# Reimagining Textbooks Through the Data Lens

Rakesh Agrawal
Search Labs, Microsoft Research

## Abstract

Textbooks are the primary vehicles for delivering subject knowledge to the students and are known to be the educational input most consistently associated with improvements in student learning. With the emergence of abundant online content, cloud computing, and electronic reading devices, textbooks are poised for transformative changes. Inspired by the emergence of the electronic 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]. 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 [3].

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 [4]. 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 [2].

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, 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] R. Agrawal, S. Chakraborty, S. Gollapudi, A. Kannan, and K. Kenthapadi. Empowering authors to diagnose comprehension burden in textbooks. In KDD, 2012.
[2] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Enriching textbooks with images. In CIKM, 2011.
[3] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Identifying enrichment candidates in textbooks. In WWW, 2011.
[4] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu. Enriching textbooks through data mining. In ACM DEV, 2010.

### Biography

Dr. Rakesh Agrawal is a Microsoft Technical Fellow, heading the Search Labs in Microsoft Research in Silicon Valley. He is a Member of the National Academy of Engineering, a Fellow of ACM, and a Fellow of IEEE. He is the recipient of the ACM-SIGKDD First Innovation Award, ACM-SIGMOD Innovations Award, IIT-Roorkee Distinguished Alumni Award, ACM-SIGMOD Test of Time Award, VLDB 10-Yr Most Influential Paper Award, and ICDE Most Influential Paper Award. Scientific American named him to the list of 50 top scientists and technologists in 2003. Dr. Agrawal has been granted more than 60 patents and has published more than 150 research papers. He has written the first and second highest cited papers in the fields of databases and data mining. Before Microsoft, he worked as an IBM Fellow at IBM Almaden and at Bell Laboratories, Murray Hill. He received his Ph.D. degree in Computer Science from the University of Wisconsin-Madison.

The 18th International Conference on Management of Data (COMAD),
14th-16th Dec, 2012 at Pune, India.