Prerequisite Relation Learning for Concepts in MOOCs

Reporter: Liangming PAN

Authors: Liangming PAN, Chengjiang LI, Juanzi LI, Jie TANG

Knowledge Engineering Group

Tsinghua University

2017-04-19

Outline

- Backgrounds
- Problem Definition
- Methods
- Experiments and Analysis
- Conclusion

Backgrounds

What?

Prerequisite Relation Learning for Concepts in MOOCs

Backgrounds

What? **Prerequisite Relation Learning for Concepts in 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.

coursera edX UDACITY

Backgrounds

What?

Prerequisite Relation Learning for Concepts in 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.

学堂在线 xuetangx.com

慕课网 imooc.com

中国大学MOOC

Backgrounds

What?

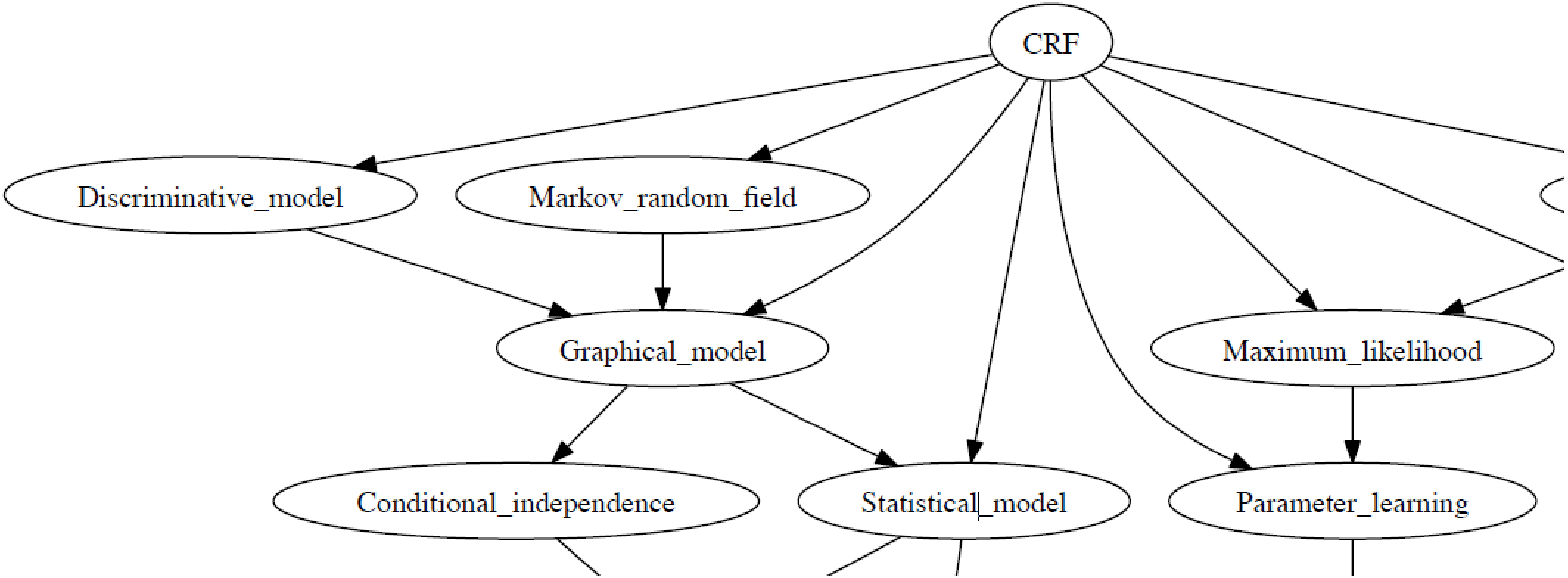
Prerequisite Relation Learning for Concepts in MOOCs

- A ***prerequisite*** is usually a concept or requirement before one can proceed to a following one.
- The prerequisite relation exists as a natural dependency among concepts in cognitive processes when people learn, organize, apply, and generate knowledge (Laurence and Margolis, 1999).

Backgrounds

What?

Prerequisite Relation Learning for Concepts in MOOCs



Partha Pratim Talukdar and William W Cohen. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. 2012.

Backgrounds

WHY?

Prerequisite Relation Learning for Concepts in MOOCs

Motivation 1. Manually building a concept map in MOOCs is infeasible

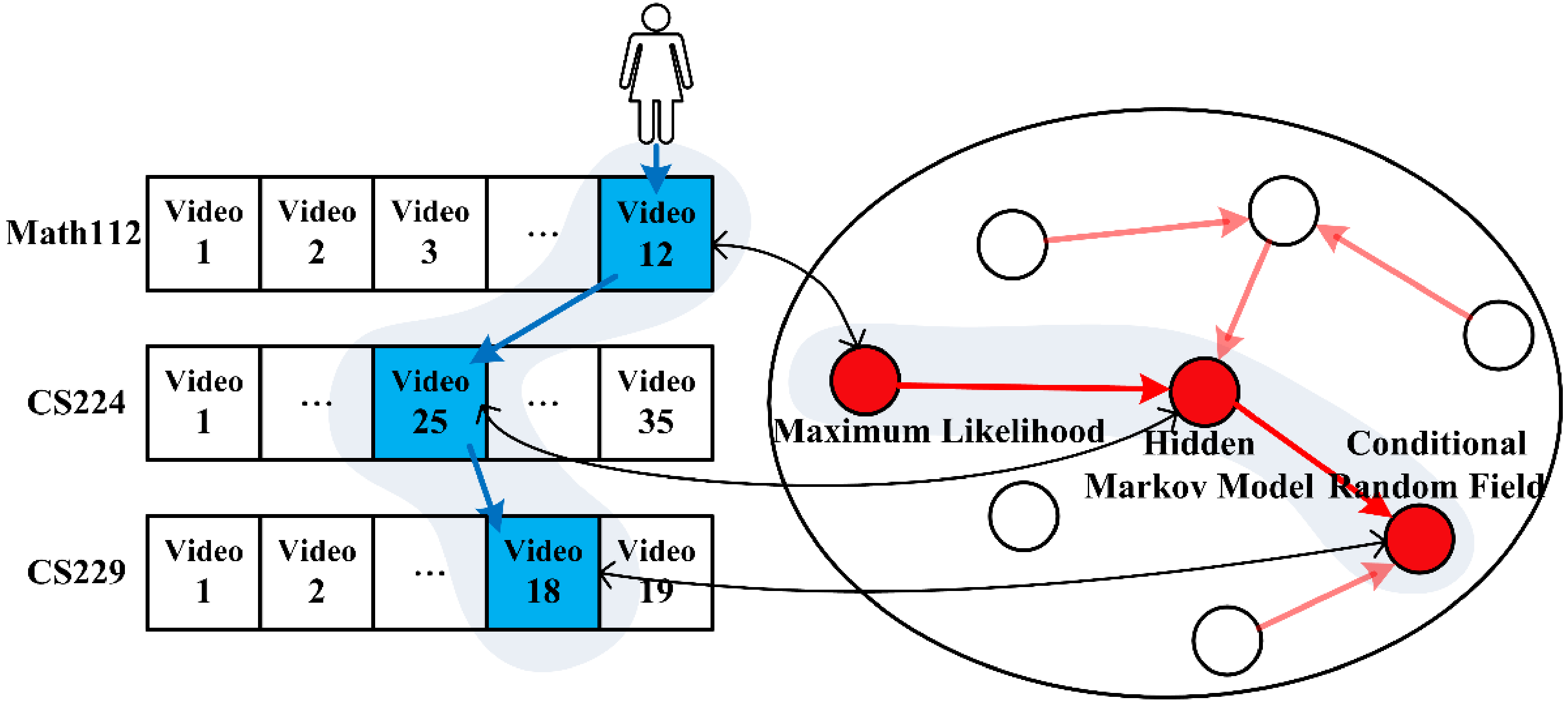
- In the era of MOOCs, it is becoming infeasible to manually organize the knowledge structures with thousands of online courses from different providers.

Motivation 2. To help improve the learning experience of students

- The students from different background can easily explore the knowledge space and better design their personalized learning schedule.

Backgrounds

Question: What should she get started if she wants to learn the concept of "conditional random field"?



Outline

- Backgrounds
- **Problem Definition**
- Methods
- Experiments and Analysis
- Conclusion

Problem Definition

- Input
  - ***MOOC Corpus*** $\mathcal{D} = \{\mathcal{C}_1, \cdots, \mathcal{C}_i, \cdots, \mathcal{C}_n\}$, *where* C_i *is one course*

    Course $\mathcal{C} = (\mathcal{V}_1, \cdots, \mathcal{V}_i, \cdots, \mathcal{V}_{|\mathcal{C}|})$ *, where* v_i *is the i-th **video** of course* C

    Video $\mathcal{V} = (s_1 \cdots s_i \cdots s_{|\mathcal{V}|})$ *, where* s_i *is the i-th **sentence** of video* v
  - ***Course Concepts*** $\mathcal{K} = \mathcal{K}_1 \cup \cdots \cup \mathcal{K}_n$ *, where* K_i *is the set of course concepts in* C_i

- Output
  - ***Prerequisite Function***

$$PF(a,b) \in \{0,1\},\ a,b \in \mathcal{K}$$

The function PF predicts whether concept a is a prerequisite concept of b

Outline

- Backgrounds
- Problem Definition
- **Methods**
- Experiments and Analysis
- Conclusion

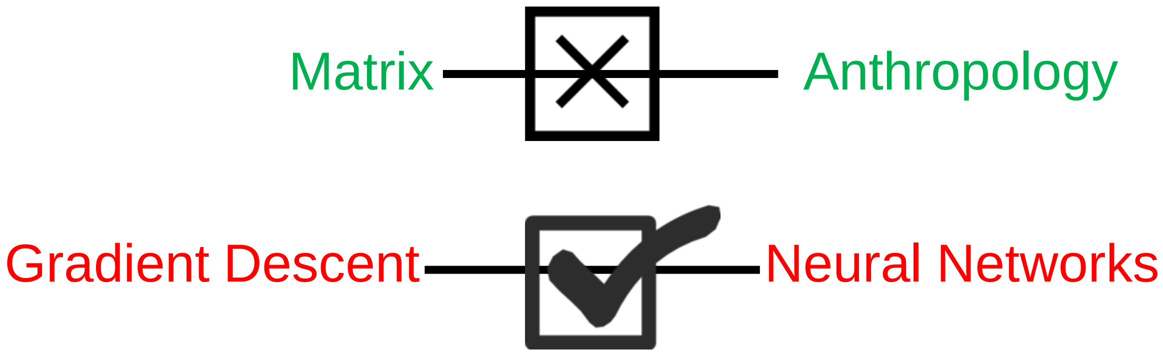
Features Overview

- **Features**
  - Semantic Features
    - Semantic Relatedness
  - Contextual Features
    - Video Reference Distance
    - Sentence Reference Distance
    - Wikipedia Reference Distance
  - Structural Features
    - Average Position Distance
    - Distributional Asymmetry Distance
    - Complexity Level Distance

Semantic Features

Features — Semantic Features — Semantic Relatedness

- Semantic Relatedness plays an important role in prerequisite relations between concepts.
- If two concepts have *very different semantic meanings*, it is *unlikely* that they have prerequisite relations.



Semantic Features

- Concept Embeddings
  - ***Wikipedia corpus***

$$OE = \langle w_1 \cdots w_i \cdots w_m \rangle$$

  - ***Procedure of Concept Embeddings***
    - 1. Entity Annotation: We label all the entities in the Wikipedia corpus based on the hyperlinks in Wiki, and get a new corpus OE' and a wiki entity set ES.

$$OE' = \langle x_1 \cdots x_i \cdots x_{m'} \rangle$$

$$ES = \{ e_1 \cdots e_i \cdots e_w \}$$

Where x_i corresponds to a word $w \in OE$ or an entity $e \in ES$

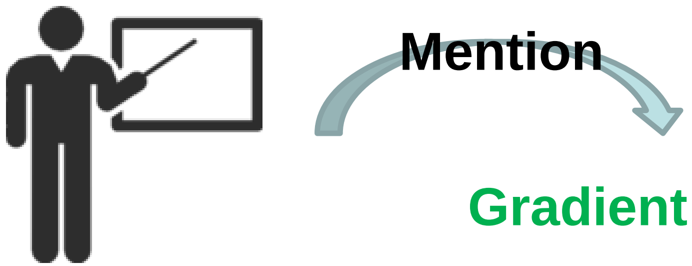
    - 2. Word Embeddings: We apply the skip-gram model to train word embeddings on OE'.
    - 3. Concept Representation: After training, we can obtain the vector for each concept in ES. For any non-wiki concept, we obtain its vector via the vector addition of its individual word vectors.

Contextual Features

Features — **Contextual Features** — **Video Reference Distance**

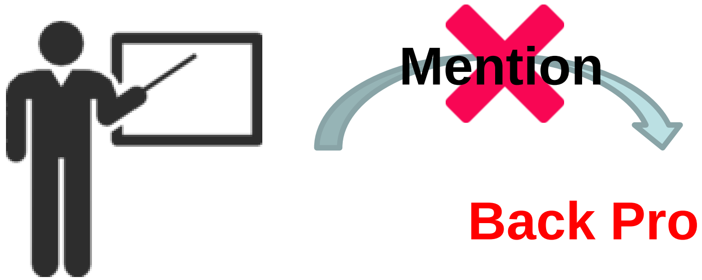
- If in videos where *concept A is frequently talked about*, the teacher *also needs to refer to concept B for a lot* but not vice versa, then B would more likely be a prerequisite of A.

Back Propagation



Gradient Descent

Gradient Descent



Back Propagation

Contextual Features

- Video Reference Distance
  - ***Video Set of the MOOC corpus***

$$V^D = V_1 \cup \cdots V_n$$

  - ***Video Reference Weight from A to B***

$$Vrw(A,B) = \frac{\sum_{v \in V^D} f(A,v) \cdot r(v,B)}{\sum_{v \in V^D} f(A,v)}$$

Where

- $f(A,v)$: the term frequency of concept A in video v
- $r(v,B) \in \{0,1\}$: whether concept B appears in video v
- It indicates how B is referred by A's videos

  - ***Video Reference Distance of (A,B)***

$$Vrd(A,B) = Vrw(B,A) - Vrw(A,B)$$

Contextual Features

- Generalized Video Reference Distance
  - ***Generalized Video Reference Weight from A to B***

$$GVrd(A,B) = \frac{\sum_{i=1}^{K} Vrw(a_i,B) \cdot w(a_i,A)}{\sum_{i=1}^{K} w(a_i,A)}$$

Where

- $\{a_1,\cdots,a_K\}$: the top-K most similar concepts of A, where $a_1,\cdots,a_K \in T$
- $w(a_i,A)$: the similarity between a_i and A
- It indicates how B is referred by A's related concepts in their videos

  - ***Generalized Video Reference Distance of (A,B)***

$$GVrd(A,B) = GVrw(B,A) - GVrw(A,B)$$

Contextual Features

- **Features**
  - Semantic Features
    - Semantic Relatedness
  - Contextual Features
    - Video Reference Distance
    - Sentence Reference Distance
    - Wikipedia Reference Distance
  - Structural Features
    - Average Position Distance
    - Distributional Asymmetry Distance
    - Complexity Level Distance

Structural Features

- **Features**
  - **Structural Features**
    - **Average Position Distance**
    - **Complexity Level Distance**
    - **Distributional Asymmetry Distance**

- In teaching videos, knowledge concepts are usually introduced based on their learning dependencies, so the structure of MOOC courses also significantly contribute to prerequisite relation inference in MOOCs.

- We investigate 3 different structural information, including *appearing positions of concepts*, *learning dependencies of videos* and *complexity levels of concepts*.

Structural Features

- Average Position Distance
  - ***Assumption***
    - In a course, for a specific concept, its prerequisite concepts tend to be introduced before this concept and its subsequent concepts tend to be introduced after this concept.
  - ***TOC Distance of (A,B)***

$$Apd(A,B) = \begin{cases} \frac{1}{|C(A,B)|} \sum_{C \in C(A,B)} (AP(A,C) - AP(B,C)) & , C(A,B) \neq \emptyset \\ 0 & , C(A,B) = \emptyset \end{cases}$$

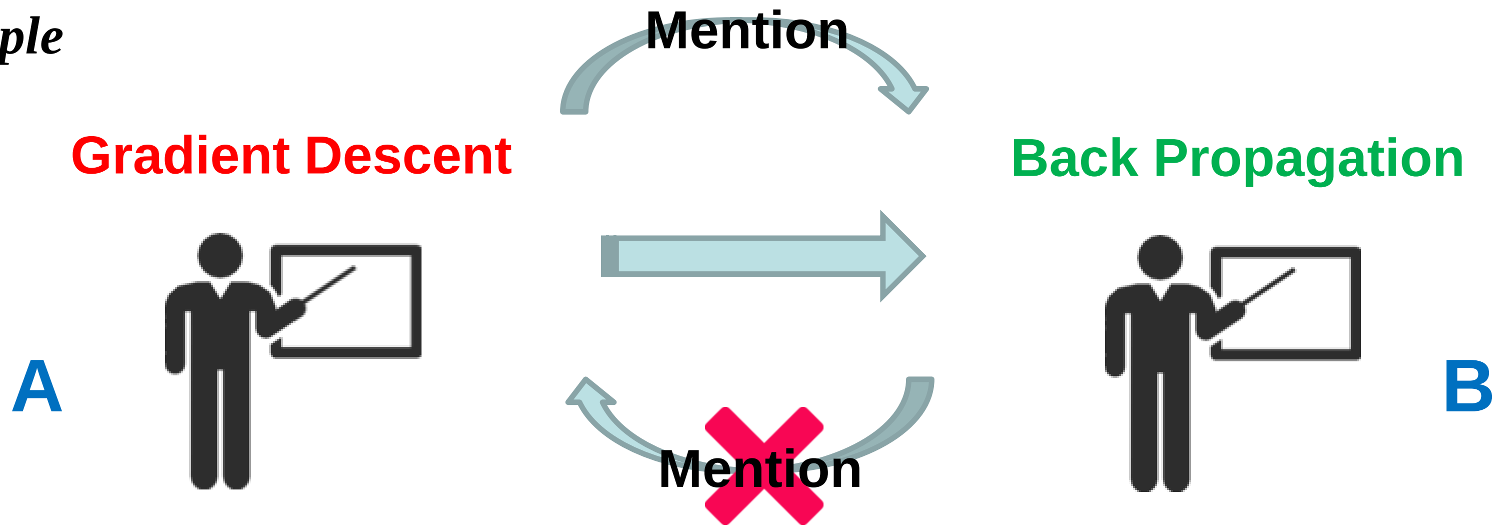
Where

- $C(A,B)$: the set of courses in which A and B both appear
- $AP(A,C)$ = the average index of videos containing concept A in course C

(***The average position of a concept A in course C)***

Structural Features

- Distributional Asymmetry Distance
  - ***Assumption***
    - The learning dependency of course videos is also helpful to infer learning dependency of course concepts.
    - Specifically, if video V_a is a precursor video of V_b, and a is a prerequisite concept of b, then it is likely that f(b, V_a) < f(a, V_b)
  - ***Example***



Structural Features

- Distributional Asymmetry Distance
  - ***All possible video pairs of*** $\langle a, b \rangle$ ***that have sequential relation***

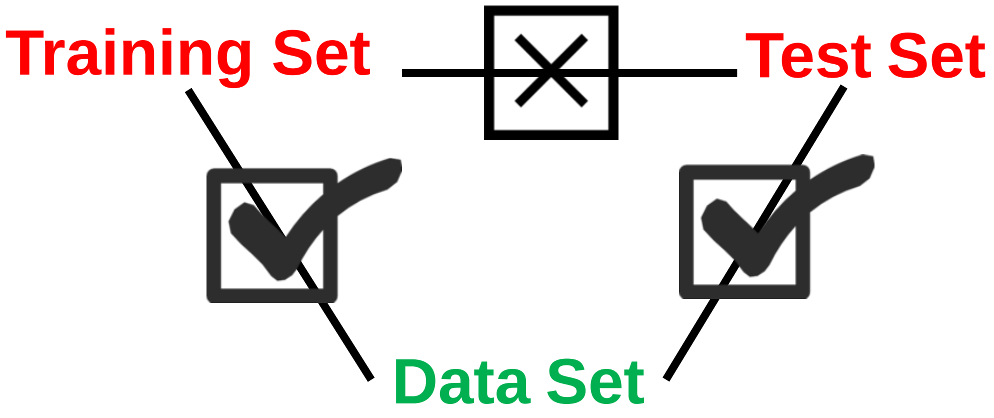
$$\mathcal{S}(\mathcal{C}) = \{(i,j) | i \in \mathcal{I}(\mathcal{C}, a), j \in \mathcal{I}(\mathcal{C}, b), i < j\}$$

  - ***Distributional Asymmetry Distance***

$$Dad(a,b) = \frac{\displaystyle\sum_{\mathcal{C} \in \mathcal{C}(a) \cap \mathcal{C}(b)} \frac{\displaystyle\sum_{(i,j) \in \mathcal{S}(\mathcal{C})} f\left(a, \mathcal{V}_i^{\mathcal{C}}\right) - f\left(b, \mathcal{V}_j^{\mathcal{C}}\right)}{|\mathcal{S}(\mathcal{C})|}}{|\mathcal{C}(a) \cap \mathcal{C}(b)|}$$

Structural Features

- Complexity Level Distance
  - ***Assumption***
    - If two related concepts have prerequisite relationship, they may have a difference in their complexity level. It means that one concept is more ***basic*** while another one is more ***advanced***.
  - ***Example***



Structural Features

Complexity Level Distance

- ***Assumption***
  - For a specific concept, if it **covers more videos** in the course or it **survives longer time** in a course, then it is more likely to be a general concept rather than a specific concept.
- ***Average video coverage of A***

$$AVC(A) = \frac{1}{C(A)} \sum_{C \in C(A)} \frac{vc(A)}{m_C}$$

- ***Average survival time of A***

$$AST(A) = \frac{1}{C(A)} \sum_{C \in C(A)} \frac{LI(A) - FI(A) + 1}{m_C}$$

- ***Complexity Level Distance of (A,B)***

$$Cld(A,B) = AVC(A) \cdot AST(A) - AVC(B) \cdot AST(B)$$

Outline

- Backgrounds
- Problem Definition
- Methods
- Experiments and Analysis
- Conclusion

Experimental Datasets

Collecting Course Videos

- “Machine Learning” (ML), “Data Structure and Algorithms” (DSA), and “Calculus” (CAL) from Coursera

Course Concepts Annotation

- Extract candidate concepts from documents of video subtitles Label the candidates as “course concept” or “not course concept”

Prerequisite Relation Annotation

- We manually annotate the prerequisite relations among the labeled course concepts.

Experimental Datasets

- Dataset Statistics
  - 3 novel datasets extracted from Coursera
    - **ML:** 5 *Machine Learning* courses
    - **DSA:** 8 *Data Structure and Algorithms* courses
    - **CAL:** 7 Calculus courses

Dataset	#courses	#videos	#concepts	#pairs		κ
				$-$	$+$	
ML	5	548	244	5,676	1,735	0.63
DSA	8	449	201	3,877	1,148	0.65
CAL	7	359	128	1,411	621	0.59

Evaluation Results

- Models
  - Naïve Bayes (NB)
  - Logistic Regression (LR)
  - SVM with linear kernel (SVM)
  - Random Forest (RF)
- Metrics
  - Precision (P)
  - Recall (R)
  - F1-Score (F1)
- 5-Fold Cross Validation

Classifier		ML		DSA		CAL	
	M	1	10	1	10	1	10
SVM	P	63.2	60.1	60.7	62.3	61.1	61.9
	R	68.5	72.4	**69.3**	67.5	**67.9**	68.3
	F_1	65.8	65.7	64.7	64.8	64.3	64.9
NB	P	58.0	58.2	62.9	62.6	60.1	60.6
	R	58.1	60.5	62.3	61.8	61.2	62.1
	F_1	58.1	59.4	62.6	62.2	60.6	61.3
LR	P	66.8	67.6	63.1	62.0	62.7	63.3
	R	60.8	61.0	64.8	66.8	63.6	64.1
	F_1	63.7	64.2	63.9	64.3	61.6	62.9
RF	P	**68.1**	**71.4**	**69.1**	**72.7**	**67.3**	**70.3**
	R	**70.0**	**73.8**	68.4	**72.3**	67.8	**71.9**
	F_1	**69.1**	**72.6**	**68.7**	**72.5**	**67.5**	**71.1**

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

Comparison with Baselines

- Comparison Methods
  - **Hyponym Pattern Method (HPM)**
    - This method simply treat the concept pairs with IS-A relations as prerequisite concept pairs.
  - **Reference Distance (RD)**
    - This method was proposed by Liang et al. (2015). However, this method is only applicable to Wikipedia concepts.
  - **Supervised Relationship Identification (SRI)**
    - Wang et al. (2016) has employed several features to infer prerequisite relations of Wikipedia concepts in textbooks, including 3 Textbook features and 6 Wikipedia features.
    - (1) **T-SRI:** only textbook features are used to train the classifier.
    - (2) **F-SRI:** the original version, all features are used.

Comparison with Baselines

- W-ML, W-DSA, W-CAL are subsets with Wikipedia Concepts
- HPM achieves relatively high precision but low recall.
- T-SRI only considers relatively simple features
- Incorporating Wikipedia-based features achieves certain promotion in performance

Method		ML	DSA	CAL	W-ML	W-DSA	W-CAL
HPM	P	67.3	71.4	69.5	**79.9**	72.3	73.5
	R	18.4	14.8	16.5	25.5	27.3	23.3
	F_1	29.0	24.5	26.7	38.6	39.6	35.4
RD	P	–	–	–	73.4	**77.8**	**74.4**
	R	–	–	–	42.8	44.8	43.1
	F_1	–	–	–	54.1	56.8	54.6
T-SRI	P	61.4	62.3	62.5	58.1	60.1	62.7
	R	62.9	64.6	65.5	67.6	65.3	67.9
	F_1	62.1	63.4	64.0	62.5	62.6	65.2
F-SRI	P	–	–	–	64.3	64.3	64.8
	R	–	–	–	62.1	65.6	65.2
	F_1	–	–	–	63.2	64.9	65.0
MOOC	P	**71.4**	**72.7**	**70.3**	72.8	68.4	71.4
	R	**73.8**	**72.3**	**71.9**	**71.3**	**72.0**	**70.8**
	F_1	**72.6**	**72.5**	**71.1**	**72.0**	**70.2**	**71.1**

Table 3: Comparison with baselines(%).

Comparison with Baselines

- Setting
  - Each time, one feature or one group of features is removed
  - We record the decrease of F1-score for each setting

	Ignored Feature(s)	P	R	F_1
Single	Sr	69.6	72.9	71.2(**-1.4**)
	GVrd	68.8	71.4	70.1(-2.5)
	GSrd	67.9	71.4	69.6(-3.0)
	Wrd	70.1	72.1	71.1(-1.5)
	Apd	69.7	70.8	70.2(-2.4)
	Dad	69.2	69.5	69.4(-3.2)
	Cld	64.9	65.6	65.2(**-7.4**)
Group	Semantic	69.6	72.9	71.2(**-1.4**)
	Contextual	66.4	68.9	67.6(-5.0)
	Structural	63.7	64.2	63.4(**-9.2**)

Table 4: Contribution analysis of different features(%).

- Conclusion
  - All the proposed features are useful
  - **Complexity Level Distance** is most important
  - **Semantic Relatedness** is least important

Outline

- Backgrounds
- Problem Definition
- Methods
- Experiments and Analysis
- Conclusion

Thanks!

Liangming Pan
KEG, THU
peterpan10211020@163.com