Course Concept Extraction in MOOCs via Embedding-Based Graph Propagation

Reporter: Liangming PAN
Authors: Liangming PAN, Xiaochen WANG, Chengjiang LI,
Juanzi LI and Jie TANG
Knowledge Engineering Group
Tsinghua University

2017-11-28

Outline

- Backgrounds
- Related Works
- Methods
- Experiments and Analysis
- Conclusion

MOOCs

Massive open online courses (MOOCs) have become increasingly popular and offered students around the world the opportunity to take online courses from prestigious universities.

- World
- China

Course Concept

Course concepts refer to the ***knowledge concepts*** taught in the course, and the ***related topics*** that help students better understand course videos.

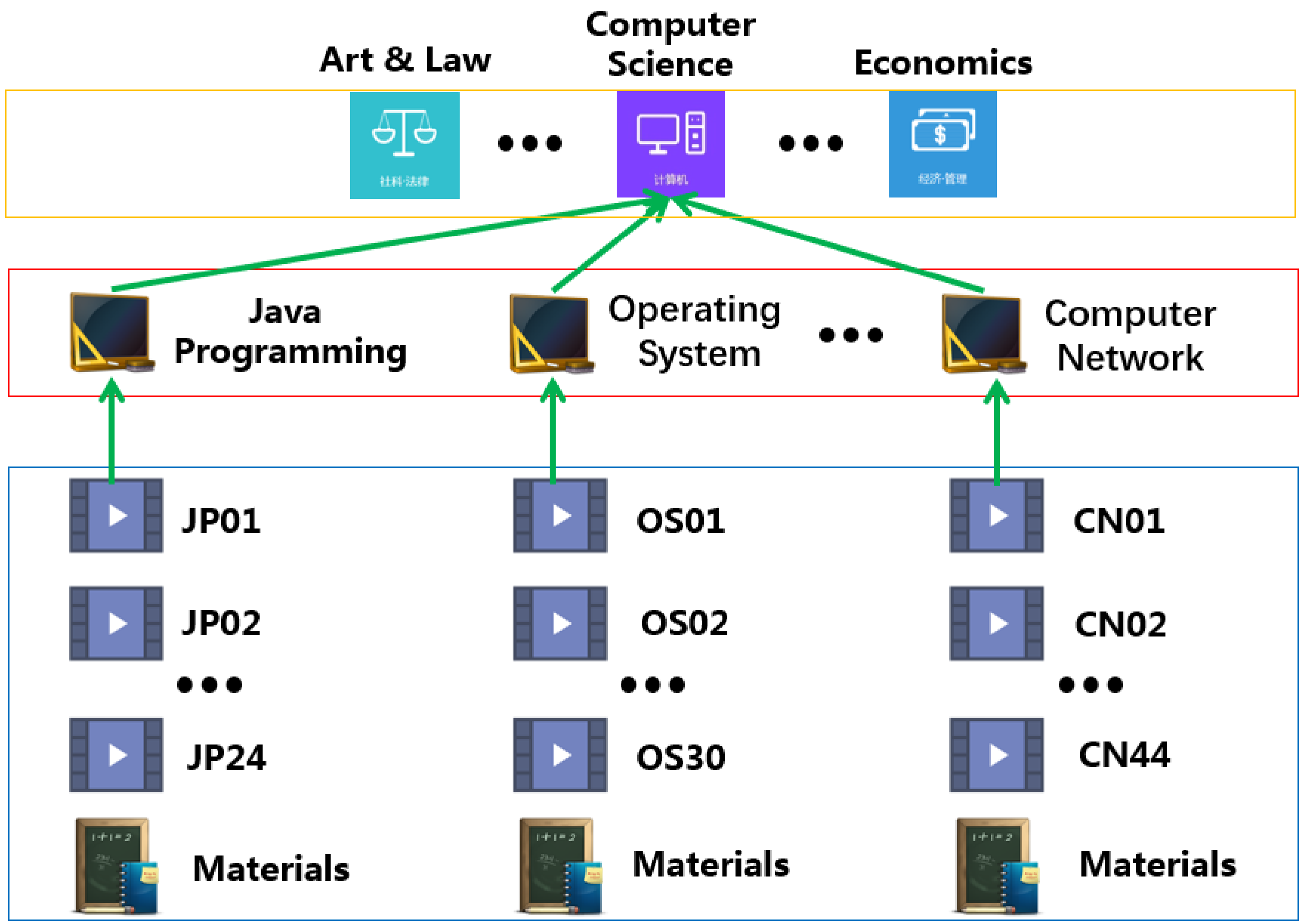
You might learn how to write a **bubble sort** and learn why a **bubble sort** is not as good as a **heapsort**. Next, we are going to talk about the **quick sort** algorithm. **Quicksort** is an algorithm invented in the 1960s by doctor Tony Hoare. It is also called the **partition exchange sort**, and is a typical algorithm based on **divide-and-conquer**.

… … …

Now we have the first version of **Q sort**. After we make an analysis on its performance, performance, we will find that **quicksort** is an **unstable sorting algorithm**. Fortunately, the **quick sort** has an average **time complexity** of n log n, and in most cases, it can achieve its optimal performance. We first estimate its performance under **independent uniform distribution**.

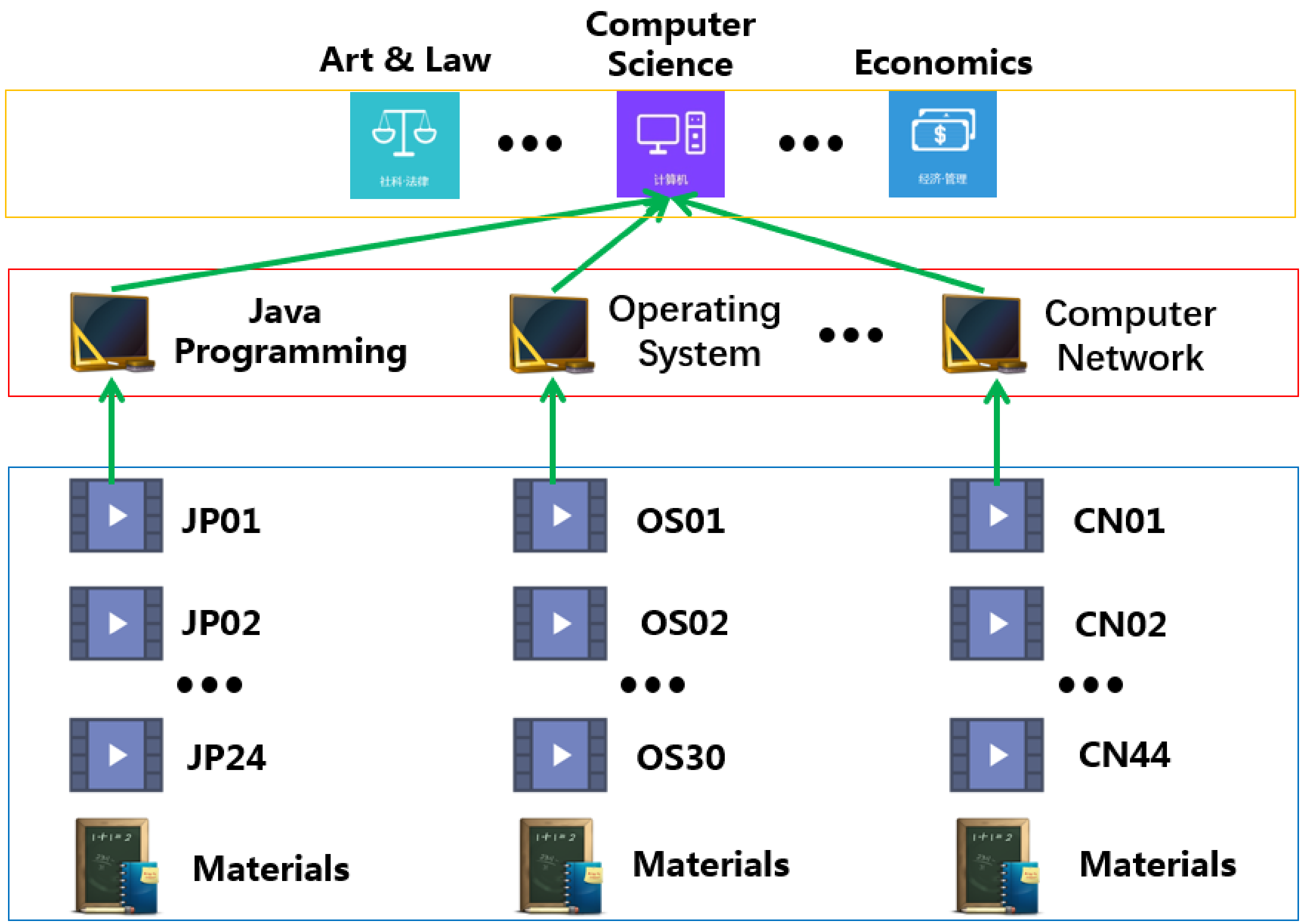
Why Course Concept Extraction?

Video-based Structure

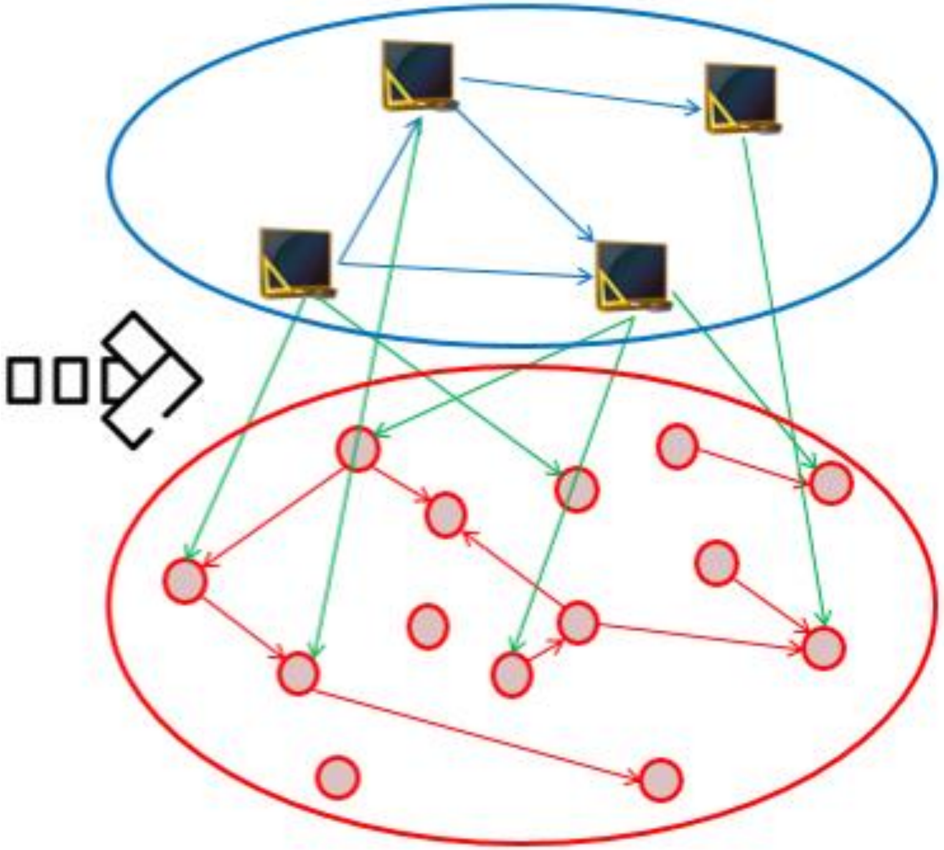


Why Course Concept Extraction?

Video-based Structure

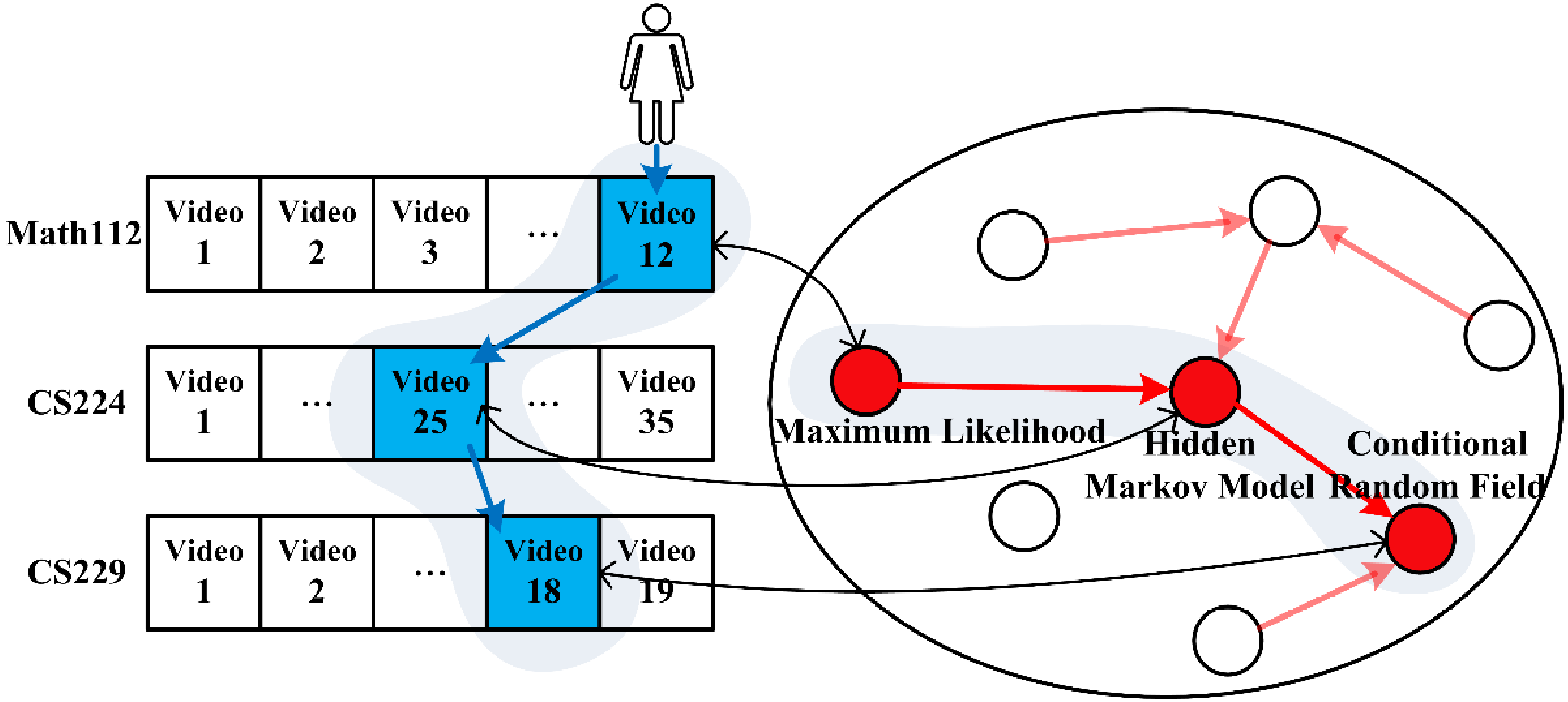


Concept-based Structure



Why Course Concept Extraction?

- **Motivation 1**. Manually extracting course concepts in MOOCs is infeasible
- **Motivation 2**. A concept map can help improve the learning experience of students



Outline

- Backgrounds
- Related Works
- Methods
- Experiments and Analysis
- Conclusion

Related Works: Keyphrase Extraction

Category	Method	Authors	Year
Supervised Learning	**Naive Bayes**	Eibe Frank et al.	IJCAI 1999
	Decision Tree	Peter D. Turney et al.	Journal of IR 2000
	Maximum Entropy	Wentau Yih et al.	WWW 2006
	SVM	Patrice Lopez et al.	2010
Graph-based Methods	**TextRank**	Mihalcea R and Tarau P	EMNLP 2004
	ExpandRank	Wan et al.	AAAI 2008
	Topical PageRank	Liu et al.	EMNLP 2010
Joint Learning-based	**Combining Text Summarization and Keyphrase Extraction**	Zha et al.	SIGIR 2002
		Wan et al.	ACL 2007

Why Course Concept Extraction *Hard*?

Low-frequency problem: Course video captions often contain many course concepts with *low frequency*, primarily for three reasons:

- Course video captions are relatively ***short documents***
- Many infrequent course concepts are from ***other prerequisite or related courses***.
- A disambiguated course concept tends to be ***expressed in various ways***, which produces many scattered infrequent terms.

You might learn how to write a **bubble sort** and learn why a **bubble sort** is not as good as a **heapsort**. Next, we are going to talk about the **quick sort** algorithm. **Quicksort** is an algorithm invented in the 1960s by doctor Tony Hoare. It is also called the **partition exchange sort**, and is a typical algorithm based on **divide-and-conquer**.

… … …

Now we have the first version of **Q sort**. After we make an analysis on its performance, we will find that **quicksort** is an **unstable sorting algorithm**. Fortunately, the **quick sort** has an average **time complexity** of n log n, and in most cases, it can achieve its optimal performance. We first estimate its performance under **independent uniform distribution**.

Properties of Course Concepts

A course concept has the following ***three*** properties:

- **Phraseness**
  - A course concept should be a semantically and syntactically correct phrase.
- **Informativeness**
  - A course concept should represent a specific scientific or technical concept.
- **Relatedness**
  - A course concept should be related to a course.

The above properties are hard to be captured by ***local statistical information*** because of the ***Low-frequency problem***.

Outline

Backgrounds

Related Works

Methods

Experiments and Analysis

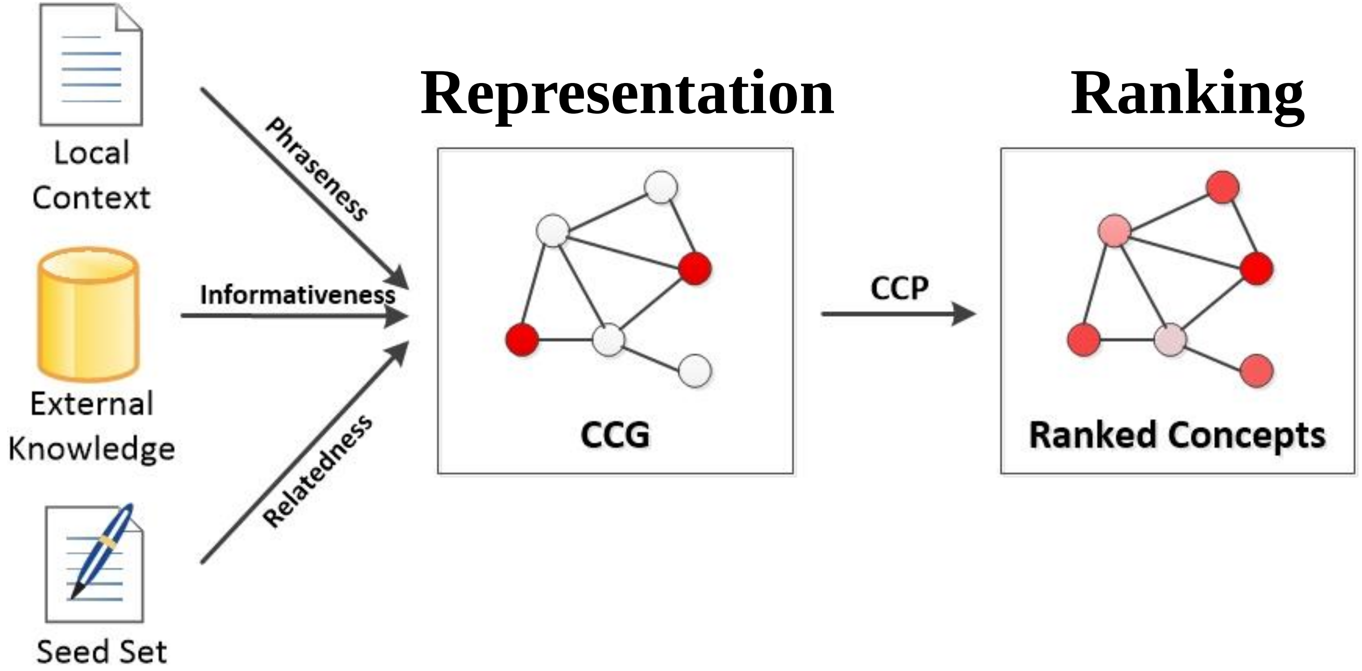
Conclusion

Method Overview

1. Candidate Extraction: Extracting ***noun phrases*** within K-grams from course video captions based on ***linguistic patterns***.

2. Representation: Incorporating ***external knowledge*** from online encyclopedia to learn ***semantic representations*** for candidate course concepts.

3. Ranking: Ranking candidate course concepts based on the representation.



Representation

- **Phraseness Measurement:** PMI-based method

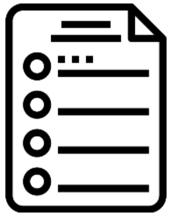
2-grams
$$Ph(w_1, w_2) = \frac{2 \times freq(w_1, w_2)}{freq(w_1) + freq(w_2)} \qquad (1)$$

N-grams (N>2)
$$Ph(t) = max\{Ph(f_i, b_i) \mid i = 1, \cdots, N-1\} \qquad (2)$$

Averaging
$$ph(c) = \alpha \cdot F[ph^D(c)] + (1 - \alpha) \cdot F[ph^E(c)] \qquad (3)$$

Representation

- **Semantic Relatedness**

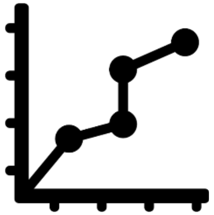


Candidate Course Concepts

Wikipedia Corpus

- ➢ **Entity Annotation**
  - **Labeling all entities** in Wikipedia Corpus
- ➢ **Word Embeddings**
  - Training **Word Embeddings** in Wikipedia
- ➢ **Concept Representation**
  - **Obtaining the vector** for each candidate
- ➢ **Semantic Relatedness**
  - Calculating SR by **cosine distance**

$$SR(a,b) = \frac{1}{2}\left(1 + \frac{v_a \cdot v_b}{|v_a| \cdot |v_b|}\right)$$



Semantic Relatedness

Course Concept Graph Construction（CCG）

The **course concept graph (CCG)** of a course is a weighted undirected fully-connected graph denoted as G = (V,E).

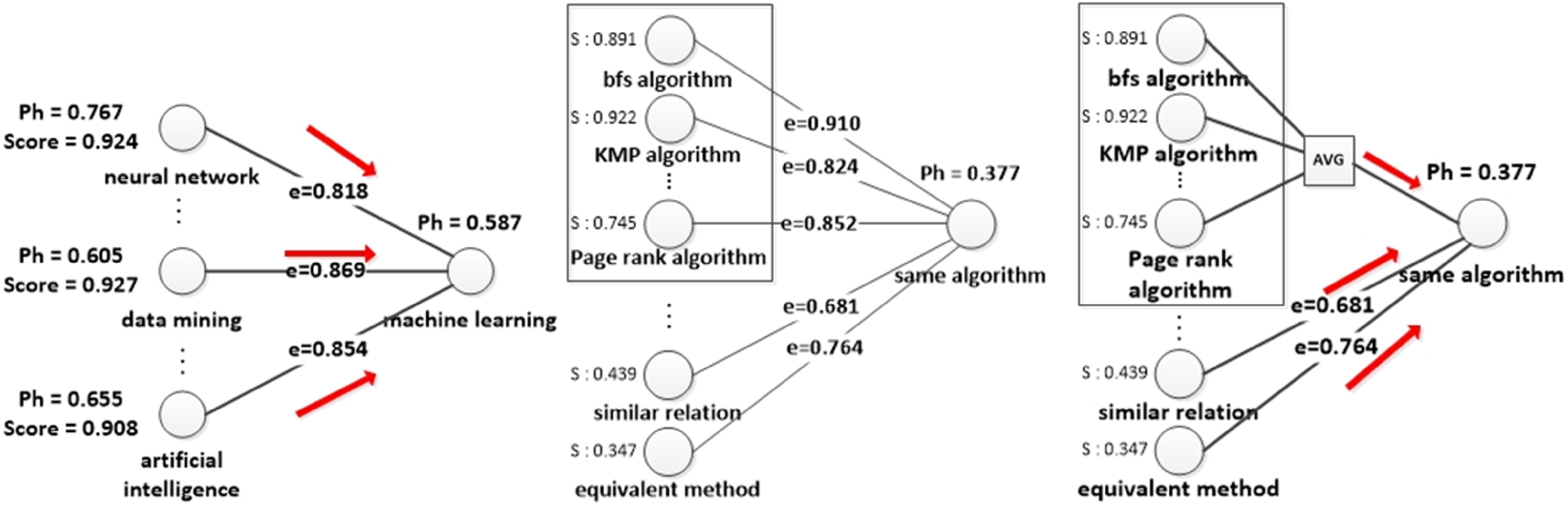
- V is the **vertex set**: Each vertex in V represents a **candidate course concept**, associated with a **phraseness score**.
- E is the **edge set**: For an edge $(c_i, c_j) \in E$, its edge weight $e(c_i, c_j) = SR(c_i, c_j)$
  $SR(c_i, c_j)$ indicates the **semantic relatedness** between c_i and c_j, i.e., the likeness of their semantic meaning.

- **Pruning**: An edge (c_i, c_j) exists in a CCG only if $SR(c_i, c_j) > \theta$.

Ranking

Assumption: In CCG, a course concept is likely to connect with other course concepts with high semantic relatedness

General Idea: Based on a small *seed set* to find more course concepts in CCG using a *graph-based propagation algorithm*.



Ranking

Propagation Process:

$$conf^{k+1}(c_i) = \frac{1}{Z}\left(\frac{\sum_{c_j \in A(c_i)} vs^k(c_j, c_i)}{|A(c_i)|}\right)$$

Voting Score: It determines how much score should a vertex receives from another vertex in each iteration.

$$vs^k(c_j, c_i) = ph(c_j) \cdot e(c_i, c_j) \cdot conf^k(c_j)$$

Generalized Voting Score: $opf(c_i, c_j) = \lambda$ if c_i and c_j are overlapping.

$$gvs^k(c_j, c_i) = opf(c_i, c_j) \cdot ph(c_j) \cdot e(c_i, c_j) \cdot conf^k(c_j)$$

Outline

- Backgrounds
- Related Works
- Methods
- **Experiments and Analysis**
- Conclusion

Experiments

Datasets

Dataset	Domain	Language	#courses	#videos	#tokens	#candidates	#labeled	correlation
CSEN	Computer Science	English	8	690	1,242,156	59,050	4,096	0.734
EcoEN	Economics	English	5	381	401,192	27,571	3,652	0.696
CSZH	Computer Science	Chinese	18	2,849	2,291,258	79,009	5,309	0.721
EcoZH	Economics	Chinese	8	455	645,016	60,566	3,663	0.646

Metrics

- **R-precision**
- **MAP (Mean Average Precision)**

Baselines

- **Statistical-based Methods** (TF-IDF，PMI)
- **Graph-based Methods** (TextRank，Topical PageRank)

Experiments

Experimental Results

- **Our method** outperforms all baselines on all datasets
- **TF-IDF & TextRank** perform worse than TPR and CCP
- **TPR** performs better than TextRank across all datasets

Method		CSEN	EcoEN	CSZH	EcoZH
TF-IDF	R_p	0.125	0.303	0.118	0.198
	MAP	0.105	0.232	0.109	0.145
PMI	R_p	0.239	0.222	0.246	0.179
	MAP	0.141	0.197	0.187	0.121
TextRank	R_p	0.151	0.290	0.142	0.161
	MAP	0.137	0.263	0.131	0.115
TPR	R_p	0.284	0.414	0.305	0.303
	MAP	0.255	**0.387**	0.267	0.288
CCP	R_p	**0.443**	**0.427**	**0.434**	**0.435**
	MAP	**0.432**	0.365	**0.416**	**0.423**

Outline

- Backgrounds
- Related Works
- Methods
- Experiments and Analysis
- **Conclusion**

Conclusion

Conclusion

- Automatically discovering course concepts in MOOCs

Future Directions

- Research on automatically course concept map generation
- Try deep learning models for course concept extraction
- Incorporating dynamic information in MOOCs (e.g., user behavior, forums, QA between students and teachers).

Thanks!

Liangming Pan
KEG, THU
peterpan10211020@163.com