

# 빅데이터 분석 1차 발표

## 1. 구매 상품 추천 시스템

### 1. 논문 소개

#### 1. 선정 논문

Gao et al., A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions, 2022

#### 2. 논문 요약

- 추천 시스템에 그래프 신경망을 적용한 경우를 다음 범주에 따라 분류
  - stage, scenario, objective, and application
- 데이터의 구조적 연결성을 이용하여 그래프 신경망을 적용하는 것에 대한 논의 진행
- 그래프 구축, 최적화, 계산 효율성 등에 대한 분석 진행 및 요약

### 2. 세부 내용

#### 1. 추천 시스템의 분류

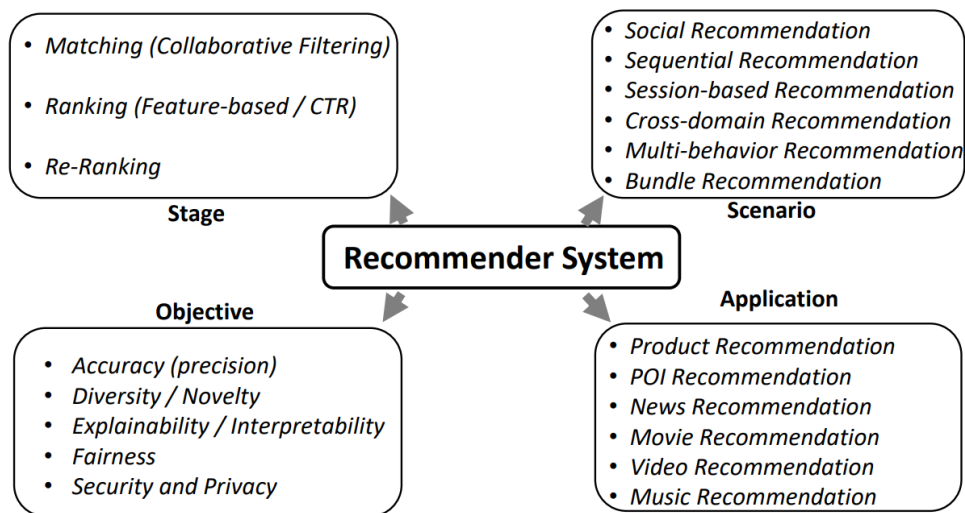


Fig. 1. An illustration of typical recommender systems (stages, scenarios, objectives, and applications)

## 2. 추천 시스템의 구조

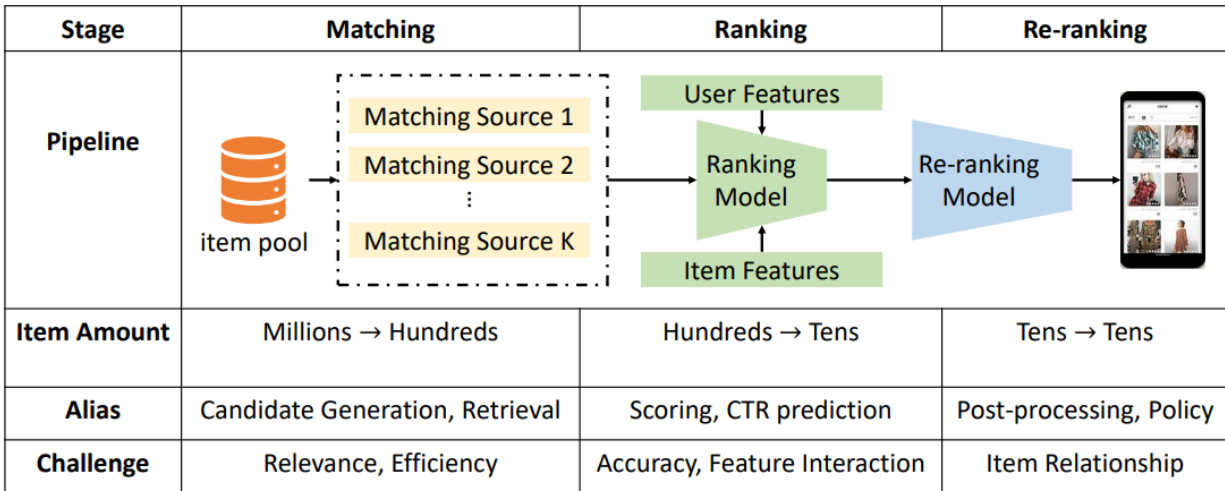


Fig. 2. The typical pipeline of recommender systems.

CTR: Click-Through Rate, 즉 웹사이트에서 무언가를 클릭할 가능성

## 3. 그래프 신경망의 큰 범주

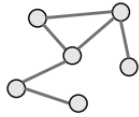
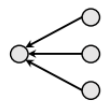
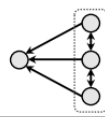
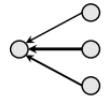
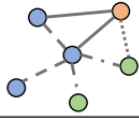
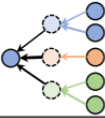

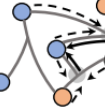
Model	Graph type of GNN	GNN category	Aggregate
GCN		Spectral	
GraphSAGE		Spatial	
GAT		Spatial	
HetGNN		Spatial	
HGNN		Spectral	

Fig. 11. Comparison of several typical GNN models. For graph type, nodes and edges type are represented by colors and line-styles respectively. For aggregation, line-width indicates neighborhood weight.

그래프 신경망 구조는 이미지 같이 유클리드 공간에서 표현되지 않는 데이터 구조에 머신러닝 기법을 적용한 것

추천 시스템은 그래프 신경망을 적용하기 좋은 분야

이 중 GCN(Graph Convolutional Networks)은 가장 간단하면서도 어느 정도 효율이 좋지만 추천 시스템에서 이용되는 아이템 개수의 규모를 생각하면 사용하기 부적절함

#### 4. Matching을 위한 그래프 신경망 모델들

Table 5. Details of GNN models in matching stage.

Model	Graph Construction	Network Design	Specifial Design for Collaborative Filtering
GCMC [8]	user-item	GCN	weight sharing among relations
NGCF [181]	user-item	GCN	enhance propagated information
DGCF [183]	user-item	GAT	disentangled representations
LightGCN [65]	user-item	LightGCN	remove transformation and nonlinearity
SGL [194]	user-item	GCN	self-supervision on graphs
NIA-GCN [160]	user-item	NIA-GCN	neighbor interaction (NI)
PinSage [222]	item-item	GCN	neighbor sampling
IMP-GCN [104]	user-item	GCN	neighbor sampling
HS-GCN [107]	user-item	GCN	re-define distance metrics between nodes
LGCN [228]	user-item	low-pass GCN	trainable kernels for spectral graph convolution

그래서 추천 시스템에 사용하기 적합한 형태로 그래프 신경망의 구조를 다소 수정한 LightGCN 을 사용하려고 고려

GCN에 들어간 여러 아이디어 중 몇 가지를 덜어내고 추천 시스템에 알맞게 수정한 독자적인 구조를 사용

LightGCN 논문을 소개하지 않고 Survey paper를 소개한 이유는 다양한 구조를 고려하며 실험을 진행하기 위함

### 3. 활용 방안

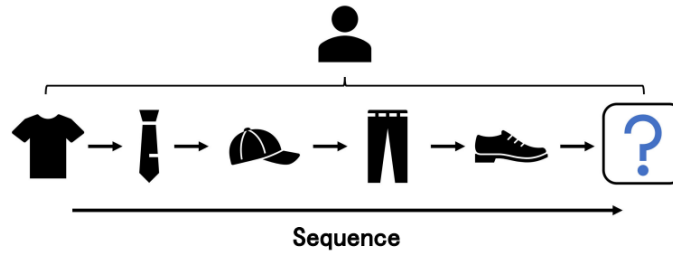


Fig. 4. An illustration of sequential recommendation. Given a user's historical sequence, recommender system aims to predict the next item.

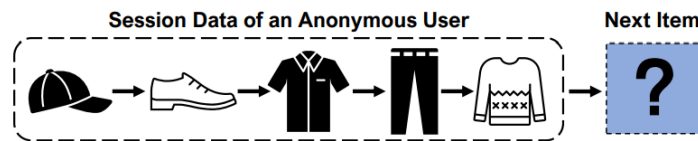


Fig. 5. An illustration of session-based recommendation. Given a anonymous short session, recommender system aims to predict the next item.

추천 시스템의 기반 아이디어 중 자주 사용되는 것이 위의 두 가지

우리는 텍스트로부터 가격, 원가, 사용 기간 등의 특성을 추출하여 한 번에 넣고 추천을 해 주는 Session-based recommendation을 적용할 예정

모든 텍스트가 완전하지 않을 수 있기 때문에 전처리가 필요할 것으로 예상

## 2. 판매 확률 예측: ML vs LLM

### 1. 논문 소개

#### 1. 선정 논문

Ko et al., A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models, 2023

#### 2. 논문 요약

- 사전 학습과 미세 조정을 거친 13억 개에서 128억 개의 매개변수를 가진 대형 언어 모델
- 비영어권 언어를 통한 미세 조정に関心

- 다양한 분야의 데이터로 학습하여 범용성 높은 모델을 구축할 수 있음 (블로그, 뉴스, 사전 등)

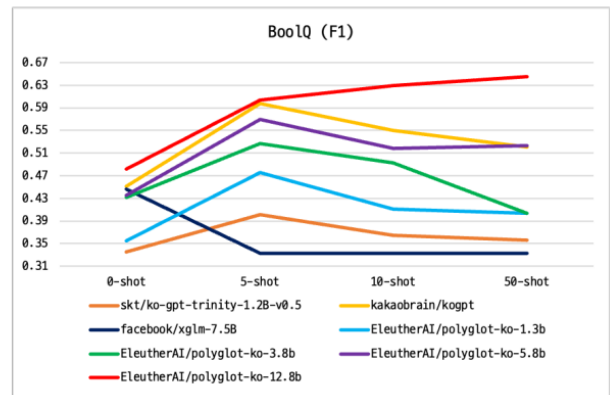
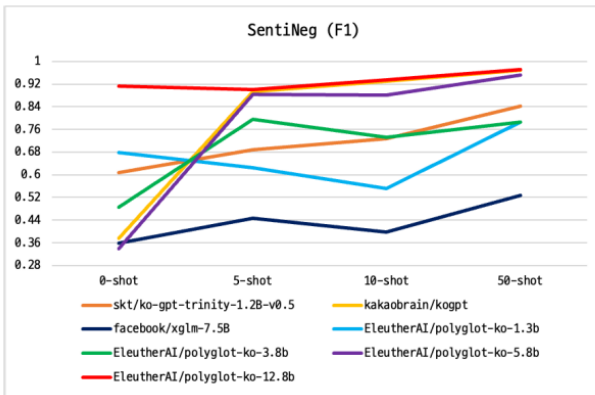
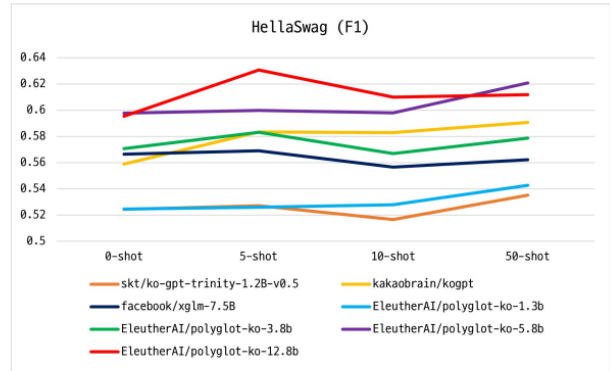
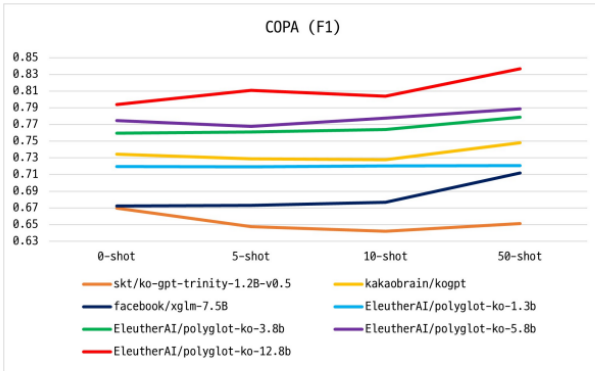
## 2. 세부 내용

### 1. 데이터셋

Source	Size (GB)
Korean blog posts	682.3
Korean news dataset	87.0
Modu corpus	26.4
Korean patent dataset	19.0
Korean Q&A dataset	18.1
KcBert dataset	12.7
Korean fiction dataset	6.1
Korean online comments	4.2
Korean wikipedia	1.4
Clova call	< 1.0
Naver Sentiment Movie Corpus	< 1.0
Korean hate speech dataset	< 1.0
Open subtitles	< 1.0
AIHub various tasks datasets	< 1.0
Standard Korean dictionary	< 1.0

Table 1: Datasets for the Korean language.

## 2. 성능



Shot: 처음 보는 임무에 대하여 거친 연습의 횟수

### 3. 활용 방안

#### 1. 파생 모델



## Update Logs [↗](#)

- 2023.06.23: [한국어 대화 평가 결과 공개](#)
- 2023.06.08: [😊 Polyglot-ko 5.8B 기반 KULLM-Polyglot-5.8B-v2 fp16 모델 공개](#)
- 2023.06.01: [구름\(KULLM\) 데이터셋 v2](#) HuggingFace Datasets 공개
- 2023.05.31: [😊 Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B-v2 fp16 모델 공개](#)
- 2023.05.30: [😊 Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B fp16 모델 공개](#)

## ☁ KULLM (구름): Korea University Large Language Model [↗](#)

KULLM(구름)은 고려대학교 [NLP & AI 연구실](#)과 [HIAI 연구소](#)가 개발한 한국어 Large Language Model (LLM) 입니다.

구름 프로젝트는 한국어 모델 뿐만 아니라, 데이터 셋까지 전면 공개하여 한국어 LLM 생태계에 기여하고자 하였습니다.

주된 활용 모델은 Polyglot 그 자체가 아니라 Polyglot을 미세 조정하여 만든 모델 중 하나인 KULLM

### 2. 성능

#### LLM Inference Results for Korean Evaluation Set [↗](#)


Type	Base-model	Model	이해 가능성 (0 - 1)	자연스러움 (1 - 3)	맥락 유지 (1 - 3)	흥미롭기 (1 - 3)	지시어 사용 (0-1)	전반적인 품질 (1-5)
Closed	GPT3.5-turbo	GPT-3.5	0.980	2.806	2.849	2.056	0.917	3.905
Closed	GPT-4	GPT-4	0.984	2.897	2.944	2.143	0.968	4.083
Open	Polyglot-ko-12.8b	<a href="#">KoAlpaca v1.1</a>	0.651	1.909	1.901	1.583	0.385	2.575
Open	LLaMA-7b	<a href="#">koVicuna</a>	0.460	1.583	1.726	1.528	0.409	2.440
Open	Polyglot-ko-12.8b	<a href="#">KULLM v2</a>	0.742	2.083	2.107	1.794	0.548	3.036

### 3. 평가 방법

- scikit learn 라이브러리를 통한 logistic regression을 진행
- Huggingface 라이브러리를 통하여 KULLM을 불러온 뒤 prompt를 통해 판매 확률을 예측해 달라고 채팅 형식으로 질문
- 두 모델의 성능을 비교

### 4. 예시

#### Prompt

두 사람 간의 대화가 주어집니다. 다음의 지시문(Instruction), 입력(Input)을 받게 될 것입니다. 그리고 지시문과  0  
당신의 작업은 응답을 평가 단계에 따라 응답을 평가하는 것입니다.  
이 평가 기준을 꼼꼼히 읽고 이해하는 것이 중요합니다. 평가하는 동안 이 문서를 계속 열어두고 필요할 때 참조해 주세요

평가 기준:

- 이해 가능성 (0 - 1): Input에 기반하여 Response를 이해 할 수 있나요?
- 자연스러움 (1 - 3): 사람이 자연스럽게 말할 법한 Instruction 인가요?
- 맥락 유지 (1 - 3): Input을 고려했을 때 Response가 맥락을 유지하나요?
- 흥미롭기 (1 - 3): Response가 지루한가요, 아니면 흥미로운가요?
- Instruction 사용 (0 - 1): Instruction에 기반하여 Response를 생성 했나요?
- 전반적인 품질 (1 - 5): 위의 답변을 바탕으로 이 발언의 전반적인 품질에 대한 인상은 어떨가요?

평가 단계:

1. Instruction, Input, 그리고 Response를 주의깊게 읽습니다.
2. 위의 평가 기준에 따라 Response를 평가합니다.

Instruction:

{{instruction}}

Input:

{{input}}

Response:

{{response}}

Result

- 이해 가능성 (0 - 1):
- 자연스러움 (1 - 3):
- 맥락 유지 (1 - 3):
- 흥미롭기 (1 - 3):
- Instruction 사용 (0 - 1):
- 전반적인 품질 (1 - 5):

위 prompt를 우리의 목적과 평가 기준에 맞추어서 수정