Concept Graph Learning from Educational Data

Yiming Yang, Hanxiao Liu, Jaime Carbonell and Wanli Ma

School of Computer Science
Carnegie Mellon University

February 3, 2015

Outline of the Talk

- Motivation
- Concept Representation Schemes
- Concept Graph Learning
- Experiments & Empirical Results
- Future Work

Scenario: Massive course materials are online available from different course providers

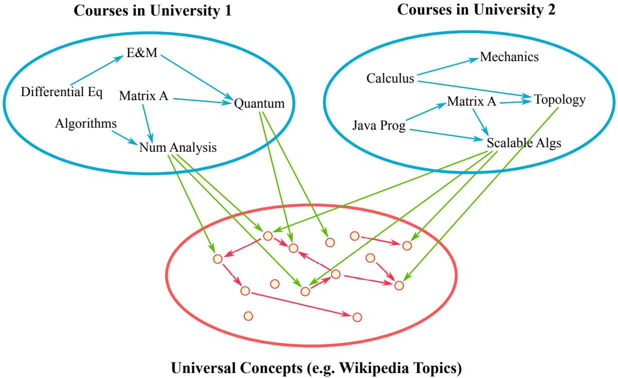
- Universities, Coursera, Edx, MIT OpenCourseWare ...

Challenge: How to integrate the scattered information?

A CMU graduate: "*After completing courses A, B on Coursera, what course shall I take next at CMU?*"

- Lack a method to measure the course overlapping and the course prerequisite relations across institutions.

We address this by putting cross-institutional courses under a canonical language—concept.



Goal: Learning a universal graph of concepts based on

1. Course-level prerequisite relations
2. Concept representation of courses

Outline of the Talk

- Motivation
- Concept Representation Schemes
- Learning the Concept Graph
- Experiments & Empirical Results
- Future Work

$\Downarrow$ crawling and parsing

```
<course>
 <id>6.080</id>
 <name>Great Ideas in Theoretical Computer Science</name>
 <tag>Electrical Engineering and Computer Science</tag>
 <description>This course provides a challenging introduction to some of the central ideas of
      theoretical computer science. It attempts ... </description>
 <keywords>computer science,theoretical computer science,logic,turing machines,computability,
      finite automata,godel,complexity,polynomial time,efficient algorithms ... </keywords>
 <calendar>Introduction # Logic # Circuits and finite automata # Turing machines # Reducibility
      and Godel # Minds and machines # Complexity # Polynomial time # P and NP # NP-
      completeness # NP-completeness in practice # Space complexity and more ... </calendar>
</course>
```

Concepts $\stackrel{\text{def}}{=}$ Words

- Vocabularies not controlled
  - CMU 10-715: shattering coefficient; MIT 15.097: growth function
- Words are in multiple granularities $\Longrightarrow$ interpretability $\downarrow$

Bag of Words

Bag of Wikipedia Categories

Wikipedia Classifier
$\Longrightarrow$

Concepts $\stackrel{\text{def}}{=}$ Wikipedia Categories

- Improved interpretability
- Controlled vocabulary at the right granularity
  - Shattering number, Growth function $\stackrel{\text{Wikipedia Classifier}}{\Longrightarrow}$ Computational Learning Theory
- Leverage oceans of knowledge in Wikipedia

Schemes based on dimensionality reduction

- Sparse Coding of Words
  - Trained on the given courses—purely unsupervised
- Distributed Word Embedding
  - Trained on Wikipedia articles—leverages exterior info

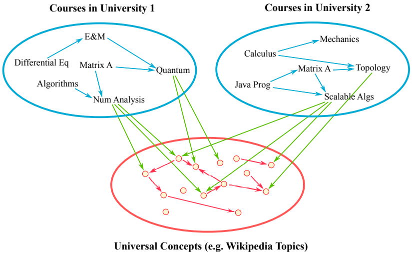
Concepts $\stackrel{\text{def}}{=}$ Dimensionality-reduced vectors

- Controlled “vocabulary”
  - Words are mapped onto a unified latent space
- Concept granularity can be controlled by latent dimensionality
- Hard to interpret

Outline of the Talk

- Motivation
- Concept Representation Schemes
- Learning the Concept Graph
- Experiments & Empirical Results
- Future Work

Problem Formulation



Observed course-level relations	$\mathcal{O}$	
Concept representation of courses	$\boldsymbol{X}$	n by p
Concept graph	$\boldsymbol{A}$	p by p

How to evaluate $\boldsymbol{A}$?

1. Map $\boldsymbol{A}$ to an estimated course graph $\boldsymbol{\Theta}$ (n by n) via $\boldsymbol{X}$.
2. Then, evaluate the quality of $\boldsymbol{\Theta}$ with $\mathcal{O}$.

Problem Formulation

How to map the concept graph $\boldsymbol{A}$ to an estimated course graph $\boldsymbol{\Theta}$?

—through a bilinear form:

$$\boldsymbol{\Theta} \stackrel{\text{def}}{=} \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top$$

Explanation:

$$\theta_{ij} \stackrel{\text{def}}{=} \sum_{u,v} a_{uv} x_{iu} x_{jv}$$

- θ_{ij}: strength from course j to course i
- a_{uv}: strength from concept v to concept u

Each course-level prerequisite θ_{ij} is defined as the cumulative effect of multiple concept-level prerequisites $\sum_{u,v} a_{uv} x_{iu} x_{jv}$

Problem Formulation

How to evaluate the estimated course graph $\boldsymbol{\Theta}$ with our observed course-level relations $\mathcal{O}$?

i.e. how to define the loss function over $\boldsymbol{\Theta}$ w.r.t. $\mathcal{O}$?

Problem

- Only positive examples are available
- Treating unobserved course relations as negative examples leads to highly skewed label set

Solution: ranking

- We hope $\theta_{ij} > \theta_{ik}$ if $j \in prereq\,(i)$ and $k \notin prereq\,(i)$

CGL objective with p^2 variables

$$\min_{\boldsymbol{A}\in\mathcal{R}^{p\times p}} \quad C\sum_{(i,j,k)} \ell\left(\theta_{ij}-\theta_{ik}\right)+\frac{1}{2}\|\boldsymbol{A}\|_{\mathrm{F}}^{2}$$
$$\text{s.t.} \quad \boldsymbol{\Theta}=\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\top}$$

Problem

- $\boldsymbol{A}$ can be a huge dense matrix (e.g. is 15,396 by 15,396 for words-based concept representation)
- Dual space? #dual variables $= O\left(n^3\right)$

Solution: derive an equivalent optimization problem with only n^2 ($n^2 \ll p^2$) variables.

Introduce slack variable $\boldsymbol{S} \in \mathcal{R}^{n \times n}$ for constraint $\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top$.

The Lagrangian is

$$\mathcal{L} = C \sum_{(i,j,k)} \ell\left(\theta_{ij} - \theta_{ik}\right) + \frac{1}{2}\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \left\langle \boldsymbol{S}, \boldsymbol{\Theta} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \right\rangle$$

$\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}}$ should vanish at the stationary point

- $\implies \boldsymbol{A}^* = \boldsymbol{X}^\top {\boldsymbol{S}^*}^\top \boldsymbol{X}$
- $\boldsymbol{A}^* \in \mathcal{R}^{p \times p}$ only has n^2 $(n^2 \ll p^2)$ degrees of freedom!

Equivalent CGL objective with n^2 variables

$$\begin{aligned} \min_{\boldsymbol{S} \in \mathcal{R}^{n \times n}} \quad & C \sum_{(i,j,k)} \ell\left(\theta_{ij} - \theta_{ik}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Theta}\boldsymbol{S}^\top\right) \\ \text{s.t.} \quad & \boldsymbol{\Theta} = \boldsymbol{K}\boldsymbol{S}\boldsymbol{K} \end{aligned}$$

We choose the squared hinge loss $\ell(x) = (\max(1 - x, 0))^2$

- large-margin property: strong generalization ability
- smoothness: allows Nesterov's accelerated GD
    - GD: 37.3min & 1490 iterations on MIT
    - accelerated GD: 3.08 min & 103 iterations on MIT

An alternative of GD: Inexact Newton Method

- To avoid the huge Hessian—use a matrix-free Conjugate Gradient to compute the Newton direction

Outline of the Talk

- Motivation
- Concept Representation Schemes
- Learning the Concept Graph
- Experiments & Empirical Results
- Future Work

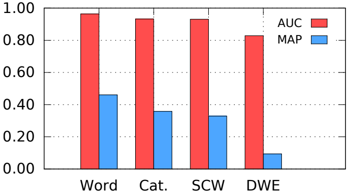
Table : Datasets Statistics[1]

University	Department	# Courses	# Prerequisites	# Words
MIT[2]	*	2322	1173	15396
Caltech	*	1048	761	5617
CMU	CS, STATS	83	150	1955
Princeton	MATH	56	90	454

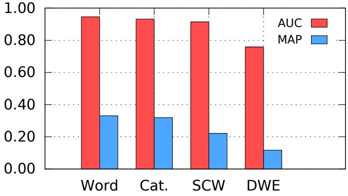
Metrics for evaluation: MAP and AUC

[1] available at http://nyc.lti.cs.cmu.edu/teacher/dataset/
[2] MIT OpenCourseWare http://ocw.mit.edu/index.htm

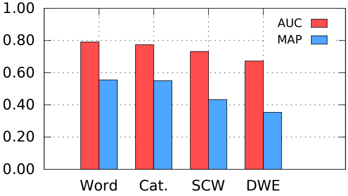
Experiments
Comparison among Concept Representation Schemes



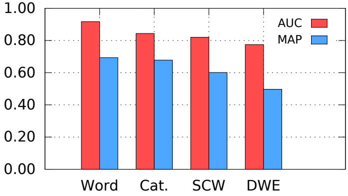
(a) CGL.Rank on MIT Data



(b) CGL.Rank on Caltech Data



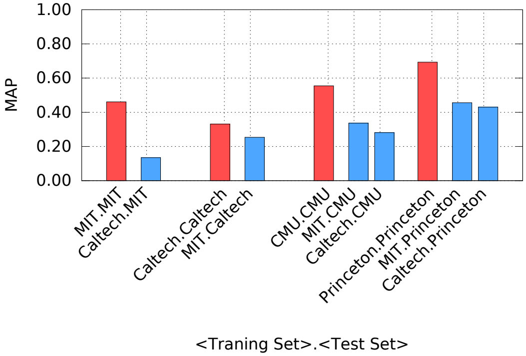
(c) CGL.Rank on CMU Data



(d) CGL.Rank on Princeton Data

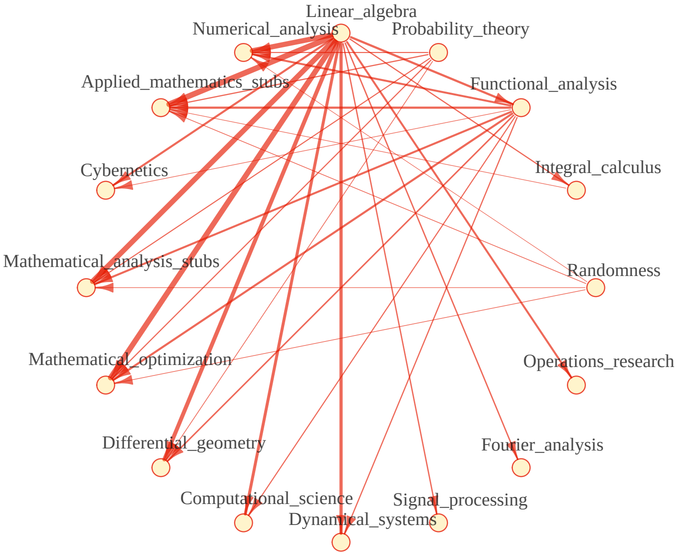
Words $\succeq$ Categories $\succ$ Sparse Coding $\succ$ Distributed Word Embedding

A good concept graph should be universal, thus should be transferable across different institutions



- There is always a performance loss if we go across institutions.
- We do get good transfer.

Empirical Results
Concept Graph for MIT



Outline of the Talk

- Motivation
- Concept Representation Schemes
- Learning the Concept Graph
- Experiments & Empirical Results
- Future Work

Future Work

- Deploying the induced concept graph for personalized curriculum planning (on-going work)
  - Student's academic background/goal $\stackrel{\text{def}}{=}$ bag-of-concepts
  - Find an optimal sequence of courses?
- Cross-language transfer learning by using Wikipedia categories (concepts) as the interlingua.

The End

Thanks!

hanxiaol@cs.cmu.edu