Electronic Textbooks and Data Mining

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

Search Labs, Microsoft Research,
Mountain View, CA, USA

Abstract. Education is known to be the key determinant of economic growth and prosperity [8,12]. While the issues in devising a high-quality educational system are multi-faceted and complex, textbooks are acknowledged to be the educational input most consistently associated with gains in student learning [11]. They are the primary conduits for delivering content knowledge to the students and the teachers base their lesson plans primarily on the material given in textbooks [7].

With the emergence of abundant online content, cloud computing, and electronic reading devices, textbooks are poised for transformative changes. Notwithstanding understandable misgivings (e.g. Gutenberg Elegies [6]), textbooks cannot escape what Walter Ong calls 'the technologizing of the word' [9]. The electronic format comes naturally to the current generation of 'digital natives' [10]. Inspired by the emergence of this new medium for "printing" and "distributing" textbooks, we present our early explorations into developing a data mining based approach for enhancing the quality of electronic textbooks. Specifically, we first describe a diagnostic tool for authors and educators to algorithmically identify deficiencies in textbooks. We then discuss techniques for algorithmically augmenting different sections of a book with links to selective content mined from the Web.

Our tool for diagnosing deficiencies consists of two components. Abstracting from the education literature, we identify the following properties of good textbooks: (1) *Focus* : Each section explains few concepts, (2) *Unity*: For every concept, there is a unique section that best explains the concept, and (3) *Sequentiality*: Concepts are discussed in a sequential fashion so that a concept is explained prior to occurrences of this concept or any related concept. Further, the tie for precedence in presentation between two mutually related concepts is broken in favor of the more significant of the two. The first component provides an assessment of the extent to which these properties are followed in a textbook and quantifies the comprehension load that a textbook imposes on the reader due to non-sequential presentation of concepts [1,2]. The second component identifies sections that are not written well and can benefit from further exposition. We propose a probabilistic decision model for this purpose, which is based on the syntactic complexity of writing and the notion of the dispersion of key concepts mentioned in the section [4].

For augmenting a section of a textbook, we first identify the set of key concept phrases contained in a section. Using these phrases, we find web articles that represent the central concepts presented in the section and endow the section with links to them [5]. We also describe techniques for finding images that are most relevant to a section of the textbook, while respecting the constraint that the same image is not repeated in different sections of the same chapter. 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 [3].

H. Gao et al. (Eds.): WAIM 2012, LNCS 7418, pp. 19–20, 2012.

We finally 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 embellishing textbooks could be a promising direction for research.

References

1. Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., Kenthapadi, K.: Empowering authors to diagnose comprehension burden in textbooks. In: KDD (2012)
2. Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., Kenthapadi, K.: Quality of textbooks: An empirical study. In: ACM DEV (2012)
3. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Enriching textbooks with images. In: CIKM (2011)
4. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Identifying enrichment candidates in textbooks. In: WWW (2011)
5. Agrawal, R., Gollapudi, S., Kenthapadi, K., Srivastava, N., Velu, R.: Enriching textbooks through data mining. In: ACM DEV (2010)
6. Birkerts, S.: The Gutenberg Elegies: The Fate of Reading in an Electronic Age. Faber & Faber (2006)
7. Gillies, J., Quijada, J.: Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. USAID Educational Quality Improvement Program, EQUIP2 (2008)
8. Hanushek, E.A., Woessmann, L.: The role of education quality for economic growth. Policy Research Department Working Paper 4122. World Bank (2007)
9. Ong, W.J.: Orality & Literacy: The Technologizing of the Word. Methuen (1982)
10. Prensky, M.: Digital natives, digital immigrants. On the Horizon 9(5) (2001)
11. Verspoor, A., Wu, K.B.: Textbooks and educational development. Technical report. World Bank (1990)
12. World-Bank. Knowledge for Development: World Development Report: 1998/99. Oxford University Press (1999)