Empower MOOCs with AI

Jie Tang
Tsinghua University

The slides can be downloaded at http://keg.cs.tsinghua.edu.cn/jietang
Big Data in MOOC

- 149 partners
- 2400+ courses
- 33,000,000 users

- 1,000+ courses
- 10,000,000 users
- Chinese EDU association

- 110 partners
- 1,800 courses
- 14,000,000 users
- 10+ MicroMaster

- ~10 partners
- 40+ courses
- 1.6 million users
- “nanodegree”

- host >1,000 courses
- millions of users
Growth of MOOCs

Source: Class Central
58M Students  700+ Universities  6850 Courses

MOOCs in 2016. Analysis by Class Central

2019/6/21
Course Distribution by Subjects

- Science: 11.3%
- Business & Management: 16.8%
- Mathematics: 4.09%
- Engineering: 6.11%
- Art & Design: 6.73%
- Programming: 7.44%
- Health & Medicine: 8.27%
- Humanities: 9.41%
- Computer Science: 9.74%
- Education & Teaching: 9.36%
- Social Sciences: 10.8%
在网上学习全世界最好的课程。

免费加入
关于此课程：Learn about artificial neural networks and how they're being used for machine learning, as applied to speech and object recognition, image segmentation, modeling language and human motion, etc. We'll emphasize both the basic algorithms and the practical tricks needed to get them to work well.

制作方：多伦多大学

教学方：Geoffrey Hinton, Professor
Department of Computer Science
Deep Learning in Python
via DataCamp

Stanford University
Machine Learning
via Coursera  •  5-7 hours a week, 11 weeks long

Goldsmiths, University of London
Machine Learning for Musicians and Artists
via Kadenze  •  7 weeks long

Google
Deep Learning
via Udacity  •  6 hours a week, 12 weeks long

University of California, Berkeley
CS188.1x: Artificial Intelligence
via edX  •  12 weeks long

Johns Hopkins University
Practical Machine Learning
via Coursera  •  4-9 hours a week, 4 weeks long

Stanford University
Introduction to Artificial Intelligence
via Udacity  •  6 hours a week, 16 weeks long

University of Toronto
Neural Networks for Machine Learning
via Coursera  •  7-9 hours a week, 16 weeks long
XuetangX

Launched in 2013
Some exciting data...

- Every day, there are 10,000+ new students.
- An MOOC course can reach 100,000+ students.
- >35% of the XuetangX users are using mobile.
- traditional-\(\rightarrow\)flipped classroom-\(\rightarrow\)online degree.
Some exciting data…

• Every day, there are 5,000+ new students
• An MOOC course can reach 100,000+ students
• >35% of the XuetangX users are using mobile
• traditional->flipped classroom->online degree
• “Network+ EDU” (O2O)
  – edX launched 10+ MicroMaster degrees
  – Udacity launched NanoDegree program
  – GIT+Udacity launched the largest online master
  – Tsinghua+XuetangX will launch a MicroMaster soon
However...

• only ~3% certificate rate
  - The highest certificate rate is 14.95%
  - The lowest is only 0.84%

• Can AI help MOOC and how?
MOOC user = Student?

How to learn more effectively and more efficiently?

- Who is who? background, where from?
- Why MOOC? motivation? degree?
- What is personalization? preference?
MOOC course = University course?

How to discover the prerequisite relations between concepts and generate the concept graph automatically?

Thousands of Courses

How to leverage the external knowledge?

artificial intelligence

data mining

machine learning

data clustering

association rule

Probability Distribution

Maximum Likelihood

Hidden Markov Model
However to improve the engagement?

User

Knowledge

- artificial intelligence
- machine learning
- association rule
- data clustering
- data mining
LittleMU (小木)

LittleMU: Intelligent Interaction

1. User analysis

   Behavior modeling
   User Profiling

2. Course analysis

   Incentive analysis
   Course recommendation
   Automated video navigation
   Question answering

3. Prerequisite relation mining

   Course Content
   Concept extraction
   Knowledge base

- What is the complexity of bubble sort?
- It has a $O(n^2)$ complexity.
- How about the best case?
- The best case is $O(n)$, but average case is $O(n^2)$.
- Then how about the space complexity?
What is XiaoMU? Another Watson?

Jill Watson: Our Newest TA

- Creation of Prof. Ashok Goel
- TA for CS 7637: Knowledge-Based Artificial Intelligence
- Based on IBM Watson platform
- Anticipate that Jill will be able to answer 40% of ~10,000 questions posted to online forum
What is XiaoMU (“小木”)

• Not a Chatbot
  – “Good morning”, “did you have the breakfast?” — NO

• Not a teacher/TA
  – “Can you explain the equation for me?” — NO

• Instead, “小木” is more like a learning peer
  – Tell you some basic knowledge in her mind
  – Tell you what the other users are thinking/learning
  – Try to understand your intention
  – Teach “小木” what you know
What is XiaoMU ("小木") Knowledge Graph?
Acrostic Poem: 小木作诗—by 九歌
But most existing systems focus on passively interactions…
XiaoMU (小木)

LittleMU: Intelligent Interaction

1. User Profiling
   - Behavior modeling
   - User analysis

2. Course analysis
   - Course content
   - Concept extraction
   - Prerequisite relation mining

3. Incentive analysis
   - Course recommendation
   - Automated video navigation
   - Question answering

Behavior logs

Knowledge base

User Modeling

Intervention

Content Analysis
MOOC user

- Who is who? background, where from?
- Why MOOC? motivation? degree?
- What is personalization? preference?
Basic Analysis

![Graphs showing data for Non-Science and Science categories with axes for Female, Bachelor, and Graduate.]
Observation 1 – Gender Difference

Model 1: Demographics vs Certificate
Model 2: Demographics + Forum activities vs Certificate

- Females are significantly more likely to get the certificate in non-science courses.
- The size of the gender difference decreases significantly after we control for forum activities.

Table 4: Regression Analysis for Certificate Rate: All Users

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-Science</td>
<td>Science</td>
<td>Non-Science</td>
<td>Science</td>
</tr>
<tr>
<td>Female</td>
<td>0.014***</td>
<td>-0.003</td>
<td>0.002*</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>New Post</td>
<td>—</td>
<td>—</td>
<td>0.004***</td>
<td>0.038***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.001)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Reply</td>
<td>—</td>
<td>—</td>
<td>0.004**</td>
<td>0.001*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Video</td>
<td>—</td>
<td>—</td>
<td>0.000***</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Assignment</td>
<td>—</td>
<td>—</td>
<td>0.003***</td>
<td>0.000***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Bachelor</td>
<td>0.014***</td>
<td>0.003*</td>
<td>0.011***</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Graduate</td>
<td>0.007***</td>
<td>0.004</td>
<td>0.013***</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Effort</td>
<td>-0.072***</td>
<td>—</td>
<td>-0.072***</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td></td>
<td>(0.003)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.286***</td>
<td>0.018***</td>
<td>0.280***</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.006)</td>
<td>(0.011)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>Obs.</td>
<td>74,480</td>
<td>19,269</td>
<td>74,480</td>
<td>19,269</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.024</td>
<td>0.001</td>
<td>0.462</td>
<td>0.363</td>
</tr>
</tbody>
</table>
Observation 2 – Ability v.s. Effort

Model 1: Demographics vs Certificate
Model 2: Demographics + Forum activities vs Certificate

- Bachelors students are significantly more likely to get the certificate in non-science courses.
- Graduate students are more likely to get the certificate in science courses. After controlling for learning activities, the size of the effect is almost doubled.
- Forum activities are good predictors for getting certificates.
Forum activity vs. Certificate

— It is more important to be presented in forum, while the intensity matters less.

“近朱者赤”（Homophily）
— Certificate probability tripled when one is aware that she has certificate friend(s)
Dynamic Factor Graph Model

Model: incorporating “embedding” and factor graphs

\[
Y^t(i) = f(W_o Z^t(i) + b_o)
\]
\[
Z^t(i) = f(W_d S^t(i) + b_d)
\]
\[
S^t(i) = [Z^{t-1}_t p(i)^T, X^t(i)^T]^T
\]

Prediction labels:
Activities we are interested in, e.g., assignments performance and getting certificates.

\[
Y^t(i) = [Y_{t,i,0}, Y_{t,i,1}, \ldots, Y_{t,i,n-1}]^T
\]

Latent learning states
Every student’s status in at time \( t \) is associated with a vector representation

\[
Z^t(i) = [Z_{t,i,0}, Z_{t,i,1}, \ldots, Z_{t,i,m-1}]^T
\]

All features: time-varying attributes:
1. Demographics
2. Forum Activities
3. Learning Behaviors

\[
X^t(i) = [X_{t,i,0}, X_{t,i,1}, \ldots, X_{t,i,d-1}]^T
\]

## Certificate Prediction

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>LRC</td>
<td>92.13</td>
<td>83.33</td>
<td>46.51</td>
<td>59.70</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>92.67</td>
<td>52.17</td>
<td>83.72</td>
<td>64.29</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>94.48</td>
<td>61.54</td>
<td>74.42</td>
<td>67.37</td>
</tr>
<tr>
<td></td>
<td>LadFG</td>
<td><strong>95.73</strong></td>
<td><strong>73.91</strong></td>
<td><strong>79.07</strong></td>
<td><strong>76.40</strong></td>
</tr>
<tr>
<td>Non-Science</td>
<td>LRC</td>
<td>94.16</td>
<td>76.93</td>
<td>89.20</td>
<td>82.57</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>93.94</td>
<td>76.96</td>
<td>88.60</td>
<td>82.37</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>94.87</td>
<td><strong>80.22</strong></td>
<td>86.23</td>
<td>83.07</td>
</tr>
<tr>
<td></td>
<td>LadFG</td>
<td><strong>95.54</strong></td>
<td>79.76</td>
<td><strong>89.01</strong></td>
<td><strong>84.10</strong></td>
</tr>
</tbody>
</table>

- LRC, SVM, and FM are different baseline models
- LadFG is our proposed model
Predicting more

- **Dropout**
  - KDDCUP 2015, 1,000+ teams worldwide

- **Demographics**
  - Gender, education, etc.

- **User interests**
  - Computer science, mathematics, psychology, etc.

- ...
User Tagging

- **Observation**: With probability 43.91%, a user will enroll in a course in the same category as the last course (s)he enrolled in.

- **Method**: Use course categories to tag users who enroll in courses under this category to aid course recommendation.
Random Walk with Restart

• Use RWR on the user-tag bipartite with # of enrolled courses in the tag (category) as edge weight to generate tag preference of users.

• Offline test in course recommendation

<table>
<thead>
<tr>
<th></th>
<th>top1</th>
<th>top3</th>
<th>top5</th>
<th>top10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.0071</td>
<td>0.0247</td>
<td>0.0416</td>
<td>0.0890</td>
</tr>
<tr>
<td>+Tag</td>
<td>0.0185</td>
<td>0.0573</td>
<td>0.1022</td>
<td>0.2198</td>
</tr>
</tbody>
</table>
XiaoMU (小木)
Knowledge Graph

- How to extract concepts from course scripts?
- How to recognize (prerequisite) relationships between concepts?

[1] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. ACL'17.
In this course, we will teach some basic knowledge about **data mining** and its application in **business intelligence**.

**Video script**

**Concept Extraction**

**Candidate Concept Extraction** → **Semantic Representation Learning** → **Graph-based Ranking**

**Vector representation**

Learned via embedding or deep learning

**data mining**

| 0.8 | 0.2 | 0.3 | ... | 0.0 | 0.0 |

**business intelligence**

| 0.1 | 0.1 | 0.2 | ... | 0.8 | 0.7 |

**Semantic Representation Learning**
Prerequisite Relationship

How to extract the prerequisite relationship?

[1] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. ACL'17.
Prerequisite Relationship Extraction

• Step 1: First extract important concepts
• Step 2: Use Word2Vec to learn representations of concepts

data mining

0.8 0.2 0.3 ... 0.0 0.0

business intelligence

0.1 0.1 0.2 ... 0.8 0.7

Vector representation
Learned via embedding or deep learning
Prerequisite Relationship Extraction

• Step 1: First extract important concepts
• Step 2: Use Word2Vec to learn representations of concepts
• Step 3: Distance functions
  – Semantic Relatedness
  – Video Reference Distance
  – Sentence Reference Distance
  – Wikipedia Reference Distance
  – Average Position Distance
  – Distributional Asymmetry Distance
  – Complexity Level Distance
### Result of Prerequisite Relationship

<table>
<thead>
<tr>
<th>Classifier</th>
<th>ML</th>
<th>DSA</th>
<th>CAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>SVM</td>
<td>P</td>
<td>63.2</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>68.5</td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>65.8</td>
<td>65.7</td>
</tr>
<tr>
<td>NB</td>
<td>P</td>
<td>58.0</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>58.1</td>
<td>60.5</td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>58.1</td>
<td>59.4</td>
</tr>
<tr>
<td>LR</td>
<td>P</td>
<td>66.8</td>
<td>67.6</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>60.8</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>63.7</td>
<td>64.2</td>
</tr>
<tr>
<td>RF</td>
<td>P</td>
<td>68.1</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>70.0</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>69.1</td>
<td>72.6</td>
</tr>
</tbody>
</table>

Table 2: Classification results of the proposed method(%).

[1] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. ACL'17.
System Deployed
XiaoMU (小木)

User Modeling
- Behavior logs
  - Behavior modeling
  - User Profiling

Intervention
- Incentive analysis
- Course recommendation
- Automated video navigation
- Question answering

Content Analysis
- Course content
- Concept extraction
- Prerequisite relation mining

Knowledge base

1. User analysis
2. Course analysis
3. Intervention
What we can do?

User modeling

Knowledge

artificial intelligence
machine learning
association rule
data clustering
data mining
• Let start with a simple case
  – Course recommendation based on user interest
Course Recommendation

With the learned user model

Course topic analysis

Low frequency
- LDA training
- User clustering
- Course perquisite modeling

High frequency
- Latent interest modeling
- Collaborative filtering

Recommendation result

Rule based adjustment

[1] Xia Jing, Jie Tang, Wenguang Chen, Maosong Sun, and Zhengyang Song. Guess You Like: Course Recommendation in MOOCs. WI'17.
## Course Recommendation

**Guess you like**

<table>
<thead>
<tr>
<th>Course Title</th>
<th>Instructor</th>
<th>Participants</th>
<th>Start Date</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company Finance</td>
<td></td>
<td>422</td>
<td>7天前</td>
<td></td>
</tr>
<tr>
<td>Management Accounting</td>
<td></td>
<td>328</td>
<td>5天前</td>
<td></td>
</tr>
<tr>
<td>University Computing Tutorial</td>
<td></td>
<td>14267</td>
<td>9个月前</td>
<td></td>
</tr>
<tr>
<td>IC Design and Method</td>
<td></td>
<td>818</td>
<td>3个月前</td>
<td></td>
</tr>
<tr>
<td>TOEFL Preparation: From Exam Taker's Perspective</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water Mechanics</td>
<td></td>
<td>2349</td>
<td>9个月前</td>
<td></td>
</tr>
<tr>
<td>Filial piety</td>
<td></td>
<td>499</td>
<td>9个月前</td>
<td></td>
</tr>
<tr>
<td>Landscaping Design</td>
<td></td>
<td>850</td>
<td>8个月前</td>
<td></td>
</tr>
<tr>
<td>Political Philosophy</td>
<td></td>
<td>214</td>
<td>4个月前</td>
<td></td>
</tr>
<tr>
<td>IELTS English Proficiency Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Further Recommendations

- **Win in Mobile Internet**: Innovations and Business Models (2017 Spring)
  - 3 months ago
  - 3,083 participants

- **u.lab 0x**: Beyond the Network System: Innovation and Co-creation of the Future
  - 8 months ago
  - 5,132 participants

  - 3 months ago
  - 1,492 participants

- **Distributed System and Data Management**: Distributed System and Data Management (Micro-Course)
  - 5 months ago
  - 1,099 participants

- **Modem Life Aesthetics**: Modern Life Aesthetics (2017 Summer)
  - 3 months ago
  - 2,907 participants
Online A/B Test

**Top-k recommendation accuracy (MRR)**

Comparison methods:
- **HCACR** – Hybrid Content-Aware Course Recommendation
- **CACR** – Content-Aware Course Recommendation
- **IBCF** – Item-Based Collaborative Filtering
- **UBCF** – User-Based Collaborative Filtering

**Online Click-through Rate**

Comparison methods:
- **HCACR** – Our method
- **Manual strategy**

---

**Performance Comparison**

![Graph showing performance comparison](image1.png)

**Online CTR Comparison**

![Graph showing online CTR comparison](image2.png)
Context based Recommendation

[Image of a webpage with a video and course recommendations]

- Hi, jietang. 我是智能学习助理小木，有什么问题要问我的呢？学习疑问、平台使用问题，我都会尽力回答哦～试试这样：
  - 如何申请电子版证书？
  - 自主课程什么意思？
  - 人工智能

[Course recommendations]

- 数据结构与算法
- 语言与沟通
- 情绪与情感
- 面试心理学
- 积极心理学

[Video]

- 智力的测量方法
- 下载字幕

遇到疑问，小木来帮忙！点击下方知识点，查看解答
• Let start the simplest case
  – Course recommendation based on user interest

• What can we else?
  – Interaction when watching video?
Smart Jump
—Automated suggestion for video navigation
Let’s begin with …

The example is that …

First, we introduce …

Next … capital assets … investment property …

On Average: 2.6 Clicks = 5 seconds
Two Numbers

According to what we have discussed we find that the fifth activity belongs to cash outflow of a business activity.

\[ 5S \times 8,000,000 \text{ users} = 1.3 \text{ years} \]
Science courses contain much more frequent jump-backs than non-science courses.

Users in non-science courses jump back earlier than users in science courses.

Users in science courses are likely to rewind farther than users in non-science courses.
Observations – User Related

- 6.6% users prefer 10 seconds
- 9.2% users prefer 17 seconds
- 6.6% users prefer 20 seconds
Video Segmentation

In the next ninth economic activity
The enterprise has paid 4,000,000 yuan

What is the money used for

Of which 2,500,000 yuan is paid for the expenditure of sales department
1,500,000  for the expenditure of administrative department

\[
\text{argmax}_{\Delta t} \frac{2 R_{e_{cj}}}{R_{e_{cj}}} \cdot \frac{R_{n_s}}{R_{n_s}}
\]

- \( R_{e_{cj}} \): rate of effective complete-jumps (start position and end position located in different segments).
- \( R_{n_s} \): rate of non-empty segments (contains at least one start position or end position of some complete-jumps).
Problem Formulation

\[
\arg \max_{\Theta} P(s_j|u, v, s_i; \Theta)
\]

Prediction Results

<table>
<thead>
<tr>
<th>Course</th>
<th>Model</th>
<th>AUC</th>
<th>P@1</th>
<th>P@3</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>LRC</td>
<td>72.46</td>
<td>35.95</td>
<td>65.54</td>
<td>80.13</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>71.92</td>
<td>35.45</td>
<td>66.15</td>
<td>81.99</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>74.02</td>
<td>37.61</td>
<td><strong>76.04</strong></td>
<td>89.59</td>
</tr>
<tr>
<td>Non-science</td>
<td>LRC</td>
<td>72.59</td>
<td>69.23</td>
<td>73.23</td>
<td>89.32</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>73.52</td>
<td>68.39</td>
<td>76.64</td>
<td>91.30</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>73.57</td>
<td>67.56</td>
<td><strong>88.43</strong></td>
<td>96.05</td>
</tr>
</tbody>
</table>

- LRC, SVM, and FM are different models
- FM is defined as follows

\[
\hat{y}(x_i) = w_0 + \sum_{j=1}^{d} w_j x_{i,j} + \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} x_{i,j} x_{i,j'} \langle p_j, p_{j'} \rangle
\]
More

- Let start the simplest case
  - Course recommendation based on user interest
- What can we else?
  - Interaction when watching video?
  - What kind of questions did the users ask?
Question Answering

User Query

Question Classification

Platform FAQ

Wikipedia

Forum Archive

Service

Others

Question Answer Assembling
Query Categories

- **PLATFORM**: XuetangX platform
- **CONTENT**: enrollments, courses, teachers
- **CONCEPT**: simple knowledge point
- **DISCUSS**: general discussion, comparison
- **FEEDBACK**: suggestions, complains
- **SMALLCHAT**: small chat
- **CUSTOMER**: personal questions (e.g., account)
- **MISC**: meaningless questions (e.g., asjedkjqw)
- **SERVICE**: poem, recommendation
Candidate Dataset

• Wikipedia: 892,185
• Forum Archive: 65,001
• Platform FAQ: 137
• Zhihu: 1,000+
• CSDN: 670
• Course Structure: 8 types
Question Classification

- #Training (March 2017 – August 2017): 2162
- #Test (September 2017): 499
  Precision: 0.77, Recall: 0.78
### Online Result

<table>
<thead>
<tr>
<th></th>
<th>#Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total_request</td>
<td>20604</td>
</tr>
<tr>
<td>feedback</td>
<td>470</td>
</tr>
<tr>
<td>Feedback_ratio</td>
<td>0.023</td>
</tr>
<tr>
<td>User-thumb_up</td>
<td>245</td>
</tr>
<tr>
<td>User-thumb_down</td>
<td>225</td>
</tr>
<tr>
<td>Thumb_ratio</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Question Retrieval

- Queries in PLATFORM category: 538
- Q-A pairs in Candidate Set: 77

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
<th>Hit @ 1</th>
<th>Hit @ 3</th>
<th>Hit @5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES (TF-IDF)</td>
<td>0.617</td>
<td>0.558</td>
<td>0.698</td>
<td>0.748</td>
</tr>
<tr>
<td>Word2vec + WMD</td>
<td>0.695</td>
<td>0.602</td>
<td>0.745</td>
<td>0.817</td>
</tr>
<tr>
<td>Word2vec + Cosine</td>
<td>0.653</td>
<td>0.577</td>
<td>0.685</td>
<td>0.726</td>
</tr>
<tr>
<td>1.0<em>WMD+1.5</em>ES</td>
<td>0.728</td>
<td>0.640</td>
<td>0.781</td>
<td>0.845</td>
</tr>
</tbody>
</table>
More

• Let start the simplest case
  – Course recommendation based on user interest
• What can we else?
  – Interaction when watching video?
  – What kind of questions did the users ask?
  – Interaction->intervention
XiaoMU would like to ask you

**Question:** What are the shortcomings of Raven Progressive Test? (3 users thumb up)

**Fundamental Challenges (3W):**
- When
- to Whom
- ask What (question)
Preliminary study—first version

Question: What are the shortcomings of Raven Progressive Test?
### Active Question

#### Positive Direct Feedback:

<table>
<thead>
<tr>
<th>Time</th>
<th>Classified Type</th>
<th>Feedback ratio (at least once)</th>
<th>Thumb_up Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0914 -- 0917</td>
<td>On/Off</td>
<td>12.4% (17/134)</td>
<td>31.2% (10/32)</td>
</tr>
<tr>
<td>0923 -- 0930</td>
<td>Social Pressure/None</td>
<td>17.5% (151/864)</td>
<td>47.1% (113/240)</td>
</tr>
</tbody>
</table>

- Each question lasts for 10 seconds;
- Displaying questions are selected manually to ensure strong connection with the on-going content;
Bandit Learning with Implicit Feedback

An online learning framework: contextual bandit

- Click/Buy etc. as reward, features of content/user/etc. as context.
- Adaptively and sequentially learning
- Successfully deployed for recommender system and ad displaying.

The problem is about the definition of rewards

- Is non-click indicates negative impression?
- Examination hypothesis:
  - Click occurs if and only if examination happens
  - Implication: no-click DOES NOT necessarily mean negative feedback
Model

Classical bandit model with linear reward:

\[ \mathbb{E}[r_{t,a} | x_{t,a}] = x_{t,a}^\top \theta_a^*. \]

- Reward is 1 if clicked, 0 if non-clicked.
- Inevitably linear regret.

Our model: E-C bandit (Examination-click bandit)

\[ \mathbb{P}(C_t = 1 | E_t = 0, x_{C,t}) = 0 \]
\[ \mathbb{P}(C_t = 1 | E_t = 1, x_{C,t}) = \rho(x_{C,t}^\top \theta_C^*) \]

Thus:

\[ \mathbb{P}(E_t = 1 | x_{E,t}) = \rho(x_{E,t}^\top \theta_E^*) \]

\[ \mathbb{E}[C_t | x_t] = \rho(x_{C,t}^\top \theta_C^*) \rho(x_{E,t}^\top \theta_E^*) \]

The common goal: regret minimization

\[ \text{BayesRegret}(T, \pi) = \sum_{t=1}^{T} \mathbb{E} \left[ \max_{a \in \mathcal{A}_t} f_{\theta^*}(x^a) - f_{\theta^*}(x^{a_t}) \right] \]
Model

E-C bandit
• A generative model of click, explicitly incorporating examination;
• Examination as a binary variable, is by nature NOT observable, thus a latent variable;

The essential problem
• Is it possible to learn E-C bandit under online learning’s paradigm?
  • Regret analysis affirms learnability to some extent.
• How to learning E-C bandit on the fly?
  • Variational approximation together with Thompson sampling
Algorithm: Parameter Estimation

The Log-likelihood of one sample:

\[
\log P(\theta_C, \theta_E | x_C, x_E, C) = \log P(C | \theta_C, \theta_E, x_C, x_E) + \log P(\theta_C, \theta_E) + \log \text{const} \\
= C \log \rho(x_C^T \theta_C) \rho(x_E^T \theta_E) + (1 - C) \log (1 - \rho(x_C^T \theta_C) \rho(x_E^T \theta_E)) \\
- \frac{1}{2} (\theta_C - \hat{\theta}_C)^T \Sigma_C^{-1} (\theta_C - \hat{\theta}_C) - \frac{1}{2} (\theta_E - \hat{\theta}_E)^T \Sigma_E^{-1} (\theta_E - \hat{\theta}_E) + \log \text{const}
\]

The variational lower bound:

- Jensen’s inequality for log-sum;
- 2-degree polynomial lower bound of log-logistic function;
- Thus, a lower bound in the form of 2-degree polynomial, which leads to an approximate Gaussian posterior when given a Gaussian prior and allows for \(O(1)\) update.

\[
\rho(x) = \frac{1}{1 + e^{-x}}
\]
Algorithm – Decision Making

Thompson sampling:
• Choose any arm by its probability of being the best among the candidate;
• Easy to implement and well integrated with our estimation procedure (Recall we have approximate Gaussian posterior of the parameters).
Algorithm 1 Thompson sampling for E-C Bandit

1: Initiate $\Sigma_C = \lambda I$, $\Sigma_E = \lambda I$, $\hat{\theta}_C = \theta_{C,0}$, $\hat{\theta}_E = \theta_{E,0}$.
2: for $k = 0, 1, 2, \ldots$ do
3: Observe the available arm set $A_k \subset A$ and its corresponding context set $X_k := \{(x^a_C, x^a_E) : a \in A_k\}$.
4: Randomly sample $\tilde{\theta}_C \sim N(\hat{\theta}_C, \Sigma_C)$, $\tilde{\theta}_E \sim N(\hat{\theta}_E, \Sigma_E)$.
5: Select:
$$a_k = \arg \max_{a \in A_k} \rho((x^a_C)^T \tilde{\theta}_C) \rho((x^a_E)^T \tilde{\theta}_E)$$
6: Play the selected arm $a_k$ and Observe the reward $C_k$.
7: Update $\Sigma_C, \hat{\theta}_C, \Sigma_E, \hat{\theta}_E$ according to Eq (3), (4), (5), (6) respectively.
8: end for

\[\Sigma_{C,\text{post}}^{-1} = \Sigma_{C}^{-1} + 2q^{1-C} \lambda(\xi_C) x_C x_C^T \quad (3)\]

\[\hat{\theta}_{C,\text{post}} = \Sigma_{C,\text{post}}(\Sigma_{C}^{-1} \hat{\theta}_C + \frac{1}{2} (-q)^{1-C} x_C) \quad (4)\]

\[\Sigma_{E,\text{post}}^{-1} = \Sigma_{E}^{-1} + 2\lambda(\xi_E) x_E x_E^T \quad (5)\]

\[\hat{\theta}_{E,\text{post}} = \Sigma_{E,\text{post}}(\Sigma_{E}^{-1} \hat{\theta}_E + \frac{1}{2} (2q - 1)^{1-C} x_E) \quad (6)\]
Regret Analysis

Sublinear regret is guaranteed if:

- MLE estimate (i.e., log-loss estimate in our 0-1 reward case) is accurate;
- Thompson sampling samples from the true posterior.
- See detailed proof in the paper and appendix.
- Proof’s framework is the same as Russo, 2014. Key proposition: aggregated empirical discrepancy is bounded within a sub-linear increasing ellipse w.h.p. (Proposition 1 in the paper.)

By experiment we demonstrate the approximation is tight, and result improving.
Evaluation - Simulation

Figure 1: Cumulative regret over 100 simulations.

Figure 2: Discrepancy bound given by Proposition 1.
Evaluation – Empirical data

Figure 4: Performance comparison on MOOC videos’ data
Conclusion

Explicitly modeling implicit feedback as composition of examination and relevance judgement provides finer modeling and leads to better result.

Further work:
• Quantitative analysis on the impact of approximated posterior on the cumulative regret;
• Generalization from one item’s recommendation to multiple items case.
XiaoMU (小木)

1. Behavior logs
   - Behavior modeling
   - User Profiling
   - User modeling

2. Course analysis
   - Incentive analysis
   - Course recommendation
   - Automated video navigation
   - Question answering

3. Content analysis
   - Course content
   - Concept extraction
   - Prerequisite relation mining

User Modeling

Intervention

Content Analysis
Recent Publications

• Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. In ACL'17.

• Xia Jing, Jie Tang, Wenguang Chen, Maosong Sun, and Zhengyang Song. Guess You Like: Course Recommendation in MOOCs. WI'17.

• Han Zhang, Maosong Sun, Xiaochen Wang, Zhengyang Song, Jie Tang, and Jimeng Sun. Smart Jump: Automated Navigation Suggestion for Videos in MOOCs. In WWW'17 Companion.

• Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. 2016. Modeling and Predicting Learning Behavior in MOOCs. In WSDM'16. 93–102.


• Jie Tang, Tracy Xiao Liu, Zhenyang Song, Xiaochen Wang, Xia Jing, Jiezhong Qiu, Zhenhuan Chen, Chaoyang Li, Han Zhang, Liangmin Pan, Yi Qi, Xiuli Li, Jian Guan, Juanzi Li, and Maosong Sun. LittleMU: Enhancing Learning Engagement Using Intelligent Interaction on MOOCs. submitted to KDD.


• 薛宇飞，敬峡，裘捷中，唐杰，孙茂松. 一种在线课程中的作业互评方法：中国，201510531490.2.（中国专利申请号）

• 唐杰, 张茜, 刘德兵. 用户退课行为预测方法及装置. 201610292389.0 （中国专利申请号）
Thank you!

Collaborators: Jian Guan, Xiuli Li, Fenghua Nie (XuetangX)
Jie Gong (NUS), Jimeng Sun (GIT)
Wendy Hall (Southampton)
Maosong Sun, Tracy Liu, Juanzi Li (THU)
Xia Jing, Zhenhuan Chen, Liangmin Pan, Jiezhong Qiu, Han Zhang,
Zhengyang Song, Xiaochen Wang, Chaoyang Li, Yi Qi (THU)

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://arnetminer.org/data
http://arnetminer.org/data-sna