

Enriching Education through Data Mining

Rakesh Agrawal, Sreenivas Gollapudi,
Anitha Kannan, and Krishnaram Kenthapadi

Search Labs, Microsoft Research
Mountain View, CA, USA
{rakesha,sreenig,ankannan,krisken}@microsoft.com

Education is acknowledged to be the primary vehicle for improving the economic well-being of people [1,6]. Textbooks have a direct bearing on the quality of education imparted to the students as they are the primary conduits for delivering content knowledge [9]. They are also indispensable for fostering teacher learning and constitute a key component of the ongoing professional development of the teachers [5,8].

Many textbooks, particularly from emerging countries, lack clear and adequate coverage of important concepts [7]. In this talk, we present our early explorations into developing a data mining based approach for enhancing the quality of textbooks. We discuss techniques for algorithmically augmenting different sections of a book with links to selective content mined from the Web. For finding authoritative articles, we first identify the set of key concept phrases contained in a section. Using these phrases, we find web (Wikipedia) articles that represent the central concepts presented in the section and augment the section with links to them [4]. We also describe a framework for finding images that are most relevant to a section of the textbook, while respecting global relevancy to the entire chapter to which the section belongs. We pose this problem of matching images to sections in a textbook chapter as an optimization problem and present an efficient algorithm for solving it [2].

We also present a diagnostic tool for identifying those sections of a book that are not well-written and hence should be candidates for enrichment. We propose a probabilistic decision model for this purpose, which is based on syntactic complexity of the writing and the newly introduced notion of the dispersion of key concepts mentioned in the section. The model is learned using a tune set which is automatically generated in a novel way. This procedure maps sampled text book sections to the closest versions of Wikipedia articles having similar content and uses the maturity of those versions to assign need-for-enrichment labels. The maturity of a version is computed by considering the revision history of the corresponding Wikipedia article and convolving the changes in size with a smoothing filter [3].

We also provide the results of applying the proposed techniques to a corpus of widely-used, high school textbooks published by the National Council of

Educational Research and Training (NCERT), India. We consider books from grades IX–XII, covering four broad subject areas, namely, Sciences, Social Sciences, Commerce, and Mathematics. The preliminary results are encouraging and indicate that developing technological approaches to enhancing the quality of textbooks could be a promising direction for research for our field.

References

1. Knowledge for Development: World Development Report 1998/99. World Bank (1998)
2. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Enriching textbooks with web images (working paper, 2011)
3. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Identifying enrichment candidates in textbooks. In: WWW (2011)
4. Agrawal, R., Gollapudi, S., Kenthapadi, K., Srivastava, N., Velu, R.: Enriching textbooks through data mining. In: First Annual ACM Symposium on Computing for Development, ACM DEV (2010)
5. Gillies, J., Quijada, J.: Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. In: USAID Educational Quality Improvement Program (EQUIP2) (2008)
6. Hanushek, E.A., Woessmann, L.: The role of education quality for economic growth. Policy Research Department Working Paper 4122, World Bank (2007)
7. Mohammad, R., Kumari, R.: Effective use of textbooks: A neglected aspect of education in Pakistan. *Journal of Education for International Development* 3(1) (2007)
8. Oakes, J., Saunders, M.: Education's most basic tools: Access to textbooks and instructional materials in California's public schools. *Teachers College Record* 106(10) (2004)
9. Stein, M., Stuen, C., Carnine, D., Long, R.M.: Textbook evaluation and adoption. *Reading & Writing Quarterly* 17(1) (2001)