ItemRAG: Item-level Retrieval-Augmented Generation for LLM-based Recommendation

Sunwoo Kim — KAIST — kswoo97@kaist.ac.kr
Geon Lee — KAIST — geonlee0325@kaist.ac.kr
Kyungho Kim — KAIST — kkyungho@kaist.ac.kr
Jaemin Yoo — Seoul National University — jaeminyoo@snu.ac.kr
Kijung Shin — KAIST — kijungs@kaist.ac.kr

Paper QR

Overview of ItemRAG

[TL;DR] We propose ItemRAG, an RAG method for LLM-based recommendation that enriches each user-purchased or candidate item with retrieved item-level information.

Input prompt for an LLM without ItemRAG

Task: Rank the candidate items based on the target user's purchase history.

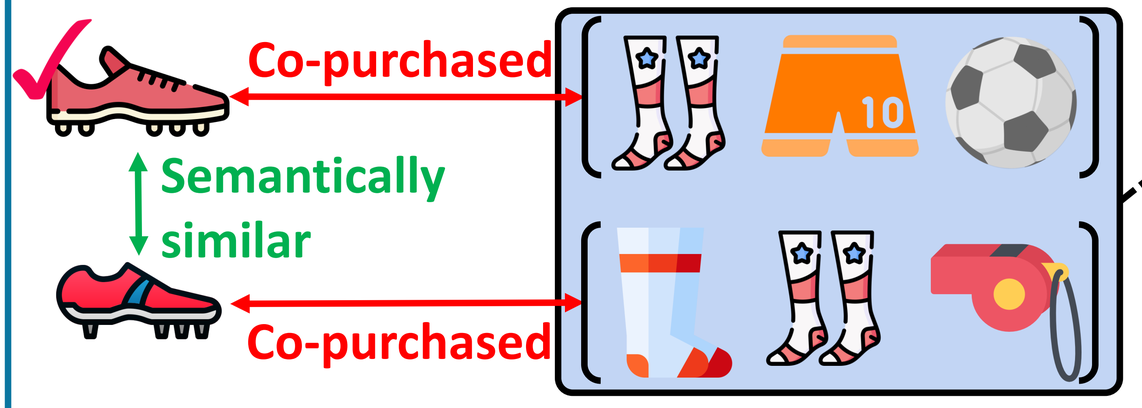
Target user purchase history:



Candidate items:



Retrieval strategy within ItemRAG



- ✓ **[Query item]** For whom to retrieve?
  - ✓ Each user-purchased item or candidate item.
- ✓ **[Retrieval strategy]** How to retrieve?
  - ✓ **Step 1:** For each query item, find items that are co-purchased with (1) the query item and/or (2) semantically similar items.
  - ✓ **Step 2:** Among these items, sample *K* items in proportion to their co-purchase frequencies.

Frequency-based Sampling

New part

Input prompt for an LLM with ItemRAG

Task: Rank the candidate items based on the target user's purchase history.

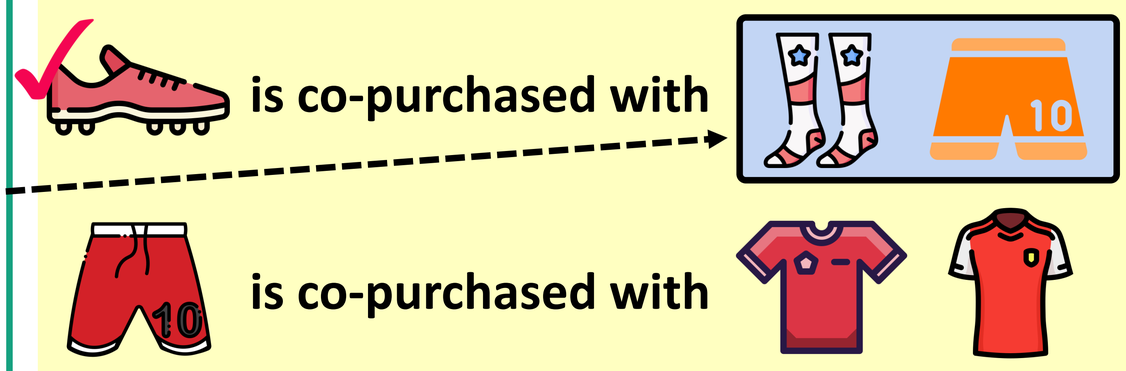
Target user purchase history:



Candidate items:

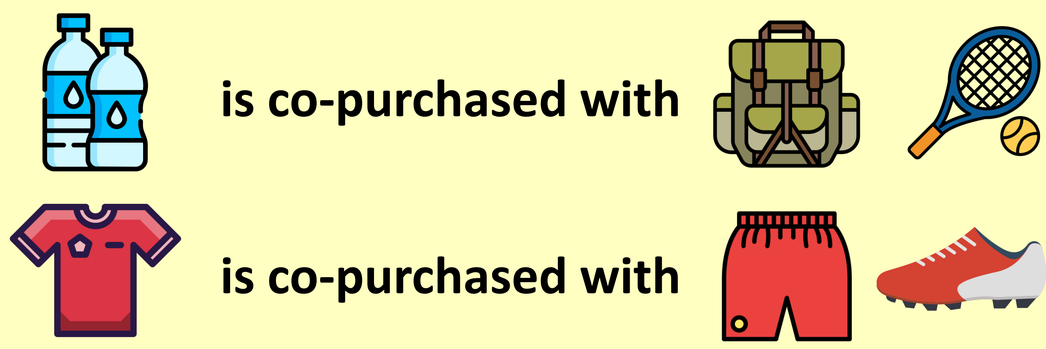


Auxiliary information of user-purchased items:



(+ Co-purchase information for other user-purchased items)

Auxiliary information of candidate items:



(+ Co-purchase information for other candidate items)

Experimental result

ItemRAG achieves the best recommendation performance in 18 out of 20 settings!

	Methods	Beauty & Personal Care					Toys & Games					Sports & Outdoors					Arts, Crafts & Sewing				
		H@1	H@3	H@5	N@3	N@5	H@1	H@3	H@5	N@3	N@5	H@1	H@3	H@5	N@3	N@5	H@1	H@3	H@5	N@3	N@5
Traditional models	LightGCN [2]	35.8*	62.8*	78.8*	51.2*	57.9*	40.9*	67.7*	81.5*	56.4*	62.1*	45.3*	70.6*	83.3*	60.2*	65.5*	51.7*	78.2*	88.0*	67.1*	71.1*
	LightGCN++ [15]	38.3*	65.2*	80.2*	53.4*	59.7*	42.2*	68.5*	82.5*	57.3*	63.1*	46.0*	70.8*	84.0*	60.5*	65.9*	53.0*	79.9*	88.8*	68.2*	72.2*
	SASRec [5]	35.2*	62.6*	78.7*	51.9*	57.5*	32.4*	58.5*	75.1*	47.0	54.6*	44.0*	68.9*	82.1*	57.9*	63.3*	47.4*	72.9*	86.5*	62.5*	69.6*
	BERT4Rec [20]	36.1*	63.2*	78.9*	52.3*	57.7*	32.0*	58.7*	75.8*	47.1*	55.5*	44.2*	70.5*	83.8*	59.4*	65.3*	48.2*	74.6*	87.1*	63.3*	69.9*
Zero-shot LLM	Zero-shot	34.6*	57.7*	72.9*	48.0*	54.2*	39.3*	62.5*	76.5*	52.7*	58.5*	43.4*	66.6*	81.6*	56.7*	62.9*	45.7*	72.2*	83.8*	61.1*	65.6*
User-level RAG for LLMs	ICL [21]	37.7*	61.9*	76.7*	51.5*	57.5*	41.3*	66.4*	80.6*	55.7*	61.1*	46.4*	70.2*	84.1*	59.8*	65.7*	46.3*	75.3*	86.7*	63.1*	67.8*
	AdaptRec [25]	34.7*	58.2*	74.9*	47.9*	55.1*	37.8*	63.3*	77.5*	52.2*	58.1*	44.4*	67.8*	83.2*	57.5*	63.6*	46.4*	74.4*	85.8*	62.9*	67.9*
	ReACT [4]	34.4*	57.7*	73.5*	47.6*	54.2*	38.5*	61.7*	75.9*	51.7*	57.8*	44.7*	68.7*	83.5*	58.2*	64.4*	46.9*	73.0*	83.8*	61.9*	66.4*
	CoRAL [23]	46.5*	69.9*	82.0*	60.2*	65.0*	44.6*	70.1*	81.2*	59.4*	64.0*	48.5*	73.2*	87.1*	62.5*	68.3*	52.4*	80.7*	91.1*	68.9*	73.3*
Variants of ItemRAG	w/o cand-aug.	36.9*	59.2*	73.8*	49.8*	55.8*	42.1*	63.7*	78.9*	54.6*	60.8*	44.8*	68.2*	81.8*	57.5*	63.3*	47.3*	72.8*	84.4*	62.0*	66.8*
	w/o co-purch.	46.1*	68.9*	82.1*	59.1*	64.7*	46.4*	70.8*	81.9*	60.6*	65.1*	50.6*	72.9*	86.2*	63.6*	68.9*	54.6*	80.1*	89.7*	69.5*	73.3*
	w/o sim-items	47.5*	70.2*	82.8*	60.8*	66.3*	49.8	71.6*	83.9*	62.4*	67.7*	50.2*	74.0*	87.7*	64.5	70.0	56.0*	82.1*	91.1*	71.0*	74.6*
	w/o co-freq.	47.7*	70.1	83.9	60.7*	66.6	48.1*	73.1*	83.5*	63.0*	67.0*	50.7*	74.1	87.9*	64.0	69.8*	56.2*	82.2*	91.0*	71.3*	75.0*
Our method	ItemRAG	49.1	71.0	83.8	61.5	66.9	49.5	74.2	85.0	63.8	68.1	51.6	74.7	88.3	64.6	70.3	57.8	82.6	91.9	72.3	76.1