Big Education in the Era of Big Data

Irwin King

Associate Dean (Education), Faculty of Engineering, CUHK
Director, Shenzhen Rich Media and Big Data Analytics and Application Key Lab, SZRI, CUHK
PI, The Knowledge and Education Exchange Platform (KEEP), CUHK

Department of Computer Science and Engineering
The Chinese University of Hong Kong

king@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~king

©2014 Irwin King. All rights reserved.
Wlodzislaw Duch

Under-Secretary of State
Ministry of Science & Higher Education
Collective Probabilistic Factor Model

\[ \mathcal{L}(U, V, Z; F^x, F^y) = \sum_{i=1}^{m} \sum_{l=1}^{d} (f^x_{il} \ln x_{il} - x_{il}) + \sum_{i=1}^{m} \sum_{j=1}^{n} (f^y_{ij} \ln y_{ij} - y_{ij}) + \sum_{i=1}^{m} \sum_{k=1}^{d} ((\alpha_k - 1) \ln (u_{ik} / \beta_k) - u_{ik} / \beta_k) + \sum_{j=1}^{n} \sum_{k=1}^{d} ((\alpha_k - 1) \ln (v_{jk} / \beta_k) - v_{jk} / \beta_k) + \sum_{l=1}^{p} \sum_{k=1}^{d} ((\alpha_k - 1) \ln (z_{lk} / \beta_k) - z_{lk} / \beta_k) + \text{const.} \]

\[ u_{ik} \leftarrow u_{ik} \frac{\sum_{j=1}^{n} (f^y_{ij} v_{jk} / y_{ij}) + \sum_{l=1}^{p} (f^x_{il} z_{lk} / x_{il}) + (\alpha_k - 1) / u_{ik}}{\sum_{j=1}^{n} v_{jk} + \sum_{l=1}^{p} z_{lk} + 1 / \beta_k} \]

\[ v_{jk} \leftarrow v_{jk} \frac{\sum_{i=1}^{m} (f^y_{ij} u_{ik} / y_{ij}) + (\alpha_k - 1) / v_{jk}}{\sum_{i=1}^{m} u_{ik} + 1 / \beta_k} \]

\[ z_{lk} \leftarrow z_{lk} \frac{\sum_{i=1}^{m} (f^x_{il} u_{ik} / x_{il}) + (\alpha_k - 1) / z_{lk}}{\sum_{i=1}^{m} u_{ik} + 1 / \beta_k} . \]
The grass is greener on the other side...

*Be inspired!*

Stories and more stories...

*Be informed!*

The devil is in the details...

*Be challenged!*
Words of Wisdom

The **BEST** universities focus on **EDUCATION**!

The **BETTER** universities focus on citation numbers and impact factors…

The **GOOD** universities focus on counting the number of publications…
VALUE!
The Value of Big Data

Over 80% of organizations say:
- Big Data is critical to meet strategic objectives.
- Sharing insights is a must-have capability for businesses.
- Big Data will amplify other technology innovations.

Nearly 60% have started to use Big Data in specific cases... but only 3% consider themselves mature.

Where people struggle:
- Beginners: 28% Data quality
- Advanced: 30% Massive data volume
- Mature: 29% Skilled manpower

58% consider improved customer engagement and performance across all lines of business as high value.

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Big Education on Lifelong Learning
LIFELONG LEARNING  BY TERGYU

DOES LIFELONG LEARNING MEAN I HAVE TO STAY IN SCHOOL FOREVER?

WWW.BITSTRIPS.COM
Once you stop learning, you start dying…

Albert Einstein
Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
MILESTONES IN E-LEARNING

1971
The Open University opens in England with an open admissions policy, and begins broadcasting lectures on television. 25,000 students enroll.

1989
University of Phoenix launches its private, for-profit online school. 12 students enroll.

1993
Criteria is created by pioneer William Graziadei III, Ph.D.; e-learning systems must be easy to use, portable, replicable, scalable, and affordable.

1999
The term ‘e-Learning’ is coined at an educational seminar.

2004
Salman Khan records instructional YouTube videos to help his cousins with math. The rising popularity of these videos leads him to found the Khan Academy, a not-for-profit, free, online educational organization.
The rise in eLearning's popularity isn't showing any signs of slowing. In fact, judging by the following Top 10 eLearning statistics for 2014, the future of the eLearning Industry is brighter than ever:
Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
BIG MONEY

BIG USERS

BIG IMPROVEMENT

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
**TOP 10 ELEARNING STATISTICS**

for 2013

- 77% of American Corporations use some form of online learning.
- 70% of the global eLearning industry is made up of the US and Europe.
- 50% more productivity can be achieved with the help of eLearning.
- 85% of every dollar spent on classroom training is spent on delivering it (instructor time, travel, etc.).
The task of the modern educator is not to cut down jungles, but to irrigate deserts.

C.S. Lewis
Trends in Big Education
Cost effectiveness

Collaboration

Customization
Multimodal Learning

Harvard @ iTunes U

Harvard @ YouTube

harvard.edu

Harvard @ edX

Harvard @ Class Central
MOOC
Massive Open Online Course
Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
MINI MOOC GUIDE: 2014 TOP U.S. MOOC PROVIDERS

OVERVIEW OF MASSIVE OPEN ONLINE COURSES

YEAR:

2008

2014

PROVIDERS:

1 - 43

2 - 7,348+

3 - 938+

EDUCATORS:

2,225

INSTITUTIONS:

183,703,144+

STUDENTS:

Now students access MOOCs from all over the world

From 2008–2014, over 11,034 courses have been offered by MOOCs

ORIGIN STORY

With a collaboration of Stephen Downes and George Siemens, the first “Official” MOOC took place in the Fall 2008.

This distributed course, CCK08, focused on the nature of Connected Knowledge

In 2012, 7.1 million Americans took at least one online course

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
1200+ courses available

- Humanities: 20%
- Engineering: 5.1%
- Math & Stats: 6.6%
- Social Sciences: 5.7%
- Education & Teaching: 8.6%
- Health & Medicine: 11%
- Science: 11%
- Business & Management: 15%
- Comp. Sci & Prog: 16%

SOURCE: Edsurge
Small Private Online Course (SPOC) with Degree
Flipped Classroom

The Flipped Classroom
Teacher's Role: Guide on the Side

ACTIVITY TODAY

WATCH lecture online tonight!

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Microlearning
Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Personalized Learning

Basic Genetics Refresher (OPTIONAL)
- Genes and alleles (10 min)
- Mendelian inheritance (13 min)
- What is DNA? (12 min)

Modern Genomics
- Modern sequencing methods (11 min)
- Genomic economics (8 min)
- Personalized medicine (13 min)

Commercial Genomics (OPTIONAL)
- Genetic testing in the commercial world (11 min)
- Protecting privacy (9 minutes)
- Case study: direct-to-consumer genetics (12 min)
- Case study: family planning (13 min)
Active Learning
Peer Learning

Coursera MOOC Participation

MOOC participation by IP address, aggregated to 23,000 sq. km hexagons

- High
- Medium
- Low
- None

3+ million of 5+ million locatable IP addresses represented on the map
Who Really Cares?
Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland

Big Education Stakeholders

- Instructors
- Students
- Administrators
- Providers (Coursera, Khan Academy, edX)
- 3rd Parties (3rd)
Students

- Save Money
- Test Mastery
- Interactive and Collaboration
- Self-paced Learning
- Access on Multiple Devices
- Keep Focused
- Get Quick Feedback
- Access from Anywhere

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Instructors

- Release heavy teaching workload
- Focus on interaction with students in the classroom
- Provide personalized help
- Track student performance

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Faculty

So what’s in it for them?

3 out of 4 professors were inspired to change the way they taught in the traditional classroom after teaching a MOOC.
MOOCs save money

Save money

Improve engagement rate

First Day of School vs. First Day of College

Administrators

Reduce failure rate

Track student performance

Big Education in the Era of Big Data @ FedCSIS 2014, September 7-10, 2014, Warsaw, Poland
Case Studies
Predict Student Performance

### Input

<table>
<thead>
<tr>
<th>demographic information</th>
<th>sex, age, school, address, habit, health status, parents’ education, job, family size, income</th>
</tr>
</thead>
<tbody>
<tr>
<td>social information</td>
<td>romantic relationship, free time after school, going out with friends, weekend/workday alcohol consumption</td>
</tr>
<tr>
<td>school related information</td>
<td>study time, past failure course, extra paid support, family support, free time after school, previous course grade</td>
</tr>
</tbody>
</table>

### Output

- Binary classification (pass/fail)
- 5-Level classification (from I - very good / excellent to V - insufficient)
- Regression, with a numeric output that range between zero (0%) and twenty (100%)
Table 3: Binary classification results (PCC values, in %; underline – best model; bold – best within the input setup)

<table>
<thead>
<tr>
<th>Input Setup</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Portugal</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
<td>RF</td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
</tr>
<tr>
<td>A</td>
<td>91.9±0.0</td>
<td>88.3±0.7</td>
<td>86.3±0.6</td>
<td>90.7±0.3</td>
<td>91.2±0.2</td>
<td>89.7±0.0</td>
<td>90.7±0.5</td>
<td>91.4±0.1</td>
<td>93.0±0.3</td>
</tr>
<tr>
<td>B</td>
<td>83.8±0.0</td>
<td>81.3±0.5</td>
<td>80.5±0.5</td>
<td>83.1±0.5</td>
<td>83.0±0.4</td>
<td>87.5±0.0</td>
<td>87.6±0.4</td>
<td>88.0±0.3</td>
<td>88.4±0.3</td>
</tr>
<tr>
<td>C</td>
<td>67.1±0.0</td>
<td>66.3±1.0</td>
<td>70.6±0.4</td>
<td>65.3±0.8</td>
<td>70.5±0.5</td>
<td>84.6±0.0</td>
<td>83.4±0.5</td>
<td>84.8±0.3</td>
<td>84.4±0.4</td>
</tr>
</tbody>
</table>

† – statistical significance under pairwise comparisons with other methods.
* – statistical significance under a pairwise comparison with NV.

Table 4: Five-level classification results (PCC values, in %; underline – best model; bold – best within the input setup)

<table>
<thead>
<tr>
<th>Input Setup</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Portugal</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
<td>RF</td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
</tr>
<tr>
<td>A</td>
<td>78.5±0.0</td>
<td>60.3±1.0</td>
<td>59.6±0.9</td>
<td>76.7±0.4</td>
<td>72.4±0.4</td>
<td>72.9±0.0</td>
<td>65.1±0.9</td>
<td>64.5±0.6</td>
<td>76.1±0.0</td>
</tr>
<tr>
<td>B</td>
<td>69.5±0.0</td>
<td>49.8±1.2</td>
<td>47.9±0.7</td>
<td>57.5±0.8</td>
<td>52.7±0.6</td>
<td>58.7±0.0</td>
<td>52.0±0.6</td>
<td>51.7±0.6</td>
<td>62.9±0.2</td>
</tr>
<tr>
<td>C</td>
<td>32.9±0.0</td>
<td>30.4±1.0</td>
<td>31.0±0.7</td>
<td>31.5±0.6</td>
<td>33.5±0.6</td>
<td>31.0±0.0</td>
<td>33.7±0.6</td>
<td>34.9±0.5</td>
<td>32.8±0.6</td>
</tr>
</tbody>
</table>

† – statistical significance under pairwise comparisons with other methods.

Table 5: Regression results (RMSE values; underline – best model; bold – best within the input setup)

<table>
<thead>
<tr>
<th>I.S.</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Portugal</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
<td>RF</td>
<td>NV</td>
<td>NN</td>
<td>SVM</td>
<td>DT</td>
</tr>
<tr>
<td>A</td>
<td>2.01±0.0</td>
<td>2.05±0.02</td>
<td>2.09±0.02</td>
<td>1.94±0.04</td>
<td>1.75†±0.01</td>
<td>1.32±0.00</td>
<td>1.36±0.04</td>
<td>1.35±0.01</td>
<td>1.46±0.03</td>
</tr>
<tr>
<td>B</td>
<td>2.80±0.0</td>
<td>2.82±0.02</td>
<td>2.90±0.02</td>
<td>2.67±0.04</td>
<td>2.46†±0.01</td>
<td>1.89±0.00</td>
<td>1.88±0.02</td>
<td>1.87±0.01</td>
<td>1.78*±0.03</td>
</tr>
<tr>
<td>C</td>
<td>4.59±0.0</td>
<td>4.41±0.03</td>
<td>4.37±0.03</td>
<td>4.46±0.04</td>
<td>3.90†±0.01</td>
<td>3.23±0.00</td>
<td>2.79±0.02</td>
<td>2.76±0.02</td>
<td>2.93±0.02</td>
</tr>
</tbody>
</table>

† – statistical significance under pairwise comparisons with other methods.
* – statistical significance under a pairwise comparison with NV.

• Use three machine learning algorithms (instance-based learning Classifier, Decision Tree, and Naive Bayes) to predict students’ performance

• Three steps

  • 1st step: Attendance information for first four weeks, grade of 1st assignment,

  • 2nd step: Attendance information for first seven weeks, grade of 1st, 2nd assignments, midterm grade

  • 3rd step: Attendance information for first ten weeks, grade of 1st, 2nd and 3rd assignments, final exam grade, midterm grade.

- Algorithms
  - K-Star
    - One of the instance-based classifiers
  - C4.5
    - An extension to ID3 algorithm
  - Naive Bayes
- Three decision schemas
  - DS1: if at least one of the algorithms classifies student as a failure than this student will be considered as failure
  - DS2: if at least two algorithms classify student as a failure than this student will be considered as failure
  - DS3: if all three algorithms classify student as a failure than this student will be considered as failure
Fig. 1. Overall accuracy of results.
Fig. 2. Overall sensitivity of results.

Fig. 3. Overall precision of results.
KEEP
Knowledge and Education Exchange Platform
KEEP Education Cloud

• Educational resources to anyone, anytime, anywhere, on any device

• An education cloud platform to provide aggregated eLearning resources for teachers and students

• Big Data analytics for education

• Knowledge aggregation and technology integration!

• Multi-year, multi-discipline, and cross-institutional project with strong partners and alliances
Some Partners and Alliances
The KEEP Education Cloud
Work in Progress & Future Works
Natural Language Processing

- Text and semantic analysis
- Summarization
- Sentiment analysis
- Automated grading
- Q&A systems
Recommendations

- Personalized learning
- Courses, tutors, peering learning partners, etc.
- Learning resources, time allocation, etc.
- Career planning
Knowledge Map

- Explore topics
- Track topic changes
- Make topic comparisons and inferences
- Better search on concepts
Algorithms & Techniques

- Machine learning
- Data analytics
- Social computing
- Web intelligence
- Multimedia information processing
Get Involved

Partners and alliances are welcomed.
“I think you’ll find that mine is bigger...”
Concluding Remarks

• **Be Inspired**
  • Big Education is the focus!

• **Be Informed**
  • Big Data in Education is the VALUE proposition!

• **Be Challenged**
  • Use technologies to transform education in the Big Data Era!
CALL FOR PAPERS (TXT)

Researchers worldwide are currently producing more and more scholarly data of various types such as papers, books, patents, etc. Such data are big data by nature. For example, the DBLP Computer Science Bibliography provides bibliographic information on major computer science journals and proceedings. Additionally, DBLP indexes more than 2.3 million articles records containing title, pages, years and authors’ information, etc. Concurrently, scholars are associated with various academic activities such as conferences, workshops, congresses, peer review, and so on. Such scenarios have motivated us to explore the Web of Scholars in the context of big scholarly data on a global scale. It is imperative and vital for researchers to drive their knowledge towards the innovative generation of values from Big Scholarly Data. The emerging worldwide Web of Scholars demands a re-evaluation of existing techniques, such as data mining, recommender systems, and social network analysis. Furthermore, there is a demand for novel ways of developing algorithms, methods and techniques to foster the analysis and interpretation of social environments such as academic collaboration networks.

In this workshop, we will explore promising areas of research in big scholarly data, with a focus on the rapidly emerging field of the Web of Scholars. This workshop also seeks to answer noteworthy research questions such as:
- How to connect scholars on the web?
- How to facilitate collaboration among scholars?
- How to find the experts in a particular field?

Researchers are welcome to submit their papers that address these questions above and other topics below which may include, but are not limited to:
- Academic social network analysis
- Scientific recommendation
- Methods and tools for analyzing big scholarly data
- Indexing, searching, and mining scholarly data
- Connecting scholars using a Web approach
- Platforms and services for the Web of Scholars
- Web tools and techniques for big scholarly data
- Paradigms to promote scientific collaboration
- Scientific trends prediction
- Applications, use cases, and evaluations of big scholarly data

IMPORTANT DATES
Author Notification: Feb 4, 2014
Final Manuscript: Feb 12, 2014
WebET 2014

The Web has long been recognized as a powerful platform for teaching and learning. The educational community was among the early adopters of the technology and has contributed to its evolution. We are at this point at a major inflection point for Web-based Education Technologies. The convergence (“a perfect storm”) of new technologies supporting search, social media, semantics, data mining (Big Data), and others along with current interest to distributed educational pedagogies such as connectivism, behaviorism, and “the flipped classroom” promises to dramatically change Web-based Education Technologies in the near future. The interest in Massive Open Online Courses (MOOCs) has been described as “a tsunami in education” and has re-kindled valuable discussions regarding the role of WebET.
The Social Media and Social Computing Series focuses on publishing high quality references in the rapidly emerging area of social media and social computing. Both experimental/practical as well as theoretical investigations are welcome. The series targets both scholars and practitioners in social media and social computing for work in the intersection of computer science, information technology, psychology, economics, education and other social sciences. The advent of the Internet and the Web has resulted in social interactions and behaviors through the use of technologies and web services, e.g., hardware devices such as smart phones, tablets, RFID, etc., software services such as wikis, blogs, micro-blogs, social network sites, recommender systems, social bookmarking, social news, multimedia sharing sites, etc. Analyzing these technologically-enabled interactions in their social context will benefit information providers and information consumers. However, the large volume and scale of user-generated contents require effective modeling methods and efficient algorithms to handle these challenging problems.

Series Editor: Irwin King

Prof. King is Associate Editor of the IEEE Transactions on Neural Networks (TNN) and IEEE Computational Intelligence Magazine (CIM). He is a senior member of IEEE and a member of ACM, International Neural Network Society (INNS), and VP & Governing Board Member of the Asian Pacific Neural Network Assembly (APNNA). He serves the Neural Network Technical Committee (NTC) and the Data Mining Technical Committee under the IEEE Computational Intelligence Society.

https://www.ece.cuhk.edu.hk/irwin.king/home

Please send Proposals to either the Series Editor or Directly to:

LEONG Li-Ming
Editor, CRC Press
240 Macpherson Road, #08-01, Pines Industrial Building, S’pore 348574
li.ming@tandf.com.sg
Tel: (65) 67415166 x 115
VeriGuide

- **Similarity text** detection system
- Developed at **CUHK**
- Promote and uphold academic honesty, integrity, and quality
- Support **English, Traditional** and **Simplified Chinese**
- Handle `.doc`, `.txt`, `.pdf`, `.html`, etc. file formats
- Generate detailed **originality report** including **readability**
Acknowledgments

- Shouyuan Chan (Ph.D.)
- Xixian Chen (Ph.D.)
- Chen Cheng (Ph.D.)
- Junjie Hu (Ph.D.)
- Baichuan Li (Ph.D.)
- Guang Ling (Ph.D.)
- Haiqin Yang (Postdoc)
- Connie Yuen (Ph.D.)
- Hongyi Zhang (Ph.D.)
- Shenglin Zhao (Ph.D.)
- Tong Zhao (Ph.D.)
- Looking for more PhD students working on machine learning, Big Data, social computing, …
Acknowledgments

- Roger Cheung
- Byron Lai
- Daisy Lau
- Patrick Lau
- Sophia Man
- Lin Tsang
- Junfeng Yang
- Raymond Yuen

- Looking for more engineers, programmers, system analysts, etc. to work on KEEP…
The Chinese University of Hong Kong
CUHK Excellence

- The only university in Hong Kong having Nobel Laureates as faculty with five Distinguished Professors-at-Large

  - Prof. Yang Chen-Ning, Nobel Laureate in Physics
  - Prof. Charles Kao, Nobel Laureate in Physics
  - Prof. Sir James A. Mirrlees, Nobel Laureate in Economic Sciences
  - Prof. Yau Shing-Tung, Fields Medalist
  - Prof. Andrew Yao, Turing Award

- Nine academicians of Chinese Academy of Sciences and Chinese Academy of Engineering
Q&A
Reference

http://blog.gmee.com/online-in-60-seconds-infographic-a-year-later/
http://www.bigdatanetwork.net/sites/default/files/field/image/value_big_data_infographic.jpg
http://www.today.mccombs.utexas.edu/sites/default/files/Open-Cover-spring-2012-Big-Data_web.jpg
http://www.carolsim.com/designworks/archive/CEO/Training-Period.gif
http://www.mbae.edu/sites/default/files/Lifelong%20Learning.jpg?1398341082
http://elearningindustry.com/top-10-e-learning-statistics-for-2014-you-need-to-know
http://degreeoffreedom.org/mooc-infographics/
http://instructor/collegeletter/2010/05/teaching-e127375852117.jpg
http://lms2012.juliadib.files.wordpress.com/2013/05/free-e-learning-udacity-coursera-mitz-edx.png
http://3.bp.blogspot.com/-RSsESvlbHRO/UV5LEKXr0II/AAAAAAAABDM/0r-g3Ra6h5s/s1600/save+money.jpg
http://www.onefortraining.org/sites/default/files/core/SelfPacedLearningButton.jpg
http://add-assets.com/asset/17.jpg
http://www.cambridgedigital.com/userfiles/collaborative-learning.jpg
http://static.guim.co.uk/sys-images/Guardian/Pix/pictures/2008/08/30/Stressed-teacher-460x276.jpg
http://extras.mnginteractive.com/live/media/site36/2014/01/04/20140104__20140105_B1_CDXXTEACHERREVALS(1)_p1.jpg
http://ctctechcrunch2011.files.wordpress.com/2013/11/user-timeline.png?w=400
http://elearninginfographics.com/the-minds-behind-the-moocs-infographic/
https://d1z850zhxs7de.cloudfront.net/topics/ml/large-icon.png
http://www.excella.com/blog/wp-content/uploads/2013/02/Code-review_11935373_m.jpg
http://www.agendaitalia.it/wp-content/uploads/2013/02/Google-Knowledge-Graph.png