: Text Mining Infrastructure in R. *In: Journal of Statistical Software,* Vol. 25, Issue 5, March 2008.
12. Ingo Feinerer: *tm: Text Mining Package*. R package version 0.5-3, URL http://tm r-forge.r-project.org/ (2010)
13. The Comprehensive R archive, URL cran r-project.org/
14. Temple Lang D.: Word Stemming in R Rstem R package version 0.3-1, URL http://www.omegahat.org/Rstem/ (2004)
15. Hornik K.: Snowball: Snowball Stemmers. R package version 0.0-7. URL http://cran r-project.org/web/packages/Snowball/index.html (2007)
16. Ioannis Giannoukos, Ioanna Lykourentzou, Giorgos Mpardis, Vassilis Nikolopoulos, Vassili Loumos, and Eleftherios Kayafas: An Adaptive Mechanism for Author-Reviewer Matching in Online Peer Assessment. *In: M. Wallace et al. (eds.): Semantics in Adaptive and Personalized Services, SCI 279, pp. 109–126*, Springer-Verlag Berlin Heidelberg (2010).
17. Raquel M. Crespo García, and Abelardo Pardo: A Supporting System for Adaptive Peer Review based on Learner's Profiles. *In: Workshop on Computer Supported Peer Review in Education*, Portsmouth (June 2010).