Data Analytics For Embellishing Educational Textbooks

Rakesh Agrawal
Microsoft Technical Fellow

Joint work with
Anitha Kannan, Krishnaram Kenthapadi, Sreenivas Gollapudi
Search Labs, Microsoft Research

December 19, 2011
The World We Live In

• 2/3 of the world’s 6 billion people live in the developing world. More than 1 in 6 live on less than $1 per day.

• Huge inequity in the availability of healthcare, education, and opportunities that condemn millions of people to lives of disease, poverty, and despair.

Inequities exist within developed societies too.
Development and Education

• Education: Primary vehicle for improving economic well-being of people

• Textbooks: Most cost-effective means of positively impacting educational quality
  – Also indispensable for fostering teacher learning and for their ongoing professional development
  – *Works by Clarke, Crossley, Fuller, Hanushek, Lockheed, Murby, Vail, and others*
Textbooks in Developing Countries

• Lack of adequate coverage of important concepts
  – [Grade IX Indian History]: The whole (medieval) period has been presented as a dull and dry history of dynasties, cluttered with the names and military conquests of kings, followed by brief acknowledgements of “social and cultural life”, “art and architecture”, “revenue administration”, and so on. The entire Mughal period (1526-1707) is disposed of in six pages.

• Lack of clarity
  – [Grade V Science, Baluchistan:] ‘Lever’ defined as a “strong rod or stick on which force is applied on its one end and can be rotated through some support and work is done on the other end”.

• Problems aggravated due to printing and distribution costs and centralized authoring [IBM05]
Outline

• Education and Data Mining
  – Embellishing textbooks
  – Research opportunities
Augmenting Textbooks with Web Content

Identify sections needing enrichment

- Decision model based on syntactic complexity of writing and dispersion of key concepts in the section [AGK+11a]

Add selective links to articles

- Determine key concepts in each section of a book and find links to authoritative web articles for these concepts [AGK+10]

Add selective images

- Find images most relevant for a section factoring in images in other sections [AGK+11b]

[AGK+11a] Identifying Enrichment Candidates in Textbooks. WWW 2011.
[AGK+10] Enriching Textbooks through Data Mining. ACM DEV 2010.
 Sections Needing Enrichment

**Decision Variables**
- Dispersion of key concepts
- Syntactic complexity of writing

**Probabilistic Decision Model**
- Enrich / Don’t / Examine

**Algorithmically Generated Training Set**
- Map a section to closest Wikipedia article version
- Impute immaturity score to section
- Perform thresholding to get labels
Many unrelated concepts in a section \( \Rightarrow \) Hard to understand

- \( V = \) set of key concepts discussed in section \( s \)
- \( rel(x,y) = true \) if concept \( x \) is related to concept \( y \)

- Dispersion\( (s) := \frac{|\{(x,y)|x,y\in V \text{ and } rel(x,y)=false\}|}{|V|(|V|-1)} \)
  - Fraction of concept pairs that are not related to each other

- Dispersion = (1 – Edge Density) of the concept graph

- Greater the dispersion, greater is the need for augmentation
Dispersion $= 1 - \frac{15}{30} = 0.5$

Dispersion $= 1 - \frac{3}{30} = 0.9$

Larger dispersion $\Rightarrow$ greater need for augmentation
Computing dispersion:

- **Concepts**: *Terminological noun phrases* [JK95, AGK+10]
  - Linguistic pattern A*N+ [A: adjective; N: noun]
  - Further refined using WordNet and Bing N-grams

- **Relation** \( \text{rel} \) between concepts:
  - Map concepts to Wikipedia articles
  - Exploit link structure to obtain the concept graph
100+ years of readability research
200+ Readability formulas
  – In widespread use (notwithstanding limitations)
Popular formulas:

<table>
<thead>
<tr>
<th>Formula</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch Reading Ease Score [17]</td>
<td>206.835 − 84.6 × S/W − 1.015 × W/T</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level [31]</td>
<td>−15.59 + 11.8 × S/W + 0.39 × W/T</td>
</tr>
<tr>
<td>Dale-Chall Grade Level [14]</td>
<td>14.862 − 11.42 × D/W + 0.0512 × W/T</td>
</tr>
<tr>
<td>Gunning Fog Index [23]</td>
<td>40 × C/W + 0.4 × W/T</td>
</tr>
<tr>
<td>SMOG Index [37]</td>
<td>3.0 + \sqrt[3]{30} × \sqrt{C/T}</td>
</tr>
<tr>
<td>Coleman-Liau Index [10]</td>
<td>−15.8 + 5.88 × L/W − 29.59 × T/W</td>
</tr>
<tr>
<td>Automated Readability Index [46]</td>
<td>−21.43 + 4.71 × L/W + 0.50 × W/T</td>
</tr>
</tbody>
</table>

Regression coefficients learned over specific datasets
  – McCall-Crabbs Standard Test Lessons
• Direct use of *Readability formulas* yielded poor results
• Variables abstracted from readability formulas:
  – Word length: Average syllables per word (S/W)
  – Sentence length: Average words per sentence (W/T)
• Larger syntactic complexity \(\Rightarrow\) greater need for augmentation
System Overview

**Decision Variables**
- Dispersion of key concepts
- Syntactic complexity of writing

**Algorithmically Generated Training Set**
- Map a section to closest Wikipedia article version
- Impute immaturity score to section
- Perform thresholding to get binary labels

**Probabilistic Decision Model**
- Enrich / Don’t / Examine

Textbooks
Probabilistic Decision Model

• Probabilistic scoring of a section needing enrichment through Binary logistic regression
• Probability that a section needs enrichment

\[ P(y = 1|z, w) = \frac{1}{1 + \exp\left\{- (b + z^T w)\right\}}. \]

• Optimal weight vector \( w \) learned from a training set of textbook sections
• Scores binned into
  – “Enrich”, “Don’t enrich”, or “Manually investigate to decide”
• Difficult to get qualified judges who would give consistent labels

• Map a textbook section to a most similar version of a similar article in a versioned repository (Wikipedia)

• Compute immaturity of this version as a proxy for that of the section

• Immaturity: function of relative edits on each day and a time window K, with more weight to recent edits (see paper)

• Immaturity computation reliable at only extreme ends

• But only few quality labels are needed

[AGK+11a] Identifying Enrichment Candidates in Textbooks. WWW 2011.
Application to Indian Textbooks

• Book corpus: 17 high school textbooks published by NCERT*
  – Grades IX – XII
  – Subject areas: Sciences, Social Sciences, Commerce, Math
  – 191 chapters, 1313 sections
• Followed by millions of students
• Available online

* National Council of Educational Research and Training
Many unrelated concepts [high dispersion]:

- Long sentences, e.g.,
  - Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be appropriate when the same business is carried on a large scale.
Results: Sections not needing enrichment

- Highly related concepts [low dispersion]:

- Written clearly with simple sentences [low syntactic complexity]
Augmenting Textbooks with Web Content

Textbooks

Identify sections that need enrichment
Decision model based on syntactic complexity of writing and dispersion of key concepts in the section

Enrich with textual web content
Determine key concepts in each section of a book and find links to authoritative web content for these concepts

Enrich with web images
Find images most relevant for a section factoring in images in other sections
1.1 Emergence of Macroeconomics

Macroeconomics, as a separate branch of economics, emerged after the British economist John Maynard Keynes published his celebrated book *The General Theory of Employment, Interest and Money* in 1936. The dominant thinking in economics before Keynes was that all the labourers who are ready to work will find employment and all the factories will be working at their full capacity. This school of thought is known as the classical tradition. However, the Great Depression of 1929 and the subsequent years saw the output and employment levels in the countries of Europe and North America fall by huge amounts. It affected other countries of the world as well. Demand for goods in the market was low, many factories were lying idle, workers were thrown out of jobs. In USA, from 1929 to 1933, unemployment rate rose from 3 per cent to 25 per cent (unemployment rate may be defined as the number of people who are not working and are looking for jobs divided by the total number of people who are working or looking for jobs). Over the same period aggregate output in USA fell by about 33 per cent. These events made economists think about the functioning of the economy in a new way. The fact that the economy may have long lasting unemployment had to be theorised about and explained. Keynes’ book was an attempt in this direction. Unlike his predecessors, his approach was to examine the working of the economy in its entirety and examine the interdependence of the different sectors. The subject of macroeconomics was born.
1.1 Emergence of Macroeconomics

Macroeconomics, as a separate branch of economics, emerged after the British economist John Maynard Keynes published his celebrated book *The General Theory of Employment, Interest and Money* in 1936. The dominant thinking in economics before Keynes was that all the labourers who are ready to work will find employment and all the factories will be working at their full capacity. This school of thought is known as the classical tradition. However, the Great Depression of 1929 and the subsequent years saw the output and employment levels in the countries of Europe and North America fall by huge amounts. It affected other countries of the world as well. Demand for goods in the market was low, many factories were lying idle, workers were thrown out of jobs. In USA, from 1929 to 1933, unemployment rate rose from 3 per cent to 25 per cent (unemployment rate may be defined as the number of people who are not working and are looking for jobs divided by the total number of people who are working or looking for jobs). Over the same period aggregate output in USA fell by about 33 per cent. These events made economists think about the functioning of the economy in a new way. The fact that the economy may have long lasting unemployment had to be theorised about and explained. Keynes’ book was an attempt in this direction. Unlike his predecessors, his approach was to examine the working of the economy in its entirety and examine the interdependence of the different sectors. The subject of macroeconomics was born.
Augmenting Textbooks with Images

Lessons from the learning literature:

• Visual material enhances comprehension and retention of information
• Most effective when presented in close proximity of the main material
• Use a small number of images that collectively best aid the understanding
Augmenting Textbooks with Images

**Image Mining**

Obtain images relevant to each section using complementary methods

*Comity*: Leverage image search provided by search engines

*Affinity*: Leverage image metadata on webpages

**Image Assignment**

Allocate most relevant images to each section such that

- Each section is augmented with at most $k$ images
- No image repeats across sections
Independent mining by complementary algorithms provides a broad selection of images to choose from.

Myopic: Section-specific image relevancy and hence images can repeat across sections within a chapter.
**MaxRelevantImageAssignment**

\[
\max \sum_{i \in I} \sum_{j \in S} x_{ij} \cdot \lambda_{ij}
\]

Relevance score of image i to section j

s.t.

\[
x_{ij} \in \{0, 1\} \quad \forall i \in I \forall j \in S
\]

=1 if image i is selected for section j else 0

\[
\sum_{i \in I} x_{ij} \leq K_j \quad \forall j \in S
\]

Constraint: At most \(K_j\) images can be assigned to section j

\[
\sum_{j \in S} x_{ij} \leq 1 \quad \forall i \in I
\]

Constraint: An image can belong to at most one section

Total relevance score for the chapter: sum of relevance scores of images assigned

Can be solved optimally in polynomial time
Value of Image Assignment

BEFORE IMAGE ASSIGNMENT

Sec 2: Magnetic field due to a current carrying conductor

Magnetic effect
Helmholtz Contour
Solenoid
Amperemeter
Galvanometer

Sec 3: Force on a current carrying conductor in a magnetic field

Magnetic effect
Electric motor cycle
Effect of magnet on domains
Meissner Effect
Descartes' magnetic field

Sec 6: Electric generator

Faraday disk generator
Magnetic effect
Two phase rotary converter
Descartes' magnetic field

AFTER IMAGE ASSIGNMENT

Sec 2: Magnetic field due to a current carrying conductor

Magnetic field
Simple electromagnet
Right hand rule
Right hand rule
Solenoid

Sec 3: Force on a current carrying conductor in a magnetic field

Electric motor cycle exploits electro
Drift of charged particles
Magnetic field around current
Electromagnet s attract paper clips...
Faraday's disk electric generator

Sec 6: Electric generator

Faraday disk generator
Single phase rotary converter
Two phase rotary converter
Three phase rotary converter
Descartes' magnetic field

Same images repeat across sections!
Richer set of images to augment the section
Evaluation on NCERT Textbooks

• User-study employing Amazon Mechanical Turk to judge the quality of results

• HIT (User task): A given image helpful for understanding the section?

• An image deemed helpful if the majority of 7 judges considered it so

• Helpfulness index:
  – Average of helpfulness score of the images over all sections
Performance

The number above a bar indicate helpfulness index for the corresponding subject (% of images found helpful)

• 94% of images deemed helpful
• Performance maintained across subjects
Recap

Textbooks

Identify sections that need enrichment

- Decision model based on syntactic complexity of writing and dispersion of key concepts in the section [AGK+11a]

Enrich with textual web content
- Determine key concepts in each section of a book and find links to authoritative web content for these concepts [AGK+10]

Enrich with web images
- Find images most relevant for a section factoring in images in other sections (Mining, Assignment, Ensembling) [AGK+11b]

- Technological solutions for
  - Diagnosing sections needing augmentation
  - Mining and optimal placement of web objects (images & articles)

- Promising results over High School textbooks across subjects and grades
Outline

- Education and Data Mining
  - Embellishing textbooks
  - Research opportunities
Textbook Augmentation

• Deeper analysis to identify key concepts discussed in a section (Discourse analysis? Formal Concept Analysis?)
• Diversity of augmentations
• Caption and placement of augmentations
• Extension to other multimedia types (video, speech)
• Evaluation methodology and performing a large field study to assess the quality of enrichments
Broader Questions

- Complementarity of algorithmic solutions to the crowdsourcing approaches
  - Tools for capturing feedback on textbooks (errors, better explanations, supplementary material, etc.)
  - Trust and ranking

- Deployment issues: making the augmented material available to students and teachers
  - Promising: Interactive DVDs [GPT’10], Low cost e-book readers, Cloud solutions
  - Study: social, behavioral, legal, cultural, policy, and political issues
Improving Education

• Identification of ill-matched material
  – Test score = f (student ability, suitability of material)
  – Learning: Item Response Theory

• Collaborative translation and localization of educational material

• Analysis of new pedagogical approaches
Summary

- Data mining has grown from solving enterprise problems to tackle problems to benefit individuals.
- The stage is set for data mining to provide fresh approaches to difficult problems hitherto unsatisfactorily addressed.
- The work on enriching education points to interesting new possibilities.
Thank you!

Questions

Search Labs’ mission is to invent next in Internet search and applications
Final Remark

Humanity’s greatest advances are not in its discoveries – but in how those discoveries are applied to reduce inequity.

Bill Gates.

Search Labs’ mission is to invent next in Internet search and applications