DATA@KU
고려대학교 디지털정보처

2021 고려대학교 대학혁신지원사업

KU-Insight Miner AI선배 강의 추천

Knowledge Graph를 활용한 R-GCN기반 제2전공 추천 시스템 개발

고려대학교 디지털정보처 데이터 hub팀

데이터 사이언티스트 이진숙

2021.07.16

R−GCN[1]

Relational Graph Convolutional Network

$$h_i^{l+1} = \sigma\left(\sum_{m \in \mathcal{M}_i} g_m(h_i^l, h_j^l)\right)$$

where

$h_i^l \in \mathbb{R}^{d^l}$: *hidden state of lth layer*

d^l : *dimension of l layer*

$\sigma(.)$: *element − wise activation function like ReLU*

$\mathcal{M}_i$: *set of incoming messages for node vi n(often identical to the set of incoming edges)*

$g_{m(.,.)}$: *accumulated and passed through dimension of l layer*

∗ note: *message specific neural network or simply a linear transformation* $g_{m(h_i, hj)} = Wh_j$

W weight matrix like [2] suggested

$Graph: G = (V, \mathcal{E}, R)$

$Node: v_i \in V$

$Relationship\ type: r \in R$

Edges : $(v_i, r, v_j) \in \mathcal{E}$

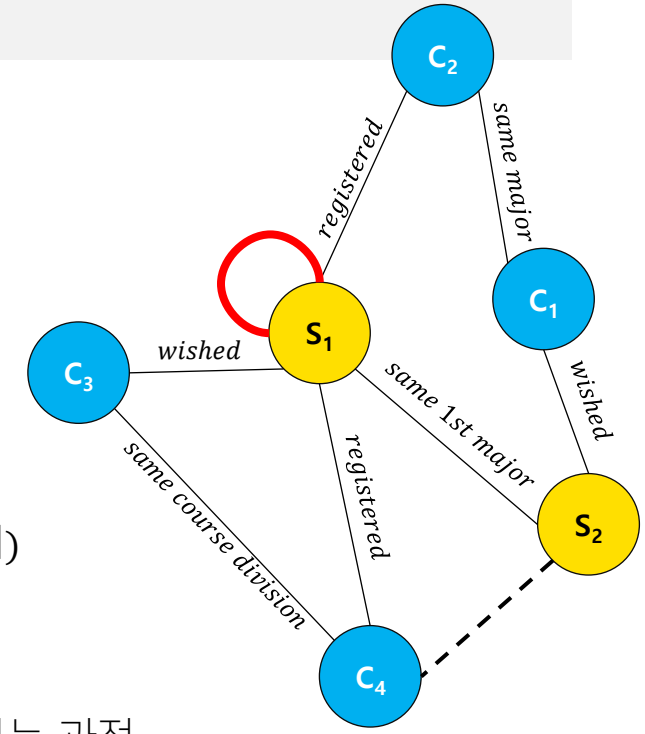
[1]Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In *European semantic web*

R-GCN[1]

Forward-pass update an entity or node denoted by v_i in a relational + directed and labeled multi-graph

Propagation model

$$h_i^{l+1} = \sigma\left(\underbrace{W_0^l h_i^l}_{self-loop} + \underbrace{\sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^l h_j^l}_{Neighbors}\right)$$



where

N_i^r : *set of neighbor indices of node i under relation* $r \in R$

$c_{i,r}$: *a problem − specific normalization* (*it can be learned or chosen in advance such as* $c_{i,r} = |N_i|$)

Neighbor들로부터 전달된 feature vector들을 normalized summation하는 과정

기존의 GCN과는 다르게 relation-specific하게 transformation(엣지의 타입이나 방향성을 고려함)

[1]Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In *European semantic web*

R−GCN[1]

Two ways of regularization

$$h_i^{l+1} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}} W_r^l h_j^l + W_0^l h_i^l\right)$$

Rare relationship에서는 overfitting 위험이 있어 Regularization 진행(Reducing parameters)

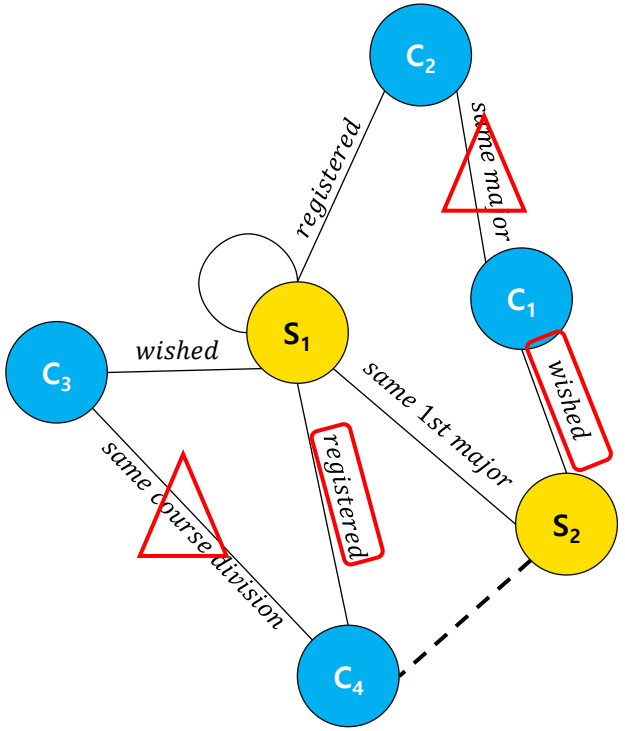
(1) Basis decomposition(linear combination of the number of components)

$$W_r^l = \sum_{b=1}^{B} a_{rb}^l V_b^l$$

$$V_b^l \in \mathbb{R}^{d^{l+1}\times d^l} \; with \; coefficients \; a_{rb}^l$$

$specifying\ the\ numbers\ of\ unique\ Ws\ that\ you\ want\ to\ have\ for\ the\ layer \rightarrow B$

$each\ of\ the\ Wr\ s\ is\ calculated\ by\ combining\ those\ components\ linearly$
$so\ they\ learn\ a\ coefficient\ for\ each\ of\ the\ components$



[1]Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In *European semantic web*

R−GCN[1]

Two ways of regularization

$$h_i^{l+1} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}} W_r^l h_j^l + W_0^l h_i^l\right)$$

Rare relationship에서는 overfitting 위험이 있어 Regularization 진행(Reducing parameters)

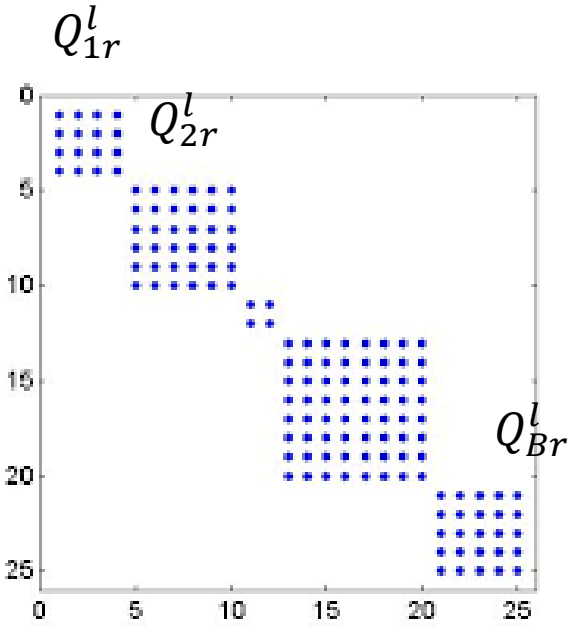
(2) Block-diagonal decomposition

$$W_r^l = \oplus_{b=1}^{B} Q_{br}^l$$
$$= diag(Q_{1r}^l, \ldots, Q_{br}^l)$$

$$Q_{br}^l \in \mathbb{R}^{(d^{l+1}/B)\times(d^l/B)}$$

defined through direct sum over a set of low dimensional matrices

Variables are strongly interconnected within a group but don't have much interactions outside of the group



[1]Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In *European semantic web*

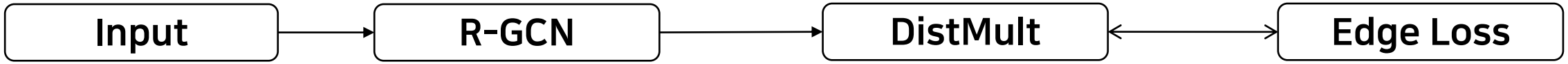
R-GCN[1]

Link Prediction Task Flow

$Graph: G = (V, \mathcal{E}, R)$
$Node: v_i \in V$
$Relationship\ type: r \in R$
Edges: $(v_i, r, v_j) \in \mathcal{E}$

Link Prediction Task Flow

$\mathcal{E}$ 대신에 불완전한 링크 $\hat{\mathcal{E}}$ 제공 , $f(s,r,o)$ 를 실제 (s,r,o)에 가깝게 예측



Encoder

$$v_i \in V \longrightarrow e_i \in \mathbb{R}^d$$
$$e_i = h_i^{(L)}$$

Decoder

$scoring\ (s,r,o)$
$s: \mathbb{R}^d \times R \times \mathbb{R}^d \rightarrow \mathbb{R}$
$f(s,r,o) = e_i^T R_r e_o$

Training

Sampling

$Negative\ Sampling\ (\omega)$

Optimization

$Cross\ entropy$

$$\mathcal{L} = -\frac{1}{(1+\omega)|\hat{\mathcal{E}}|} \sum_{(s,r,o,y)\in\mathcal{T}} ylog\big(f(s,r,o)\big) + (1-y)\log(1 - l\big(f(s,r,o)\big))$$

$\mathcal{T}: total\ set\ of\ real\ and\ corrupted\ triples$
$l: logistic\ sigmoid\ function$
$y: y = 1\ for\ positive\ triples\ and\ y = 0\ for\ negative\ ones$

[1]Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In *European semantic web*

Knowledge Graph 정의 및 데이터

14학번~21학번의 수강이력, 관심과목등록, 학적정보, 제1전공, 2전공 커리큘럼

이용 데이터

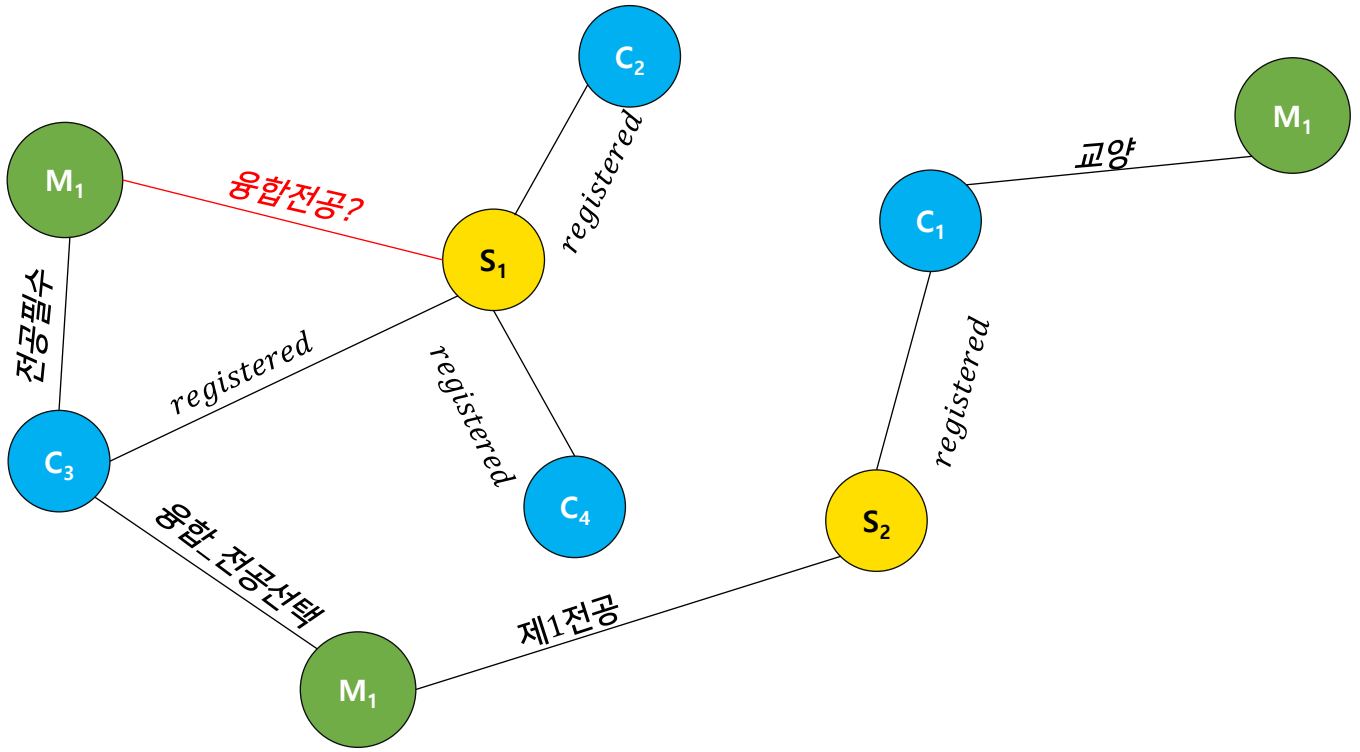
1. 수강이력
2. 기준 1) 2014~2021학번 서울캠퍼스학생
기준 2) 2014~2021년도 수강이력

2. 제1전공, 제2전공 내역
3. 학과, 융합전공, 학생설계전공 커리큘럼

Knowledge Graph

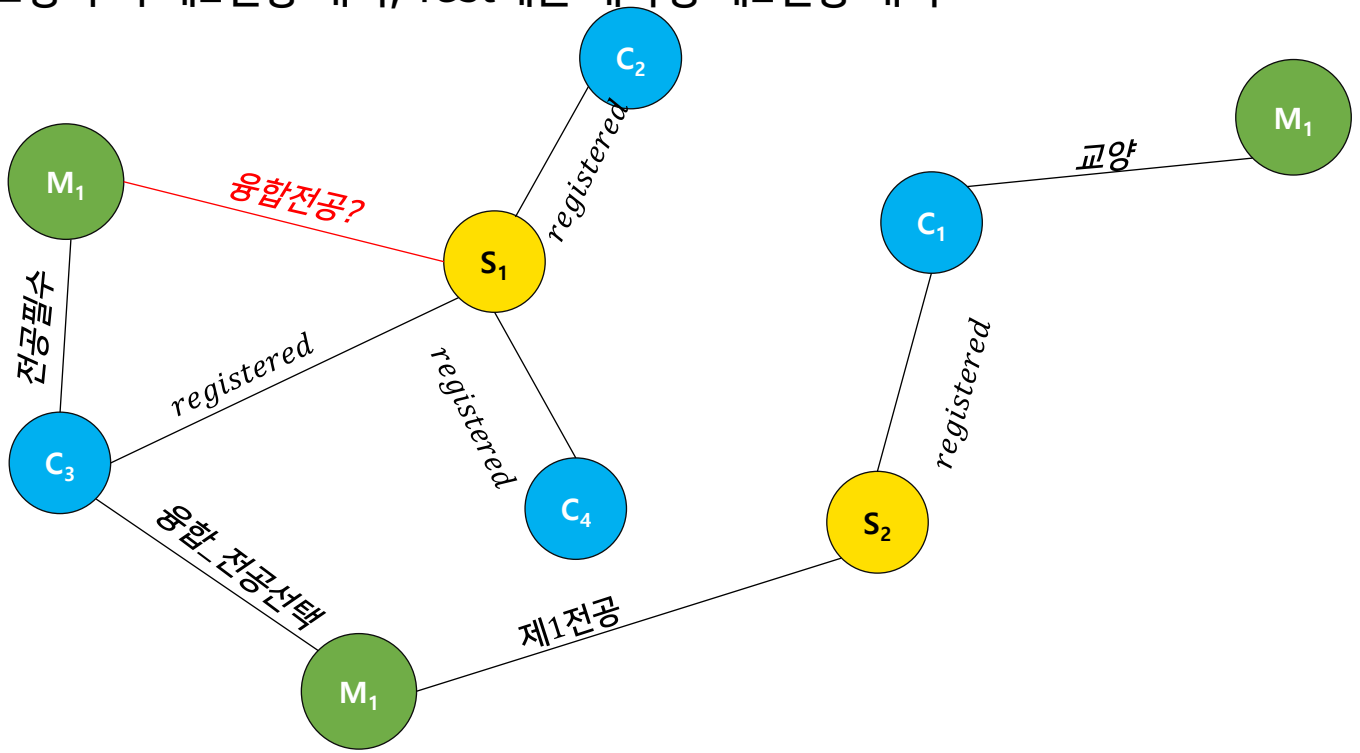
1. Vertex 정의
- 학생
- 강의
- 학과, 제2전공

2. Relation 정의
- 수강
- 전공, 제2전공 이수
- 어떤 이수구분으로 커리큘럼 포함



* IMPORTANT * *Train and test set split*

Validation에는 졸업생 및 수료생의 의 제2전공 내역, Test에는 재학생 제2전공 내역



*교양의 경우에 학과는 교양교육원

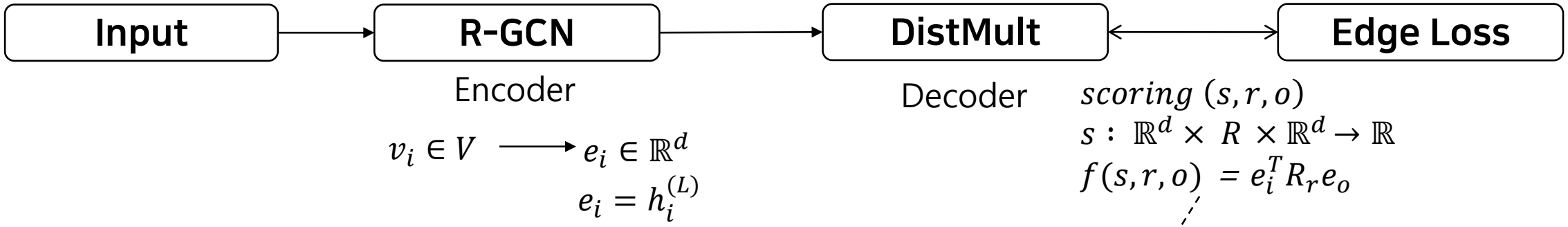
Train triplets(*N* = 1,410,295)	Valid triplets (*N* = 7,409)	Test triplets (*N* = 26,937)
1. 수강이력 기준 1) 2014~2021학번 서울캠퍼스학생 기준 2) 2014~2021년도 수강이력	**제2전공 내역(졸업, 수료)** 기준 1) 2014~2021학번 서울캠퍼스학생	**제2전공 내역(재학, 휴학생)** 기준 1) 2014~2021학번 서울캠퍼스학생
2. 제1전공, 제2전공 내역		
3. 학과, 이중, 융합, 학생설계 커리큘럼		

이용모델: R-GCN

Subject Perturbation과 Object perturbation

Link Prediction Task Flow

$\mathcal{E}$ 대신에 불완전한 링크 $\hat{\mathcal{E}}$ 제공 , $f(s,r,o)$ 를 실제 (s,r,o)에 가깝게 예측



DistMult

Relation Embedding의 한 방법으로 가중치의 대각행렬을 이용하여 score계산

$\mathrm{Merkel}: \mathrm{h} = (1,0)^{\mathrm{T}}$

$\mathrm{Germany}: \mathrm{t} = (1,0)^{\mathrm{T}}$

$\mathrm{Is\_leader}: \mathrm{W_r} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

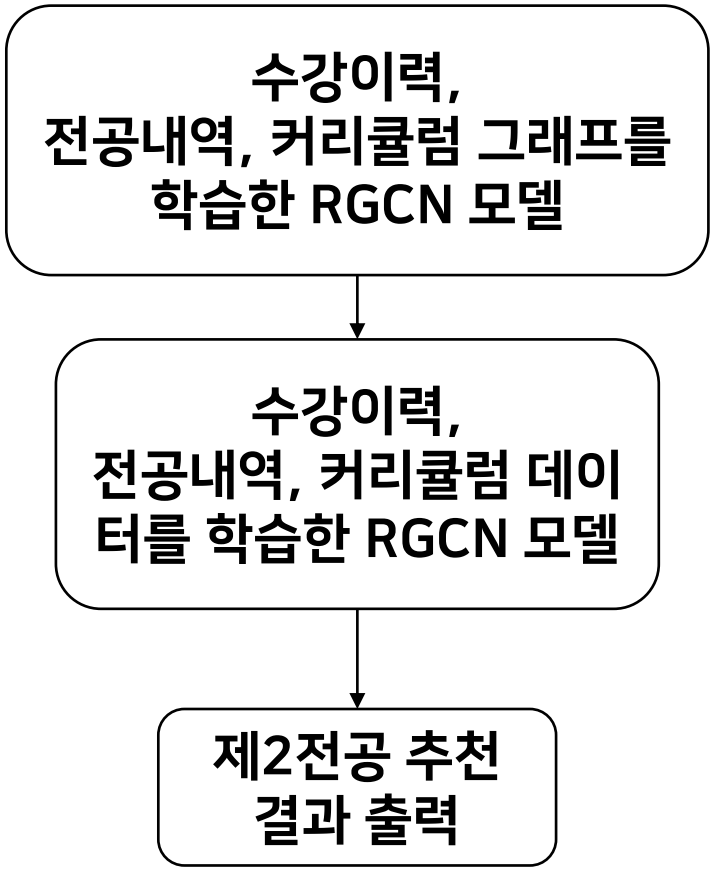
$\mathrm{Score} = \mathrm{h}^{\mathrm{T}}\mathrm{W_r}\mathrm{t}$

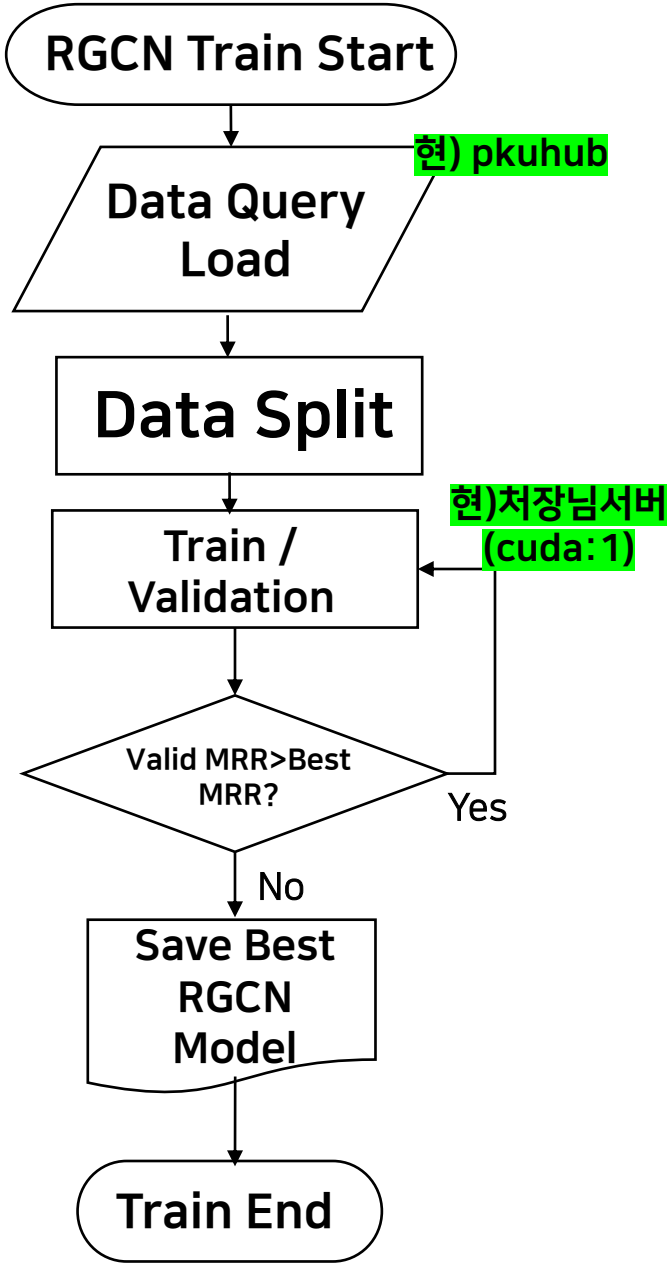
$= \sum(\mathrm{h} \odot \mathrm{t}) \odot \mathrm{diag}(\mathrm{W_r})$

추천 Flow

Object perturbation score만 사용

Recommendation Task Flow





Result 1. Metric & Hyperparameter setting

Hit@n, Recall@n, MRR

MRR(Mean Reciprocal Rank)

$$MRR = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{1}{rank_i}$$

Recall

$$Recall = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{|courses_i \cap recommended_i|}{|courses_i|}$$

Hit@1,3,10

N개 안에 포함

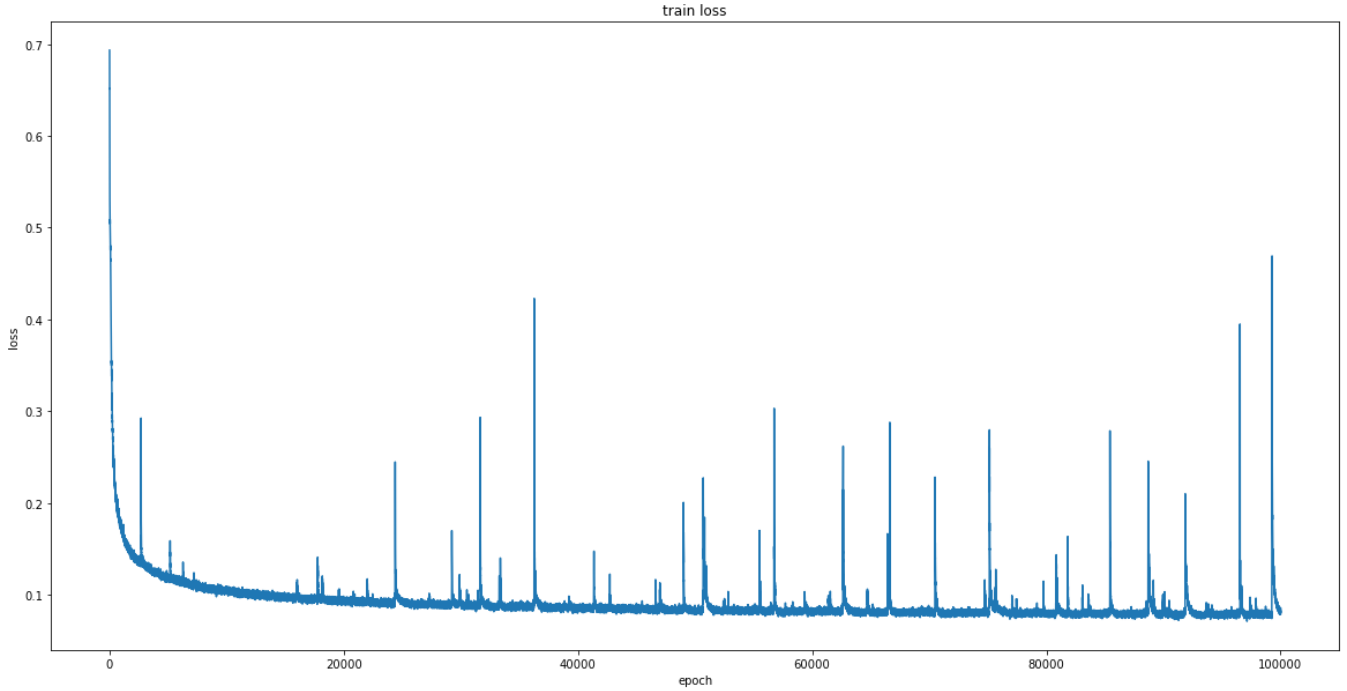
Epoch	Dropout	LR	n-bases	Negative Sample	Composition
~100000	0.2	0.01	4	o	Basic

Evaluate every 1000

Loss 대신에 Valid MRR로 Best epoch 설정

Experiment 1. Hit@10의 경우 98% 정확도

Hit@n, MRR



num_entity: 43389
num_relation: 17
num_train_triples: 1410295
num_valid_triples: 7409
num_test_triples: 26937

Best Model Description(Evaluated every 1000 epoch)

Epoch	Dropout	LR	n-bases	Train Loss	Valid MRR	Test MRR	Hit@1	Hit@3	Hit@10
90000	0.2	1e-2	4	0.1378	0.7473	0.8104	0.687	0.928	0.984

Example 1. 제2전공 추천

훈련셋에 들어가지 않은 데이터 테스트, top10개 추천 리스트 출력

mmajor_nm	score
바이오의공학부	0.994280
의과학융합전공	0.576477
뇌인지과학융합전공	0.492002
경제통계학부 빅데이터전공	0.489575
인공지능융합전공	0.445383
전기전자공학부	0.399804
식품영양학과	0.346129
메디컬융합공학융합전공	0.254796
신소재공학부	0.199610
언어학과	0.079861

1전공 바이오의공학부,
2전공 바이오의공학부
심화전공 학생

mmajor_nm	score
사학과	0.912904
중어중문학과	0.673528
정치외교학과	0.608415
일어일문학과	0.350359
국어국문학과	0.269022
통일과국제평화융합전공	0.239582
자유전공학부	0.221816
의료인문학융합전공	0.219849
인문학과정의융합전공	0.129227
소셜커뮤니케이션학생설계전공	0.086313

1전공 사학과,
2전공 정치외교학과
이중전공 학생

mmajor_nm	score
공공거버넌스와리더십융합전공	0.977941
경제학과	0.976081
법과행정융합전공	0.720717
경영학과	0.177749
사회규범과 행정 융합전공	0.100891
행정학과	0.083138
인문학과정의융합전공	0.045330
금융공학융합전공	0.043269
자유전공학부	0.009594
정치외교학과	0.006152

1전공 경제학과,
2전공 공공거버넌스와리더십
융합전공 학생

mmajor_nm	score
암호학융합전공	0.996524
수학과	0.996165
보험과위험관리학생설계전공	0.992712
금융공학융합전공	0.983837
통계학과	0.956303
지구환경과학과	0.936540
소프트웨어벤처융합전공	0.924663
경영학과	0.915762
기술창업융합전공	0.856390
식품자원경제학과	0.802311

1전공 수학과,
2전공 보험과위험관리학생설계전공
학생설계전공 학생

Example 1. 제2전공 추천

훈련셋에 들어가지 않은 데이터 테스트, top10개 추천 리스트 출력

mmajor_nm	score
교육학과	0.998676
인적자원개발학학생설계전공	0.998085
국어교육과	0.948781
다문화한국어교육융합전공	0.947611
패션디자인및머천다이징융합전공	0.925751
인문학과정의융합전공	0.614305
인문심리학학생설계전공	0.598036
소비자분석학학생설계전공	0.583741
가정교육과	0.488956
심리학부	0.331412

1전공 교육학과,
2전공 인적자원개발학학생설계전공

mmajor_nm	score
독어독문학과	0.999130
경영학과	0.996465
서어서문학과	0.935053
인문학과문화산업융합전공	0.634863
통일과국제평화융합전공	0.322756
불어불문학과	0.218219
중국학부	0.142102
글로벌학부 한국학전공	0.108632
인적자원개발학학생설계전공	0.074516
금융공학융합전공	0.071468

1전공 독어독문학과,
2전공 경영학과
이중전공 학생

mmajor_nm	score
건축사회환경공학부	0.905345
융합에너지공학과	0.167512
기술창업융합전공	0.137871
산업경영공학부	0.006059
전기전자공학부	0.002546
기계공학부	0.000880
신소재공학부	0.000695
응용통계학과	0.000497
소프트웨어벤처융합전공	0.000393
화공생명공학과	0.000358

1전공 건축사회환경공학부,
2전공 건축사회환경공학부

mmajor_nm	score
미디어학부	0.779057
자유전공학부	0.333001
소프트웨어벤처융합전공	0.108209
융합보안융합전공	0.040923
기술창업융합전공	0.023030
미디어문예창작학과	0.021631
패션디자인및머천다이징융합전공	0.016423
컴퓨터학과	0.006101
인적자원개발학학생설계전공	0.004414
스마트보안학부	0.003717

1전공 미디어학부,
2전공 소프트웨어벤처융합전공

Example 1. 제2전공 추천

다전공이 없는 같은 사회학과 학생

mmajor_nm	score
통일과국제평화융합전공	0.880532
사회학과	0.765507
철학과	0.712162
한국사학과	0.656078
인문학과정의융합전공	0.376938
독어독문학과	0.306265
국어국문학과	0.278535
정치외교학과	0.113022
일어일문학과	0.054413
독일문화학과	0.030368

mmajor_nm	score
사회학과	0.264107
경제학과	0.048913
통계학과	0.025963
인문학과정의융합전공	0.004163
교육학과	0.004130
의료인문학융합전공	0.003865
보험과위험관리학생설계전공	0.002365
금융공학융합전공	0.001933
통일과국제평화융합전공	0.001701
소비자분석학학생설계전공	0.001600

mmajor_nm	score
사회학과	0.517796
의료인문학융합전공	0.046174
교육학과	0.015661
통일과국제평화융합전공	0.013948
과학기술학융합전공	0.004029
소셜커뮤니케이션학생설계전공	0.003905
인문학과정의융합전공	0.001393
사학과	0.000722
철학과	0.000366
소비자분석학학생설계전공	0.000328

mmajor_nm	score
사회학과	0.414091
교육학과	0.000314
통일과국제평화융합전공	0.000275
자유전공학부	0.000099
사학과	0.000055
인문학과정의융합전공	0.000032
한국사학과	0.000030
의료인문학융합전공	0.000012
국어국문학과	0.000005
과학기술학융합전공	0.000004

맞춤형 추천 가능성 확인

결론 및 보완점

- 강의 추천보다 성능이 뛰어난 것 확인 → 커리큘럼이 중요한 역할을 한 것으로 사료됨
- 다전공이 없는 학생의 콜드 스타트 문제 해결
- 같은 학과여도 다른 추천 결과 → 맞춤형 추천 가능 확인

- 서울캠퍼스 학생이 세종캠퍼스의 제2전공을 선택하거나 세종캠퍼스의 학생이 소속변경을 한 경우에 추천 리스트에 세종캠퍼스에서만 이수가능한 제2전공이 추천됨 → 쿼리로 필터링

2021 고려대학교 대학혁신지원사업

KU-Insight Miner AI선배 강의 추천

Knowledge Graph를 활용한 R-GCN기반 제2전공 추천 시스템 개발

감사합니다!

고려대학교 디지털정보처 데이터 hub팀
데이터 사이언티스트 이진숙

2021.07.16