Data-Driven Education: Technologies and Directions

Rakesh Agrawal
Data Insights Laboratories, San Jose, CA
Visiting Professor, DIAS Lab, EPFL, Switzerland

November 2, 2017
Datafying Education: A Research Expedition

How to enhance the quality of the electronic textbooks?

How to form teams of students in a class?

How to create study plans for courses?
Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research
Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research
Data Mining for Enhancing Electronic Textbooks

Algorithmic enhancement of textbooks for enriching reading experience

- **References to selective web content**: Links to authoritative articles [AGK+10], images [AGK+11b] and videos [ACG+14], based on the focus of the section.

- **References to prerequisites**: Links to concepts necessary for understanding the present section, derived using a model of how students read textbooks [AGK+13].

Diagnostic tools for identifying weaknesses in textbooks

- **Within section deficiencies**: Complexity of writing and dispersion of concepts in the section [AGK+11a].

- **Across sections deficiencies**: Comprehension burden due to non-sequential presentation of concepts [ACG+12].

Validation on textbooks from U.S.A and India, on different subjects, across grades.

Joint work with Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, et al.
Data Mining for Enhancing Electronic Textbooks

**Algorithmic enhancement of textbooks for enriching reading experience**

- **References to selective web content**
  - Links to authoritative articles [AGK+10], images [AGK+11b] and videos [ACG+14], based on the focus of the section

- **References to prerequisites**
  - Links to concepts necessary for understanding the present section, derived using a model of how students read textbooks [AGK+13]

**Diagnostic tools for identifying weaknesses in textbooks**

- **Within section deficiencies**
  - Complexity of writing and dispersion of concepts in the section [AGK+11a]

- **Across sections deficiencies**
  - Comprehension burden due to non-sequential presentation of concepts [ACG+12]

**Validation on textbooks from U.S.A and India, on different subjects, across grades**

Joint work with Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, et al.
Identification of Deficient Sections

**Decision Variables**
- Dispersion of concepts
- Complexity of writing

**Algorithmically Generated Training Set**
1. Map a section to closest Wikipedia article version
2. Impute immaturity score to section
3. Perform thresholding to get labels

**Probabilistic Decision Model**
- Deficient / Good / Examine

See our WWW-2011 paper
Concept Dispersion

Many unrelated concepts → Hard to understand section

Dispersion(s) := Fraction of unrelated concept pairs

○ (1 – Edge Density) of the concept graph
Concept Dispersion

(a) A section with very low dispersion (Grade IX Mathematics book)

(b) A section with low dispersion (Grade IX Science book)

(c) A section with very large dispersion (Grade XII Sociology book)
Writing Complexity

➢ Readability Formulas (~100 years of research)

➢ More than 200 formulas in widespread use

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

➢ Sentence Length:

○ Avg. number of words per sentence

➢ Word Length:
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.
Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research
Optimal Grouping of Students in a Large Class

Given:

- a class of $N$ students
- each exhibiting a different ability level, $\theta_i \in \mathbb{R}_{\geq 0}$

How to partition them into $k$ groups, each of size $n$, so that the overall gain from peer learning is maximized ($N = k \times n$)

Ability Score $\theta_i$:
- Measured via a test (e.g. using Item Response Theory)

Work with Sharad Nandanwar & M.N. Murty (Under Submission)
Extension of work with Behzad Golshan & Evimaria Terzi (KDD 2014)
Prevalent Approaches

- Stratified (Ability-based grouping)
  - Put best with the best

- Pseudo-random (Diversity-based grouping)
  - Group students of all abilities together

- Inconclusive verdict from the empirical studies on the effectiveness [Richer76, Kulik92, Grossen96]

- Any computational alternative?
Model

• Every student gains from higher-ability peers [Vygotski]

• Learning gain for student $i$ in group $g$: $L_i(g) = R_i(g) - \theta_i$

  $\rightarrow R_i(g)$ is a function of $i$’s superior peers in group $g$

  $\rightarrow R_i(.)$ is different for different students in the same group

  $\rightarrow R_i(.)$ is different for the same student in different groups

  $\rightarrow$ Examples of $R_i(.)$: Mean, Median, $p$-percentile
Illustration: $R_i(g) = \text{mean of the scores of the superior peers}$

$R_{1}(g) = (2+3+4+5+6+7+8)/7 = 5 \quad L_{1}(g) = 5 - 1 = 4$

$R_{4}(g) = (5+6+7+8)/4 = 6.5 \quad L_{4}(g) = 6.5 - 4 = 2.5$

Box $i$ has the score of student $i$
Illustration: \( R_i(g) = \text{median of the scores of the superior peers} \)

Box \( i \) has the score of student \( i \)
Grouping Algorithm: Magic Partitions

Optimal for:

- $R_i(g) = \text{mean of the scores of the superior peers}$
- $R_i(g) = \text{median of the scores of the superior peers}$
- $R_i(g) = p$-percentile scores of the superior peers where $100-p$ divides 100
Assume scores are unique and $k$ divides $N$

1. Sort the scores in descending order

2. Partition the sorted scores into $N/k$ blocks of $k$ scores (students) each

3. Assign randomly from each block exactly one score (student) to each group
Experiment: Learning Gain

Data Distribution

- SSC (Normal)
- GATE (Log Normal)
- J&K Board (Truncated Normal)

Average Learning Gain

- MagicPartitions
- Random
- Stratified

Graphs showing the distribution of learning gain for different group sizes and data distributions.
Experiment: Group Structure

SS C
GATE
J & K Board
Learning gain only for below average students gained.
Time complexity of the proposed algorithm left open for future work.
Grouping Students (Recap)

- Computational approach points to a grouping strategy better from the conventional strategies
- For the objective of maximizing overall gain, the proposed partitions are optimal for a variety of reasonable learning models
- Low complexity of the algorithm
- Ongoing work: Incorporate learning gain from teaching, social constraints
Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research
Synthesizing Study Plans

Imagine you are an instructor who wants to offer a new course

You know the concepts you want to teach in the course, but need help with formulating the study plan:

a. What concepts should you cover in one session

b. The sequencing of sessions

Work with Behzad Golshan and Evangelos Papalexakis (EDM 2106)
Study Plans

Input

Study Plan #1
- Speed
- Displacement
- Acceleration
- Harmonic Motion
- Waves
- Sound
- Mass
- Force

Study Plan #2
- Speed
- Displacement
- Acceleration
- Harmonic Motion
- Sound
- Waves
- Mass
- Force
Axioms

• Learning Unit
  • *A group of coherent concepts suitable to be covered together*
  • Cohesion: Concepts within a learning unit must be closely related
  • Isolation: Concepts in different learning units must be independent
  • Unity: A concept should be covered in one unit

• Study plan
  • *An ordering of some number of learning units*
  • Prerequisite compliance: $L_1 < L_2 \implies$ concepts in $L_2$ not needed for $L_1$
  • Locality of references: $L_2$ builds upon $L_1 \implies$ $L_2$ should come soon after $L_1$
Problem Statement

Given a set of concepts,

- Partition them into a given number of learning units, and
- Provide a sequencing of learning units such that an objective function \( f \) is minimized

**Problem 1 (Study Plan Design Problem).** Given a concept graph \( G = \langle V, E \rangle \) with \( n > 0 \) nodes, and the number of desired learning units \( m \) (\( m \leq n \)), output an ordered vector of learning units \( \mathcal{L} = \langle L_1, L_2, \ldots, L_m \rangle \) to

\[
\text{Minimize: } f(\mathcal{L})
\]

\[
\text{s.t. } \forall i : L_i \subseteq V, L_i \neq \emptyset, \text{ and } \bigcup_i L_i = V.
\]
Objective Function

\[
f(L) = \sum \pi(u) > \pi(v) \& (u, v) \in E \ (\pi(u) - \pi(v)) C_p
\]
\[
+ \sum \pi(u) < \pi(v) \& (u, v) \in E \ (\pi(v) - \pi(u)) C_r
\]
\[
+ \sum \pi(u) = \pi(v) \& (u, v) \not\in E C_c
\]

Penalize if \( u \) is taught after \( v \) but \( v \) is a prerequisite for understanding \( v \)

Penalize if \( u \) is a prerequisite for \( v \) but \( u \) is taught much earlier

Penalize if \( u \) and \( v \) placed in the same learning unit but are unrelated

\( \pi(u) \): gives the position of the learning unit in which the concept \( u \) is covered

Prerequisite Compliance Violation

Locality of Reference Violation

Cohesion Violation (Also Isolation)

Unity Violation penalized by first two terms
Problem Complexity

NP-Complete

Minimum Linear Arrangement (minLA) problem reduces to our problem

See [AGP EDM2016] for our solution
Experiment

Input: 139 high school physics concepts from CK12.org
## Synthesized Study Plan

<table>
<thead>
<tr>
<th>Unit 1 (20 concepts)</th>
<th>Unit 2 (21 concepts)</th>
<th>Unit 3 (14 concepts)</th>
<th>Unit 4 (18 concepts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>buoyancy</td>
<td>acceleration</td>
<td>atom</td>
<td>calorimetry</td>
</tr>
<tr>
<td>euclidean vector</td>
<td>angular momentum</td>
<td>bohr model</td>
<td>change of state</td>
</tr>
<tr>
<td>force</td>
<td>angular velocity</td>
<td>conservation of energy</td>
<td>combined gas law</td>
</tr>
<tr>
<td>free body diagram</td>
<td>centripetal force</td>
<td>elastic collision</td>
<td>conversion of units</td>
</tr>
<tr>
<td>friction</td>
<td>circular motion</td>
<td>inelastic collision</td>
<td>double-slit experiment</td>
</tr>
<tr>
<td>impulse</td>
<td>displacement</td>
<td>kinetic energy</td>
<td>energy</td>
</tr>
<tr>
<td>inclined plane</td>
<td>keplers laws of ...</td>
<td>mass versus weight</td>
<td>energy conversion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit 5 (27 concepts)</th>
<th>Unit 6 (28 concepts)</th>
<th>Unit 7 (11 concepts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ammeter</td>
<td>beat</td>
<td>doppler effect</td>
</tr>
<tr>
<td>capacitor</td>
<td>color</td>
<td>general relativity</td>
</tr>
<tr>
<td>capacitors in circuits</td>
<td>concave lens</td>
<td>half-life</td>
</tr>
<tr>
<td>electric charge</td>
<td>conduction</td>
<td>length contraction</td>
</tr>
<tr>
<td>electric current</td>
<td>curved mirror</td>
<td>mathematical physics</td>
</tr>
<tr>
<td>electric field</td>
<td>diffraction</td>
<td>newtons law of univer...</td>
</tr>
<tr>
<td>electric power</td>
<td>diffraction grating</td>
<td>rc time constant</td>
</tr>
</tbody>
</table>
User Study

• Recruited 9 domain experts (Physics teachers, Graduate students)

• They were given the following tasks:
  1) Count the number of odd concepts in each learning unit that you believe do not belong to the unit
  2) Without changing any of the learning units proposed, what order do you suggest?
Results of the User Study

• Number of concepts that do not belong in the respective unit:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th># Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1</td>
<td>1</td>
<td>6</td>
<td>3.0</td>
<td>3.4</td>
<td>20</td>
</tr>
<tr>
<td>Unit 2</td>
<td>0</td>
<td>3</td>
<td>1.0</td>
<td>1.1</td>
<td>20</td>
</tr>
<tr>
<td>Unit 3</td>
<td>1</td>
<td>7</td>
<td>3.5</td>
<td>3.7</td>
<td>14</td>
</tr>
<tr>
<td>Unit 4</td>
<td>0</td>
<td>5</td>
<td>2.0</td>
<td>1.8</td>
<td>18</td>
</tr>
<tr>
<td>Unit 5</td>
<td>0</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>26</td>
</tr>
<tr>
<td>Unit 6</td>
<td>0</td>
<td>3</td>
<td>0.5</td>
<td>0.9</td>
<td>28</td>
</tr>
<tr>
<td>Unit 7</td>
<td>0</td>
<td>5</td>
<td>1.0</td>
<td>1.4</td>
<td>11</td>
</tr>
</tbody>
</table>

• Only two participants ordered the units somewhat differently
• The high school Physics teacher: our study plan was very clever
Synthesizing Study Plans (Recap)

• Formalized the problem of synthesizing study plans automatically
• Provided a novel and pragmatic solution
• The proposed method did not use domain specific knowledge
  • Generalizing to other areas seems promising
• Experimental results as well as the user study show that the problem of creating study plans is amenable to computational approaches
Further Work

- Incorporate user modeling into the system
  - Creating study plans that suit students' background/interests/abilities
- Investigate how human input (implicit or explicit) can improve the quality of generated study plans
Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research
Future Research Opportunities

- Validation of experimental results through deployment
- Synergies with crowd-sourcing approaches
- Use of logs of interactions data and personalization
- Performance evaluation methodologies and benchmarks
- Issues related to privacy, security, confidentiality, copyright, royalty

... 

*Magic happens when what is desperately needed meets what is technically feasible*
Selected References


Questions?