Data Mining Fall 2024 Survey Project

Innovative Data Mining and Machine Learning Approaches for E-commerce Optimization and User Experience Enhancement

Student ID: 313551135	Student ID: 313551099
Student Name: 林念慈	Student Name: 李以恩

Table of Content

List of all papers

#01. Arnab Dutta, Gleb Polushin, Xiaoshuang Zhang, and Daniel Stein. 2024. Enhancing E-commerce Spelling Correction with Fine-Tuned Transformer Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 4928–4938. https://doi.org/10.1145/3637528.3671625

#02. Xiaochen Wang, Xiao Xiao, Ruhan Zhang, Xuan Zhang, Taesik Na, Tejaswi Tenneti, Haixun Wang, and Fenglong Ma. 2024. Mitigating Pooling Bias in E-commerce Search via False Negative Estimation. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 5917–5925. https://doi.org/10.1145/3637528.3671630

#03. Jianke Yu, Hanchen Wang, Xiaoyang Wang, Zhao Li, Lu Qin, Wenjie Zhang, Jian Liao, and Ying Zhang. 2023. Group-based Fraud Detection Network on e-Commerce Platforms. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 5463–5475. https://doi.org/10.1145/3580305.3599836

#04. Yankai Chen, Quoc-Tuan Truong, Xin Shen, Jin Li, and Irwin King. 2024. Shopping Trajectory Representation Learning with Pre-training for E-commerce Customer Understanding and Recommendation. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 385–396. https://doi.org/10.1145/3637528.3671747

#05. Andrea Nestler, Nour Karessli, Karl Hajjar, Rodrigo Weffer, and Reza Shirvany. 2021. SizeFlags: Reducing Size and Fit Related Returns in Fashion E-Commerce. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp; Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 3432–3440. https://doi.org/10.1145/3447548.3467160

#06. Abhinav Anand, Surender Kumar, Nandeesh Kumar, and Samir Shah. 2023. CADENCE: Offline Category Constrained and Diverse Query Generation for E-commerce Autosuggest. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 3703–3712. https://doi.org/10.1145/3580305.3599787

#07. Abhirup Mondal, Anirban Majumder, and Vineet Chaoji. 2022. ASPIRE: Air Shipping Recommendation for E-commerce Products via Causal Inference Framework. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 3584–3592. https://doi.org/10.1145/3534678.3539197

#08. Sen Li, Fuyu Lv, Ruqing Zhang, Dan Ou, Zhixuan Zhang, and Maarten de Rijke. 2024. Text Matching Indexers in Taobao Search. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 5339–5350. https://doi.org/10.1145/3637528.3671654

#09. Xiaowen Shi, Fan Yang, Ze Wang, Xiaoxu Wu, Muzhi Guan, Guogang Liao, Wang Yongkang, Xingxing Wang, and Dong Wang. 2023. PIER: Permutation-Level Interest-Based End-to-End Re-ranking Framework in E-commerce. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 4823–4831. https://doi.org/10.1145/3580305.3599886

#10. Eleanor Loh, Jalaj Khandelwal, Brian Regan, and Duncan A. Little. 2022. Promotheus: An End-to-End Machine Learning Framework for Optimizing Markdown in Online Fashion E-commerce. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 3447–3457. https://doi.org/10.1145/3534678.3539148

#11. Ziming Wang, Qianru Wu, Baolin Zheng, Junjie Wang, Kaiyu Huang, and Yanjie Shi. 2023. Sequence As Genes: An User Behavior Modeling Framework for Fraud Transaction Detection in E-commerce. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 5194–5203. https://doi.org/10.1145/3580305.3599905

#12. Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-commerce Product Attribute Value Extraction. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 2910–2914. https://doi.org/10.1145/3626772.3661357

#13. Guipeng Xv, Chen Lin, Wanxian Guan, Jinping Gou, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. 2023. E-commerce Search via Content Collaborative Graph Neural Network. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 2885–2897. https://doi.org/10.1145/3580305.3599320

#14. Chi Chen, Hui Chen, Kangzhi Zhao, Junsheng Zhou, Li He, Hongbo Deng, Jian Xu, Bo Zheng, Yong Zhang, and Chunxiao Xing. 2022. EXTR: Click-Through Rate Prediction with Externalities in E-Commerce Sponsored Search. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 2732–2740. https://doi.org/10.1145/3534678.3539053

#15 Shaowei Yao, Jiwei Tan, Xi Chen, Juhao Zhang, Xiaoyi Zeng, and Keping Yang. 2022. ReprBERT: Distilling BERT to an Efficient Representation-Based Relevance Model for E-Commerce. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 4363–4371. https://doi.org/10.1145/3534678.3539090

[Paper #01]

Enhancing E-commerce Spelling Correction with Fine-Tuned Transformer Models

I. Target Problem

The search process is a typical behavior of users in e-commerce. Statistically, spelling errors by users are widespread. About 10% of all queries with spelling errors are reported by Google. Hence, it's worth noting that spelling correction plays a vital role in shaping the users' search experience by rectifying erroneous query inputs, thus facilitating more accurate retrieval outcomes. Also, the spelling correction mechanism is crucial to identifying user intent, correcting errors, and delivering desired search results. To achieve this, the author aims to enhance the existing state-of-the-art discriminative model performance with generative modeling strategies. Moreover, they also intend to address the engineering concerns associated with real-time online latency.

II. Dataset

- Training Set
  - Twitter Dataset: https://luululu.com/tweet/#cr
  - Webis Dataset: https://dl.acm.org/doi/pdf/10.1145/1146847.1146848 (AOL search logs)
  - https://aclanthology.org/2021.bea-1.4/
  - The authors used the above sources to create a combination dataset that captures the most typical misspelling.
- Golden Set
  - The authors manually created the golden sets in five different languages: English, German, French, Italian, and Spanish, which are the primary languages in their top revenue-generating sites.

III. Data Mining Workflow

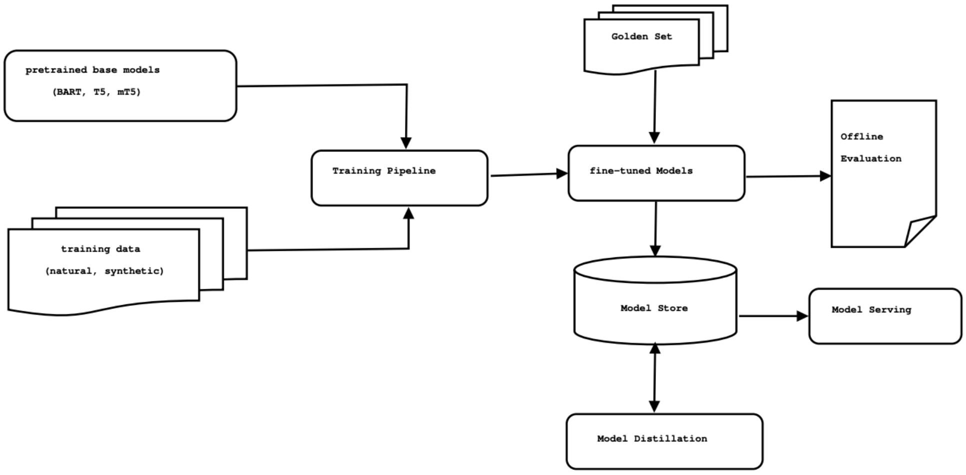


Figure 4: Modular flow of the pipeline

The pre-trained models, such as BART, T5, and mT5 (for multilanguage), are employed during the training stage. The process refines the model and evaluates it against their golden datasets. The training runtime configurations were iteratively adjusted for relative improvements, and the optimal models were checkpointed and stored internally in the Model Store. Model Distillation would access Model Store, persisting the distilled models. The model would be pruned in the distillation process by dropping decoder layers and fine-tuning them. In addition, the authors used a no-teacher distillation scheme, which was observed to produce better model performance compared to the typical teacher-student distillation process. Ultimately, the model serving platform is a CPU/GPU-based Java framework designed for deploying deep learning models. It leverages load-balancing techniques to distribute candidate models across multiple endpoints, supporting internal testing and production use. Finally, in their experimentation evaluation, the 2-layer decoder model was deployed through the process, which integrates the exposed endpoint with the platform.

IV. Results

This work not only implements fine-tuning off-the-shelf Transformer models for spelling correction but also handles the latency issues. There are some crucial results as follows. For the correction task, BART and T5 models outperformed LSTM models. To compare BART and T5 models, non-optimized BART shows faster response times than T5, prompting further BART optimizations for latency. As mentioned in the mining workflow, reducing decoder layers would improve online model latency, enabling real-time traffic serving. Regarding multilanguage, the authors' fine-tuned model for US and UK sites enhances user experiences, which was proved by reduced session exit rates and a drop in zero-result search pages. This work also shows the potential of extending the modeling approach to non-English sites with A/B testing in progress.

[Paper #02]

Mitigating Pooling Bias in E-commerce Search via False Negative Estimation

I. Target Problem

In E-commerce, efficient and accurate product relevance assessment is critical for user experience and business success. However, current methods introduce pooling bias by mistakenly sampling false negatives, which diminishes performance. Furthermore, training such sophisticated search relevance models typically need datasets of high quality and ample quantity, which is usually impractical in e-commerce because annotating training data is time-consuming and labor-intensive. Also, query-product pairs for annotation are typically pre-collected through an introductory information retrieval system, which results in difficulties in that the relevant pairs dominate labeled data. In contrast, irrelevant cases are scarce, which harms the performance of the relevance assessment.

II. Dataset

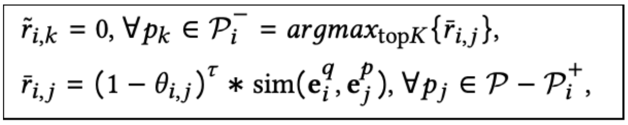
- To examine the efficacy of False Negative Estimation, the authors conducted an experiment on the Semantic Textual Similarity (STS) task using the benchmark dataset. (http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark)
- Training and evaluating the cross-encoder model: The dataset is from Instacart's ongoing data collection efforts to benchmark search relevance, where a random sample of search queries and the products ranked in top display positions are sent for human annotation every month to judge the relevance between queries and products.

III. Data Mining Workflow

- The input data is a random sample of search queries and the products. The annotators categorize each (query, product) pair into 5 relevant categories: strongly relevant, relevant, somewhat relevant, not relevant, and offensive.
- To alleviate pooling bias, they proposed two models: sampling regularization and pseudo label generation. They utilize False Negative Estimation as an additional regularization term to generate negative samples. To further eliminate the remaining pooling bias, they designed pseudo label generation to conduct a dual insurance mechanism to guarantee the extinction of all false negatives. The authors established the Bias-mitigating Hard Negative Sampling (BHNS) by combining the above two modules.
- In sampling regularization, first, calculate the likelihood that pj is relevant to qi.

$$\boxed{\theta_{i,j} = \frac{1}{T_j}\sum_{t=1}^{T_j} r_{t,j} * \mathrm{sim}(\mathbf{e}_i^q, \mathbf{e}_t^q), \forall p_j \in \mathcal{P}_i^-,}$$

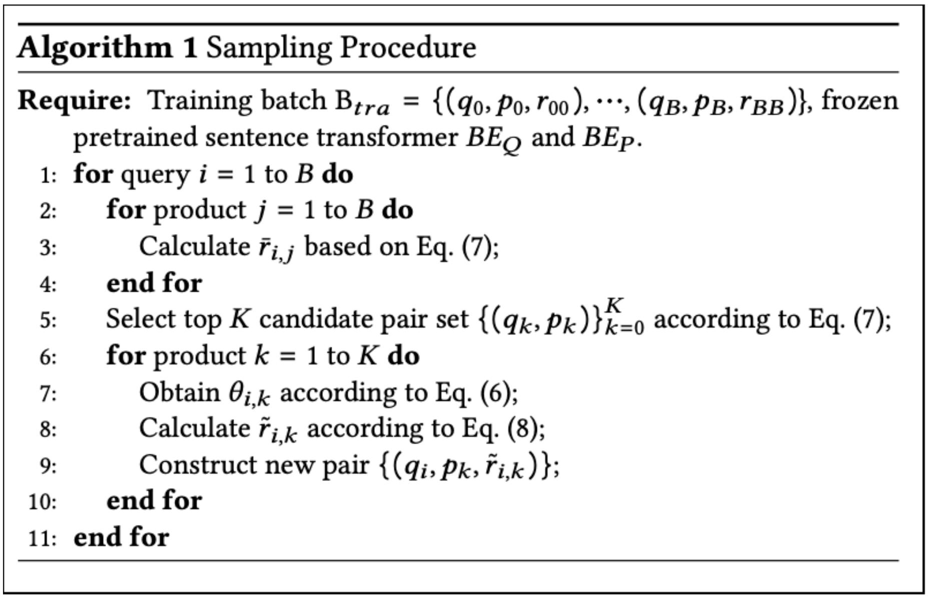
Then, ensure that the pair has less chance of being erroneously selected as a hard negative sample, thus balancing informativeness and robustness and mitigating pooling bias.



$$\tilde{r}_{i,k} = 0, \forall p_k \in \mathcal{P}_i^- = argmax_{\text{top}K}\{\bar{r}_{i,j}\},$$

$$\bar{r}_{i,j} = (1-\theta_{i,j})^{\tau} * \text{sim}(\mathbf{e}_i^q, \mathbf{e}_j^p), \forall p_j \in \mathcal{P} - \mathcal{P}_i^+,$$

- In pseudo label generation, they used the false negative sample to train the model and generate pseudo label.
- After the BHNS, the false negative sample is used to train the cross-encoding model.



Algorithm 1 Sampling Procedure

```
Require: Training batch $B_{tra} = \{(q_0, p_0, r_{00}), \cdots, (q_B, p_B, r_{BB})\}$, frozen
    pretrained sentence transformer $BE_Q$ and $BE_P$.
for query $i = 1$ to $B$ do
    for product $j = 1$ to $B$ do
        Calculate $\bar{r}_{i,j}$ based on Eq. (7);
    end for
    Select top $K$ candidate pair set $\{(q_k, p_k)\}_{k=0}^{K}$ according to Eq. (7);
    for product $k = 1$ to $K$ do
        Obtain $\theta_{i,k}$ according to Eq. (6);
        Calculate $\tilde{r}_{i,k}$ according to Eq. (8);
        Construct new pair $\{(q_i, p_k, \tilde{r}_{i,k})\}$;
    end for
end for
```

IV. Results

Through the authors' experimental results and case study, their proposed BHNS demonstrates superior performance, thus providing strong evidence that validates the hypothesis underlying False Negative Estimation and supports the correctness of the rationale behind the design of BHNS. The success of BHNS implies its applicability extends beyond the e-commerce domain. Also, they have demonstrated that BHNS significantly enhances the performance of the cross-encoder model, which holds considerable promise for its application in the e-commerce setting.

[Paper #03]

Group-based Fraud Detection Network on e-Commerce Platforms

I. Target Problem

Nowadays, with the rapid technological and commercial innovation on e-commerce platforms, there are an increasing number of frauds that bring great harm to these platforms. Although the high concealment and strong destructiveness of group-based fraud exist, no research can completely exploit the information within the transaction networks of e-commerce platforms for group-based detection. The fraudulent attacks affect the platform's reputation, influence the user experience, and even lead to the loss of platform users. Therefore, in this paper, the authors aim to design an end-to-end learning-based model for group-based fraud detection on attributed bipartite graphs. Also, they seek to find fraudulent clicks or fraudulent users on e-commerce platforms.

II. Dataset

- Dataset for "Ride Item's Coattails" attack detection
  - TC: https://tianchi.aliyun.com/dataset/123862
  - TB: a large-scale attributed bipartite customer-product graph on the Alibaba e-commerce platform Taobao
- Dataset for STARS attack detection
  - Bitcoin Alpha & Bitcoin OTC: The two datasets are made by https://ieeexplore.ieee.org/document/7837846
  - The original link of the dataset: https://www.cs.umd.edu/~srijan/wsn/
  - The two datasets are user-to-user trust networks of Bitcoin users trading using Alpha platform and OTC platform, and they are made bipartite by splitting each user into a "rater" with all its outgoing edges and each "product" with all incoming edges.

III. Data Mining Workflow

- The two types of fraud approach the authors aim to detect
  - "Ride Item's Coattails" attack creates fake clicks by groups of fraudsters to establish the deceptive correlation between popular and low-quality products. It, therefore, promotes the recommendation of low-quality products to other customers.
  - Sockpuppet-based Targeted Attack on Reviewing Systems (STARS) attack: it aims at the review systems of platforms, and it's usually conducted by groups of fraudsters, which initiates fake ratings of target products, thus changing the rating of the target products and fraudulently promoting the products to other legitimate users.

- The authors proposed an end-to-end semi-supervised model Group-based Fraud Detection Network(GFDN) for group-based fraud detection on attributed bipartite graphs. The model consists of two main parts: a structural feature generation module and a communication-aware fraud detection network.
- First, in the structural feature generation module, they use (α, β)-core distribution to obtain the structural information. α limits the minimum degree of the vertex set, such as the customer vertex set, and β limits the minimum degree of the other vertex set, such as the product vertex set. (α, β)-core is used to effectively obtain subgraphs with different sparsity by varying the values of α and β. The subgraph is used to generate expressive structural features of the whole graph.
- Second, in the fraudster community detection, they proposed a community-aware graph neural network for bipartite graphs with the inspiration of SDCN, which achieves SOTA performance for clustering on the unipartite graphs. Later, they proposed Bipartite Deep Clustering Network(BDCN), which can better detect communities based on structural and attribute information than SDCN. In addition, BDCN can be trained in a self-supervised fashion while preserving the information from the input features with an autoencoder. The autoencode preserves the structural and attribute information from initial features with MLP. Furthermore, they use graph neural networks(GNN) and the autoencoder in BDCN to capture the abundant information within the attributed bipartite graph further.
- To further optimize GFDN for both types of fraud, they designed a multi-task training objective for edge(fraud detection) and vertex classification(fraudster detection).

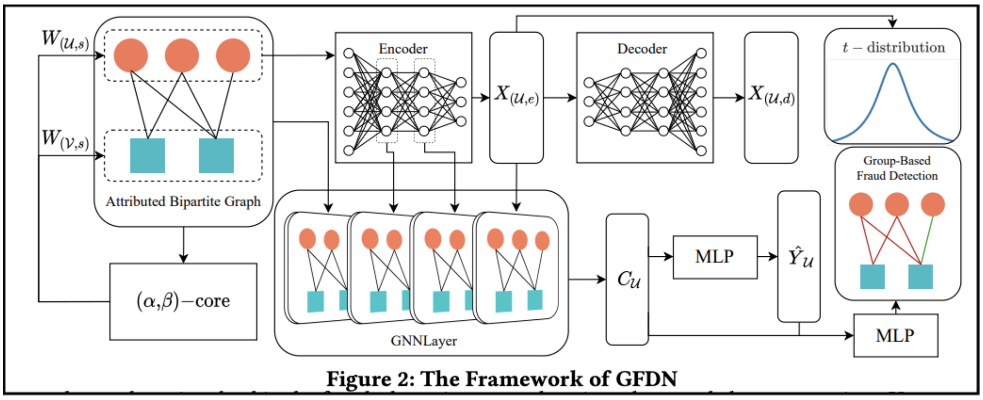


Figure 2: The Framework of GFDN

IV. Results

In this paper, the authors proposed GFDN. This novel end-to-end model adaptively utilizes the cohesive subgraph distribution, structural, attribute, and community information in the attributed bipartite graph for group-based fraud detection. Moreover, they also conducted extensive experiments for the fraud detection of “Ride Item’s Coattails” attack and STARS attack on real-life datasets. The results of these experiments demonstrate a significant performance improvement of GFDN compared with the existing methods of group-based fraud detection. Ultimately, they also conducted an in-depth analysis to evaluate the effectiveness of each component in GFDN.

[Paper #04]

Shopping Trajectory Representation Learning with Pre-training for E-commerce Customer Understanding and Recommendation

I. Target Problem

In e-commerce, understanding customer behavior is crucial for improving service quality. The ever-growing volume of products bombards online shoppers, making it difficult to identify items of interest. Recommender systems address this challenge by providing personalized suggestions throughout the customer shopping journey, from browsing to checkout. In pursuit of personalization for various scenarios, the key prong of delivering high-quality services is rooted in reliable and comprehensive customer understanding. Therefore, it motivated the authors to effectively unveil customers' insights via mining information from various historical shopping engagements. In this paper, the authors proposed C-STAR, a new framework that learns compact representations from customer shopping journeys, with good versatility to fuel multiple downstream customer-centric tasks.

II. Dataset

- The authors used customer engagements for 28 days to fully anonymize the data. For the following three downstream tasks, they collected the following datasets for fine-tuning and evaluation.
- Task 1: Customer Segmentation. About 1M data were collected following rule-based semantic similarity.
- Task 2: Shopping Trajectory Completion. They collected about 5M new shopping trajectories and randomly hid 20% of trajectory elements.
- Task 3: Shopping Intent Identification. They merged 5M additional trajectories from the purchase data.

III. Data Mining Workflow

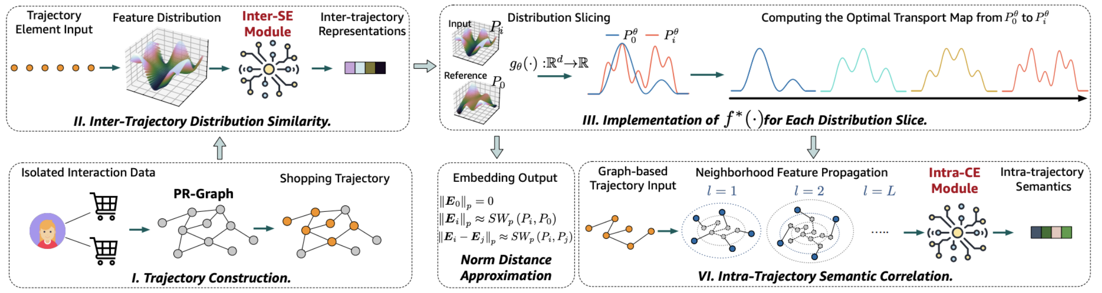

Trajectory Element Input
Feature Distribution
Inter-SE Module
Inter-trajectory Representations
II. Inter-Trajectory Distribution Similarity.
Input P_i
Reference P_0
Distribution Slicing
$g_\theta(\cdot):\mathbb{R}^d\rightarrow\mathbb{R}$
P_0^θ P_i^θ
Computing the Optimal Transport Map from P_0^θ to P_i^θ
III. Implementation of $f^*(\cdot)$for Each Distribution Slice.
Isolated Interaction Data
PR-Graph
Shopping Trajectory
I. Trajectory Construction.
Embedding Output
$\|\boldsymbol{E}_0\|_p = 0$
$\|\boldsymbol{E}_i\|_p \approx SW_p(P_i, P_0)$
$\|\boldsymbol{E}_i - \boldsymbol{E}_j\|_p \approx SW_p(P_i, P_j)$
Norm Distance Approximation
Graph-based Trajectory Input
Neighborhood Feature Propagation
$l=1$
$l=2$
$l=L$
Intra-CE Module
Intra-trajectory Semantics
VI. Intra-Trajectory Semantic Correlation.

Figure 2: Illustration of C-STAR for coarse-to-fine trajectory representation learning (best view in color).

- The authors used a Product-Relation Graph(PR-Graph), an internal knowledge base of product categories and relations organized in the graph format. The graph's nodes represent product categories, and edges capture strong correlations like co-purchases.

- In inter-trajectory distribution similarity, the Optimal Transport Theory measures the similarity between shopping trajectories. The trajectories are modeled as distributions, and sliced-Wasserstein distance is used for efficient and accurate similarity measurement.
- In intra-trajectory semantic correlation, Graph Convolutional Networks (GCNs) encode semantic relationships among trajectory elements using neighborhood structures in the PR-Graph.
- To enhance model performance, they conducted pre-training. They designed two pre-training objectives:
  - Inter-trajectory Element Overlaps
    Rank trajectories based on shared elements to learn customer preferences.
  - Intra-trajectory Contextual Relations
    This aims to ensure that these matching scores surpass those computed from other negative nodes that do not appear in the trajectory.
- The framework of C-STAR can used to downstream tasks:
  - Customer segmentation
  - Shopping trajectory completion
  - Shopping intent identification

IV. Results

The authors defined the notion of shopping trajectory that encompasses customer interaction at the level of product categories. They proposed C-STAR, a new framework, to learn compact representations from the customer shopping journeys, and it, along with good versatility, fuels multiple downstream customer-centric tasks. C-STAR excels at modeling both inter-trajectory distribution similarity, the structural similarities between different trajectories, and intra-trajectory semantic correlation, the semantic relationships within individual ones. They also conducted extensive evaluations on large-scale industrial and public datasets, which demonstrates the effectiveness of C-STAR across three diverse customer-centric tasks. These tasks empower customer profiling and recommendation services for enhancing personalized shopping experiences on their E-commerce platform.

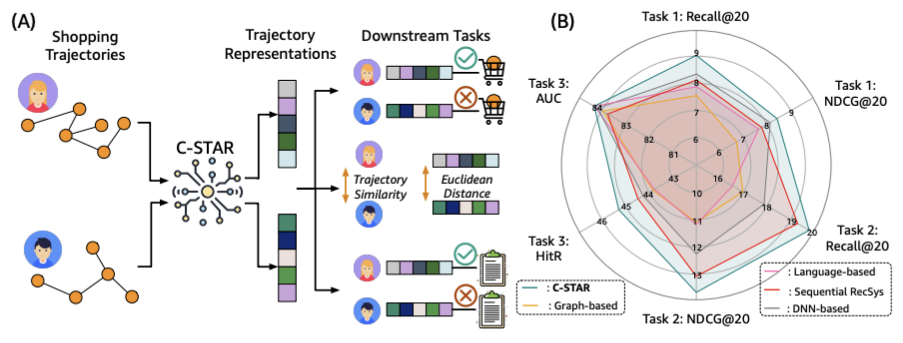


Figure 1: (A) C-STAR framework illustration; (B) performance comparison (%) with selected competitive methods.

[Paper #05]

SizeFlags: Reducing Size and Fit Related Returns in Fashion E-Commerce

I. Target Problem

Nowadays, e-commerce is growing rapidly, and the fashion industry has recently witnessed a noticeable shift in customers' order behavior towards stronger online shopping. However, fashion articles ordered online do not always find their way to a customer's wardrobe. A large share of them end up being returned. Finding clothes that fit online is very challenging and accounts for one of the main drivers of increased return rates in fashion e-commerce. What's worse, size and fit related returns severely impact the customers' experience and dissatisfaction with online shopping, the environment through an increased carbon footprint, and the profitability of online fashion platforms. Over the past years, customer returns have been growing significantly, reaching a 50% increase year over year for specific categories.

II. Dataset

The dataset used in this paper is large-scale, weakly annotated data derived from customer returns in the fashion e-commerce platforms. Specifically, it includes subjective and noisy feedback from customers regarding return reasons, which is used as input to determine the fit of fashion articles. This data is supplemented by priors from human expert feedback and computer vision techniques to address the size and fit recommendation challenge￼.

III. Data Mining Workflow

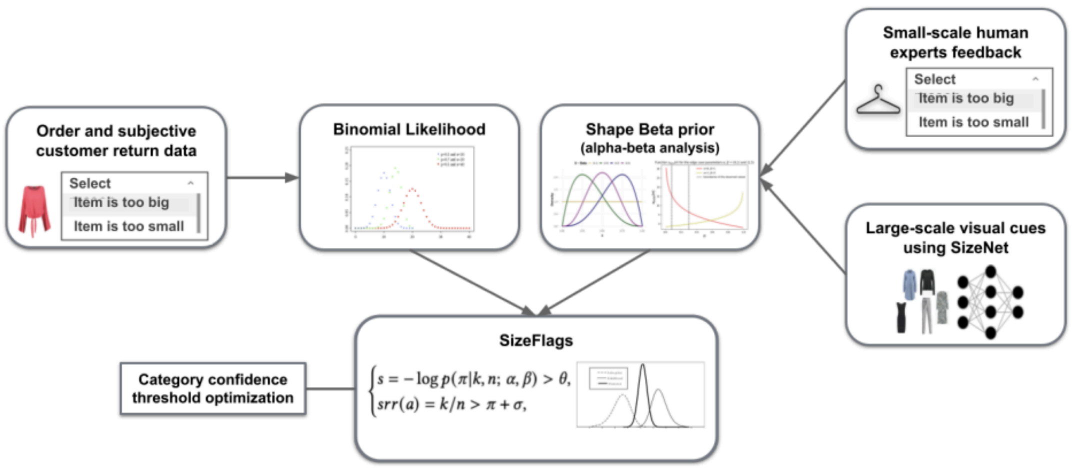
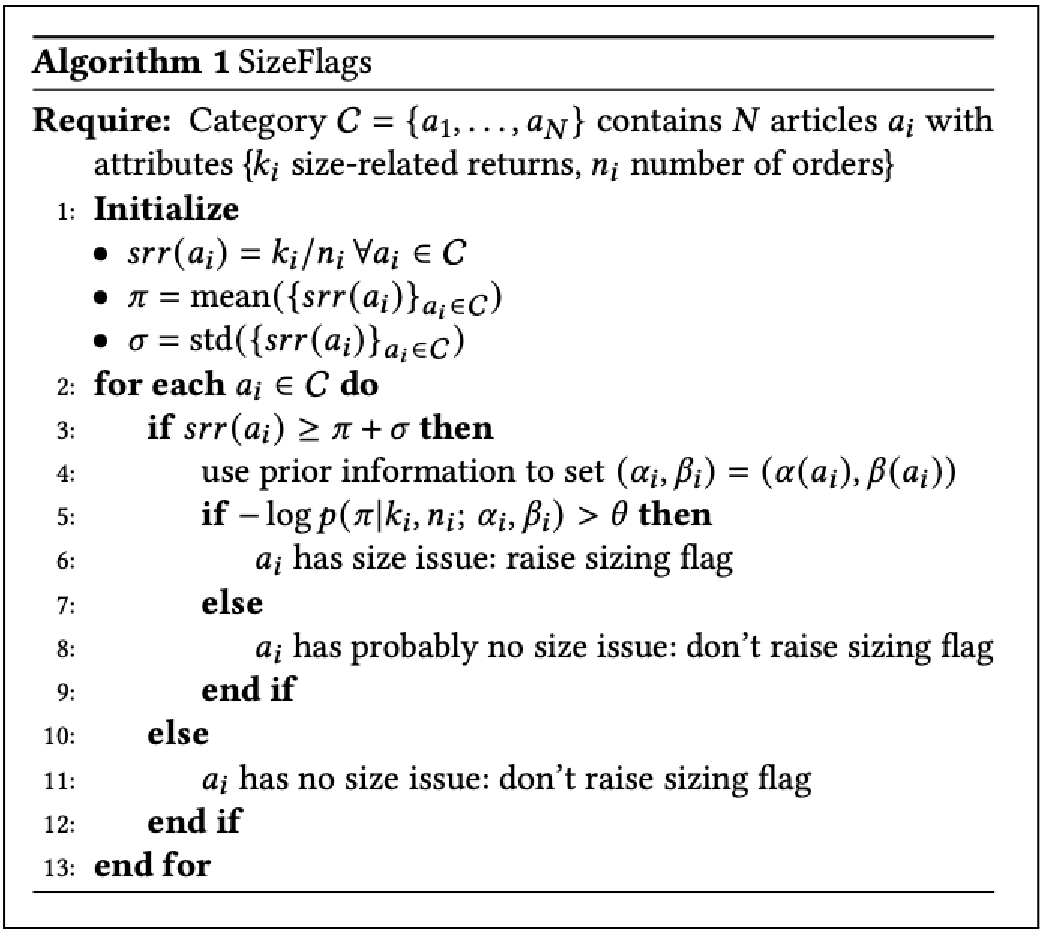

Small-scale human experts feedback
Select
Item is too big
Item is too small
Order and subjective customer return data
Select
Item is too big
Item is too small
Binomial Likelihood
Shape Beta prior (alpha-beta analysis)
Large-scale visual cues using SizeNet
SizeFlags
$s = -\log p(\pi | k, n; \alpha, \beta) > \theta,$
$srr(a) = k/n > \pi + \sigma,$
Category confidence threshold optimization

Figure 1: High level overview of the SizeFlags approach.

- The workflow uses a large-scale, weakly annotated dataset, which includes customers' feedback on articles' size and fit issues.
- They first simplify the complex problem of determining the size and fit of fashion articles into 3-class classification problems, such as no size issue, too small, and too big. In addition, to determine whether an article has a size issue, they used a Bayesian model based on a binomial likelihood distribution. Separate Bayesian

models predict whether an article is “too big” or “too small.” If neither condition is met, the article is flagged as having “no size issue.”

- The model utilized a Beta distribution as the prior. Priors are enriched with human expert feedback and computer vision techniques that analyze article images for size and fit characteristics.
- They also set a conservative threshold that makes the algorithm more robust to noise and increases confidence.
- Priors are used to tackle the challenge of limited data for new articles.
- The model was evaluated by A/B testing and continuous evaluation.



Algorithm 1 SizeFlags

```
Require: Category $C = \{a_1, \ldots, a_N\}$ contains $N$ articles $a_i$ with
    attributes $\{k_i$ size-related returns, $n_i$ number of orders$\}$
Initialize
    • $srr(a_i) = k_i / n_i \; \forall a_i \in C$
    • $\pi = \text{mean}(\{srr(a_i)\}_{a_i \in C})$
    • $\sigma = \text{std}(\{srr(a_i)\}_{a_i \in C})$
for each $a_i \in C$ do
    if $srr(a_i) \geq \pi + \sigma$ then
        use prior information to set $(\alpha_i, \beta_i) = (\alpha(a_i), \beta(a_i))$
        if $-\log p(\pi | k_i, n_i; \alpha_i, \beta_i) > \theta$ then
            $a_i$ has size issue: raise sizing flag
        else
            $a_i$ has probably no size issue: don't raise sizing flag
        end if
    else
        $a_i$ has no size issue: don't raise sizing flag
    end if
end for
```

IV. Results

The authors proposed SizeFlags, a probabilistic Bayesian model. Leveraging the advantages of the Bayesian framework, they extended their model to successfully integrate rich priors from human experts' feedback and computer vision intelligence. Ultimately, they conducted extensive experiments of A/B testing and continuous evaluation of the model, and they demonstrated the strong impact of the proposed approach in robustly reducing size-related returns in online fashion platforms over 14 countries. The results showed that SizeFlags effectively reduces size-related returns in online platforms. Moreover, using optimized thresholds and rich priors greatly reduced the number of ordered and returned articles necessary for reducing size-related returns.

[Paper #06]

CADENCE: Offline Category Constrained and Diverse Query Generation for E-commerce Autosuggest

I. Target Problem

In e-commerce platforms, Query AutoComplete(QAC) or AutoSuggest(AS) is the first place of user interaction with an e-commerce search engine. Therefore, the QAC system must suggest relevant and well-formed queries for multiple possible user intents. However, much of the recent work generates synthetic candidates using models trained on uer queries, and thus, there are three issues. The first one is the cold start problem, as new products in the catalog fail to get visibility due to a lack of representation in user queries. The second problem is the poor quality of generated candidates due to concept drift. The last one is low diversity or coverage of attributes such as brand, color, and other facets in generated candidates. Hence, the authors propose an offline neural query generation framework, CADENCE(Constrained And DiversE Neural Candidate GenErator), to address the issues.

II. Dataset

- The query dataset contains unique user queries over two months. There are 42 categories and about 4000 verticals in their e-commerce catalog.
- After the data preprocessing, the authors excluded user queries with impressions of less than 5 and product titles having a length of more than 10 tokens.
- The final query dataset has about 6 million unique queries, and the catalog dataset contains about 4 million unique titles.

III. Data Mining Workflow

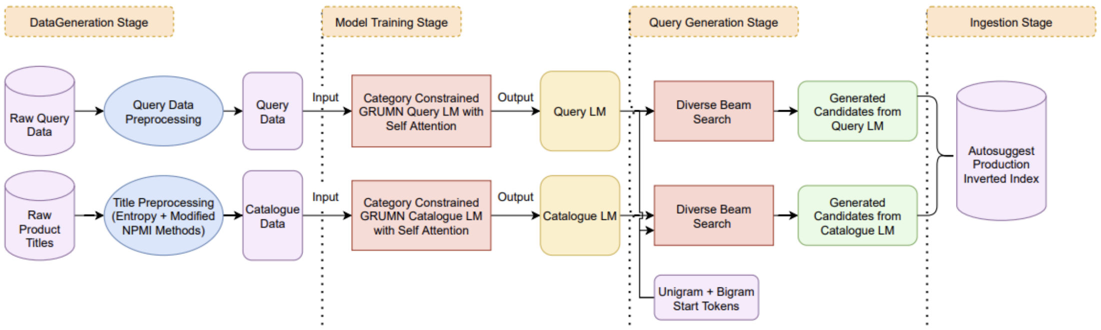

DataGeneration Stage
Model Training Stage
Query Generation Stage
Ingestion Stage
Raw Query Data
Query Data Preprocessing
Query Data
Input
Category Constrained GRUMN Query LM with Self Attention
Output
Query LM
Diverse Beam Search
Generated Candidates from Query LM
Autosuggest Production Inverted Index
Raw Product Titles
Title Preprocessing (Entropy + Modified NPMI Methods)
Catalogue Data
Input
Category Constrained GRUMN Catalogue LM with Self Attention
Output
Catalogue LM
Diverse Beam Search
Generated Candidates from Catalogue LM
Unigram + Bigram Start Tokens

Figure 2: CADENCE Framework

- First, they collected the data from the past two months of unique user queries, and queries appearing less than five times were filtered out. Also, they process the data with spell correction, stopword removal, singularization, and unit correction. Moreover, they use entropy and PMI(pointwise mutual information) as two statistical measures to identify the non-informative tokens and reduce noise.

- In the model training stage, they trained two separate neural language models, Query Language Model(Query LM) and the Catalogue Language Model(Catalogue LM). Also, they create a deep constrained GRU-MN(Gated Recurrent Unit Memory Network) based RNN neural network by adding category constraints to each layer to prevent concept drift and ensure relevance.
- In the query generation stage, they implement dynamic beam search for diverse query suggestions. In addition, they generate candidates offline by combining outputs from Query LM and Catalogue LM.
- In the ingestion stage, they filtered non-performing queries that led to a null search and ingested generated candidates into the autosuggest production system.

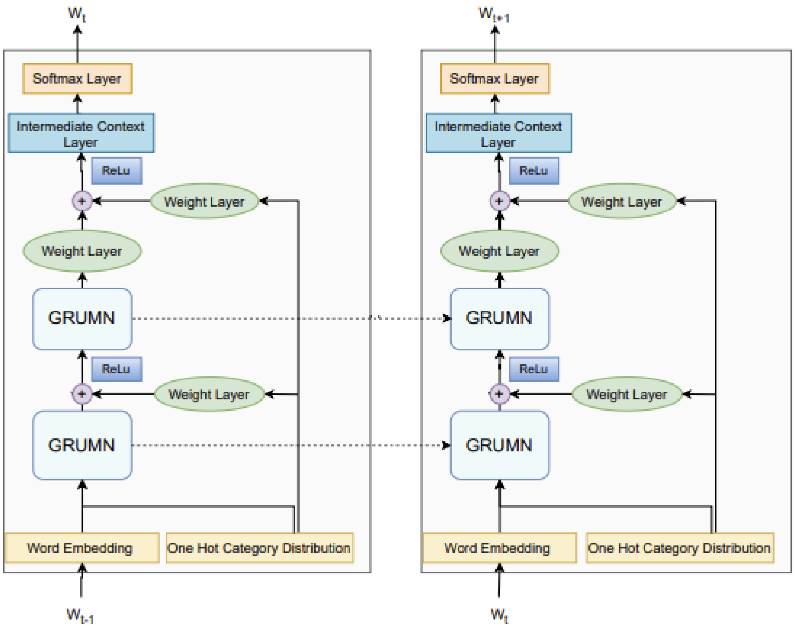


Figure 3: Category constrained model architecture

IV. Results

In this paper, they propose an offline neural query generation framework, CADENCE, to address challenges. Besides solving for cold start and rare/unseen prefix coverage, CADENCE also increases the coverage of the existing query prefixes through more relevant and diverse query suggestions. Specifically, the framework prevents concept drift by enforcing category constraints during training and generation. Also, they developed a product Catalogue Language Model from the products' short and noisy title text to address the cold start problem. Moreover, they brought diversity to their query candidates.

[Paper #07]

ASPIRE: Air Shipping Recommendation for E-commerce Products via Causal Inference Framework

I. Target Problem

In online shopping platforms, delivery speed is critical for e-commerce platforms' success—faster delivery promise to the customer results in increased conversion and revenue. However, in the absence of sufficient data, business decisions are often based on intuitive, broad-brush "back of the envelop" rules without a detailed evaluation of the impact of the decisions. A principled data-driven response would help determine the appropriate quantum of discount to maximize the revenue per unit discount. The authors presented a machine learning-based framework to recommend product air-shipping eligibility in such conditions.

II. Dataset

They used more than 40 million product page views from an emerging marketplace over a period of three months in 2019. The dataset includes contextual features such as product attributes, customer details, merchant details, and page-view-level data. It trains and evaluates machine learning models for propensity score estimation and conversion prediction.

III. Data Mining Workflow

- In propensity score matching, a propensity score model is built to estimate the likelihood of a fast shipping promise being offered. They also used a LightGBM-based classifier to predict conversion.
- At the stage of doubly robust estimation, they used it to ensure unbiased estimates of ITE and ATE. They trained the classification model for each bin, defined by the propensity score.
- Next is the estimated causal impact of fast promise on conversion.
- In air capacity allocation, products are ranked based on benefit scores. In this part, the authors are interested in maximizing the revenue earned through air recommended products. A knapsack optimization algorithm was used to select products for air shipping within the available air capacity constraints.
- The last part was offline and online experiments.
  - In offline evaluation, they compared ASPIRE with baseline policies. Also, metrics such as revenue uplift and product weights were analyzed.
  - In online A/B testing, ASPIRE is applied to the treatment cohort, while the control cohort follows a rule-based policy.

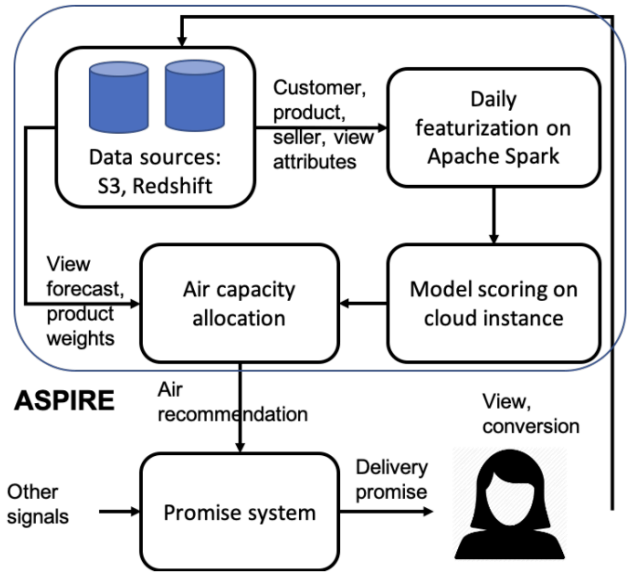


Figure 8: High-level system architecture of *ASPIRE*.

IV. Results

The authors presented ASPIRE(Air ShipPIng REcommendation), which is a causal modeling framework to determine offline the air eligibility of products based on their historical performance. ASPIRE can balance the trade-off between revenue or conversion and delivery cost to decide whether a product should be shipped via air. In this paper, the authors considered the problem of air capacity allocation across many products. Importantly, this is the first paper that addresses the problem systematically. Their proposed framework addresses the impact of confounding factors such as replication, available air capacity, and demand forecasts. Ultimately, they conducted extensive offline and online A/B testing experiments to measure the performance of ASPIRE. A/B testing indicated that the ML-based policy results in a 79 basis points improvement in revenue compared to the incumbent rule-based policy.

[Paper #08]

Text Matching Indexers in Taobao Search

I. Target Problem

In Taobao, the largest e-commerce platform in China, product search is an essential service. Users can easily find products relevant to their specific needs through this service. Coping with titanium size query loads, Taobao product search has traditionally relied on classical term-based retrieval models due to their powerful and interpretable indexes. In essence, efficient retrieval hinges on adequately storing the inverted index. Recent successes involve reducing the inverted index size, but the construction and deployment of lossless static index pruning in practical product searches still pose non-trivial challenges.

II. Dataset

- Large-scale industrial offline dataset
  - The authors randomly sampled 1.5 million search interaction records on the next day. The size of the candidate item set of inverted index is about 2 billion, and the semantic index is about 100 million
- Online dataset
  - The size of the item candidate set is about 2 billion. The size of the semantic index of the item pool is 100 million.

III. Data Mining Workflow

- First, they analyzed user logs to capture higher-order term dependencies and preferences based on purchase, click, exposure, and relevance patterns.
- In hypergraph construction, User-Query-Item(UQI) interactions were modeled into hypergraphs to understand the dependencies.
- Next, a novel TermRank algorithm was used to compute values for terms and items within the hypergraphs, capturing their importance based on user preferences.
- The traditional inverted index is pruned by a user-aware term scoring function, which integrates user search preferences and query-independent item quality metrics.
- In semantic retrieval, a multi-granularity semantic model was incorporated to address vocabulary mismatches.
- Ultimately, they combined the pruned inverted index and semantic retrieval models to ensure comprehensive coverage of relevant items.
- In the evaluation part, they conducted offline evaluations and online A/B testing.

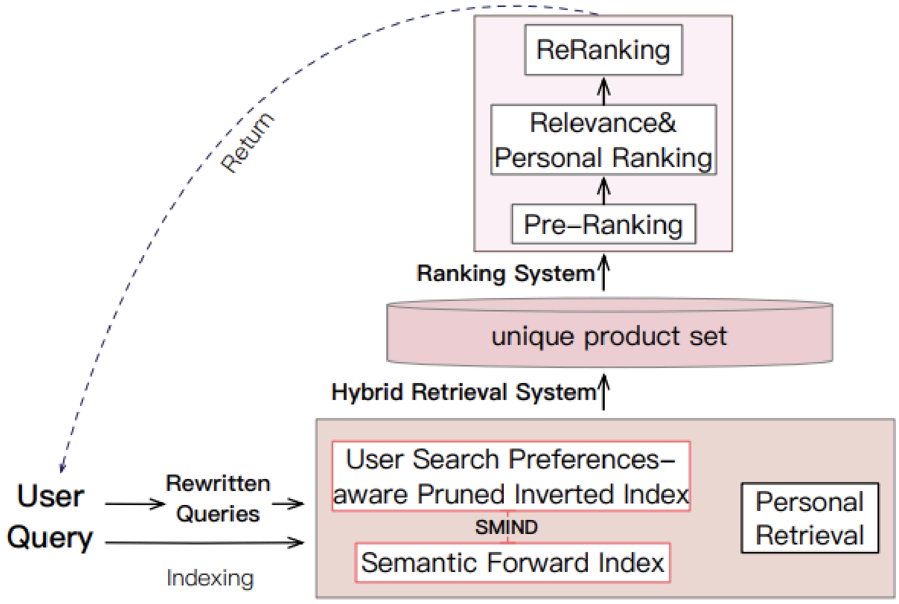


Figure 3: Overview of the architecture of the Taobao search engine with the proposed indexing system SMIND.

IV. Results

The authors introduced SMIND(SMart INDexing) to retrieve relevant, high-quality items. SMIND is an efficient billion-scale product search indexing solution. The key idea is to minimize information loss during the static inverted index pruning process by incorporating higher-order term dependencies from user search preferences. They conducted a term-level analysis of multiple user search preferences, which were employed to prune inverted indexes and bridge the gap between user queries and pruned inverted indexes. The authors also offered new insights that user search preferences encompass higher-order term dependencies beyond text relevance. Ultimately, they evaluated SMIND's performance through offline evaluation and online A/B testing. The result showed that SMIND effectively mitigates the Matthew effect of user queries and has been in service for hundreds of millions of daily users since November 2022.

[Paper #09]

PIER: Permutation-Level Interest-Based End-to-End Re-ranking Framework in E-commerce

I. Target Problem

In order to improve the user's decision-making efficiency in E-commerce applications, a slate which contains limited items is usually provided based on the user's interest. In order to model the influence of the arrangement of displayed items on user behaviors, the re-ranking stage is introduced to rearrange the initial list from the ranking stage. Many existing re-ranking methods directly take the initial ranking list as input, and generate the optimal permutation through a well-designed context-wise model, which brings the evaluation-before-reranking problem. Also in the two-stage architecture, for the generation stage, heuristic methods only use point-wise prediction scores and lack an effective judgment; for the evaluation stage, most existing context-wise evaluation models only consider the item context and lack more fine-grained feature context modeling. This paper presents a novel end-to-end re-ranking framework named PIER to tackle the above challenges.

II. Dataset

In order to verify the effectiveness of our framework, they conduct sufficient experiments on both public dataset and industrial dataset.

- Avito dataset (public dataset)
  The public Avito dataset contains user search logs and metadata from avito.ru, which contains more than 36M ads, 1.3M users and 53M search requests. The full features include user search information and ad information.
- Meituan dataset (industrial dataset)
  The industrial Meituan dataset is collected on Meituan food delivery platform during April 2022, which contains user information(e.g., userid, gender, age), ad information(e.g., adid, categoryid, brandid and so on)

III. Data Mining Workflow

The paper proposes a novel end-to-end re-ranking framework named Permutation-Level InterestBased End-to-End Re-ranking (PIER). The framework still follows the two-stage paradigm which contains two modules named Fine-grained Permutation Selection Module (FPSM) and Omnidirectional Context-aware Prediction Module (OCPM). The goal of PIER is to enhance recommendation quality by re-ranking based on user click behaviors. The steps are as follows:

- Input: Given an initial ranked list *C* and a sequence of user click behaviors *B*
- Generate Candidate Permutations: Create multiple candidate permutations *G* using a permutation algorithm.
- Fine-Grained Selection Module (FPSM): Select the top *K* permutations based on time-sensitive Hamming distance, calculated with SimHash, weighted by time to reflect current user interests.

- Omnidirectional Context-Aware Prediction Module (OCPM): Predict scores (e.g., click-through rate) for each permutation and choose the highest-scoring one as the output.
- Integration: FPSM and OCPM form a unified framework, working collaboratively for optimal re-ranking results.

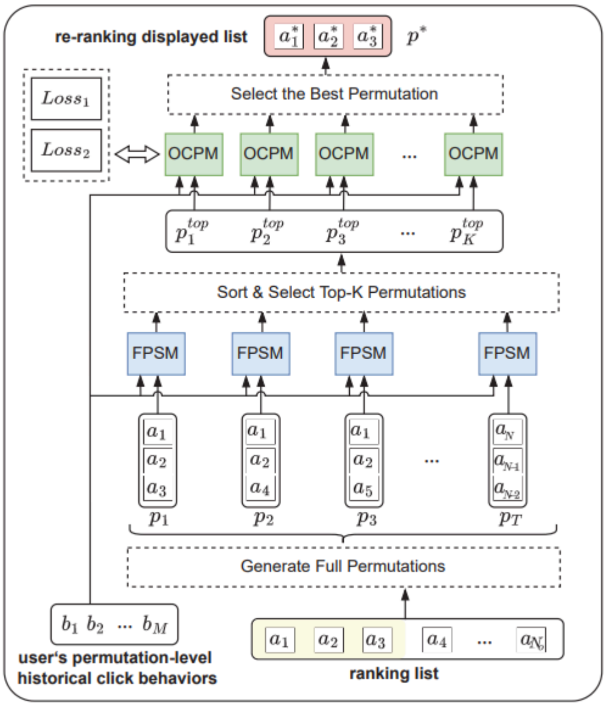

re-ranking displayed list
a_1^* a_2^* a_3^*
p^*
$Loss_1$
$Loss_2$
Select the Best Permutation
OCPM
OCPM
OCPM
...
OCPM
p_1^{top} p_2^{top} p_3^{top} ... p_K^{top}
Sort & Select Top-K Permutations
FPSM
FPSM
FPSM
FPSM
a_1 a_2 a_3
p_1
a_1 a_2 a_4
p_2
a_1 a_2 a_5
p_3
...
a_N a_{N-1} a_{N-2}
p_T
Generate Full Permutations
b_1 b_2 ... b_M
user's permutation-level historical click behaviors
a_1 a_2 a_3 a_4 ... a_N
ranking list

Figure 2: Overview of our framework PIER. PIER takes ranking list and user's permutation-level historical click behaviors as input, and outputs a re-ranking list to display with the help of FPSM and OCPM.

IV. Results

PIER is compared with other models (e.g. base ranking model, PRM and so on) and all deployed on Meituan food delivery platform through online A/B test with 1% of whole production traffic from April 09, 2022 to April 25, 2022 (one week). As a result, PIER gets CTR and GMV increase by 5.46% and 5.83% respectively. Also in the aspect of time costs which determine whether it can be applied to a large scale of industrial scenarios, compared with beam-search, PIER has improved CTR by 2.29% and GMV by 2.31% while the time-out ratio effect increases little, which is acceptable to the system. Now, PIER has been deployed online and serves the main traffic, and contributes to significant business growth.

[Paper #10]

Promotheus: An End-to-End Machine Learning Framework for Optimizing Markdown in Online Fashion E-commerce

I. Target Problem

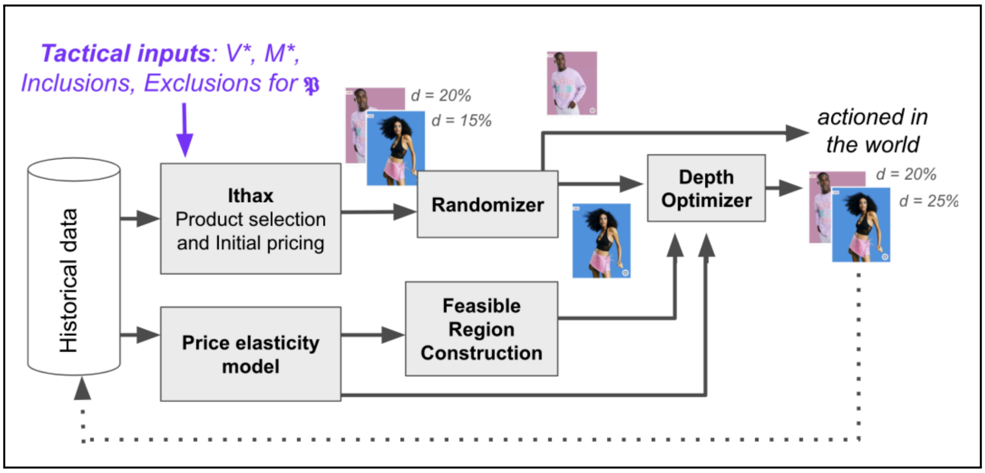
Managing discount promotional events (”markdown”) is a significant part of running an ecommerce business, and inefficiencies can significantly impact a retailer’s profitability. Traditional approaches for tackling this problem rely heavily on price elasticity modelling, yet due to the incomplete data, many retailers still rely on rule-based methods, missing out on potential profitability gains from machine learning. In this paper, the authors introduce two novel end-to-end markdown management systems for optimising markdown at different stages of a retailer’s journey. The first system, "Ithax," enacts a rational supply-side pricing strategy without demand estimation, and can be usefully deployed as a “cold start” solution to collect markdown data while maintaining revenue control. The second system, "Promotheus," presents a full framework for markdown optimization with price elasticity.

II. Dataset

The model is trained by several years of historical data. Also because seasonality strongly modulates product demand in the fashion sector, they also include several seasonality covariates, and they train the models with several years of historical data to expose the model to several epochs of a seasonal trend.

III. Data Mining Workflow

Promotheus comprises multiple components. Firstly, "Ithax" is a supply-side markdown algorithm that selects products and sets initial depths. In order to measure incrementality and actively sample the action space more broadly, they randomly hold-out a proportion of selected products whose prices are set by Ithax (Randomizer). For all other products, depths are adjusted later on (Depth Optimizer) before the decisions are actioned in the world. They train a price elasticity model, and use offline validation to define the feasible region for decision making (Feasible Region Construction), as inputs to the Depth Optimizer. Although Ithax is a sub-process of Promotheus, it is itself a robust end-to-end markdown management system that can be deployed with financial control. Both Promotheus overall and Ithax specifically outperform manual pricing decisions in an online test, with Promotheus showing the best performance overall.



IV. Results

The performance of the system was evaluated over 3 months of price reduction campaigns, with the following results:

1. Goal Achievement: Ithax successfully hit stock depth and stock value targets for every single markdown event, converging within 25 iterations and completing within 30 minutes.
2. Discount Allocation: Products with poorer performance (i.e. higher cover) were assigned deeper discount depths by Ithax. Within each markdown event, Ithax also allocated more products to middling depths.
3. Flexible Strategy Adaptation: The system also provides Include/Exclude Feature and Group Prioritization.

[Paper #11]

Sequence As Genes: An User Behavior Modeling Framework for
Fraud Transaction Detection in E-commerce

I. Target Problem

With the explosive growth of e-commerce, detecting fraudulent transactions in real-world scenarios is becoming more and more important. Many supervised approaches have been proposed to use user behavior sequences, which record the user's track on platforms and contain rich information for fraud transaction detection. However, these methods always suffer from the scarcity of labeled data. Some pre-training methods in NLP and CV domains help to solve the problem, but user behavior sequences differ intrinsically from text, images, and videos. In this paper, the authors propose a novel and general user behavior pre-training framework, named Sequence As GEnes (SAGE), which provides a new perspective for user behavior modeling and is inspired by the nature of DNA expression.

II. Dataset

1. Pre-training datasets

   They implement the user behavior modeling framework on real-world data collected from multiple online e-commerce scenarios in Alibaba and split the dataset for different pre-training tasks.

   - MAM: the Mask Action Modeling task for pre-training the snapshot encoder
   - SeqMuta: sequential mutation task
   - SeqReco: sequential recombination task

2. Evaluation datasets

   They conduct experiments for fraud transaction detection on four real business scenarios, i.e., Tmall Supermarket, Tmall Global, Taobao, and Ali Health.

III. Data Mining Workflow

The authors propose a self-supervised pre-training framework to learn representation from user behavior data for fraud transaction detection. The framework draws inspiration from the perspective of "Sequence As Genes" and can fall into several vital components.

1. Input Data Paradigm
   - Long-term user behavior sequences are organized by intercepting only the crucial segments, referred to as snapshots.
   - The approach is inspired by non-coding DNA, ensuring shorter input sequences while preserving essential information over extended time spans.
2. Two-Stage Transformer Architecture
   - Snapshot Encoder: A transformer that embeds each snapshot into a real-valued vector.
   - Summarizer: A second transformer that aggregates the snapshot embeddings into a sequential representation for modeling long-term user behaviors.

3. Two-Stage Pre-Training Technique
   - Stage 1: Pre-train the snapshot encoder using an MLM-like (Masked Language Modeling) approach.
   - Stage 2: Pre-train the summarizer with two novel self-supervised learning tasks:
     - Sequential Mutation Task: Inspired by genetic mutations, this task involves adding noise to the sequence and training the model to reconstruct the original.
     - Sequential Recombination Task: Inspired by genetic recombination, this task involves using contrastive learning techniques to bring similar sequences (from the same fraud ring) closer in representation space.

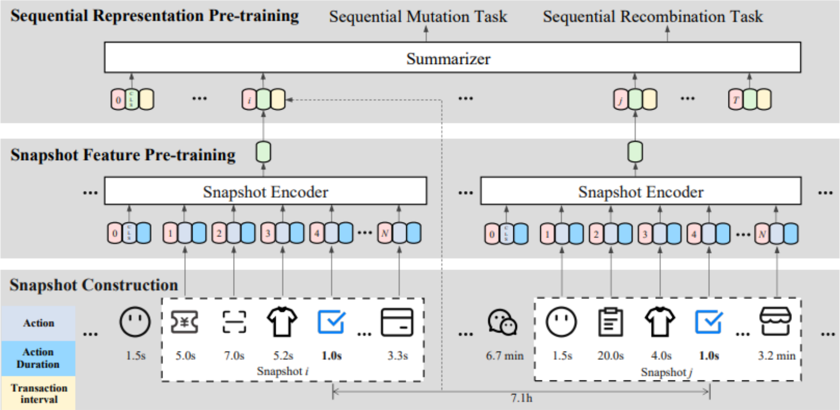


Figure 2: Overview of the *Sequence-As-Genes* (SAGE) framework for user behavior modeling.

IV. Results

The SAGE model was compared with three advanced sequence modeling methods: BiLSTM-max, HConvNet, and transformer. SAGE demonstrated significant performance improvements in fraud detection in all four scenarios. At a precision of 0.9, SAGE achieved a recall rate increase of approximately +4.93% compared to transformer+ and +3.44% compared to the best non-pretrained model HConvNet+. These results confirm the effectiveness of SAGE's pre-training approach.

[Paper #12]

LLM-Ensemble: Optimal Large Language Model Ensemble Method
for E-commerce Product Attribute Value Extraction

I. Target Problem

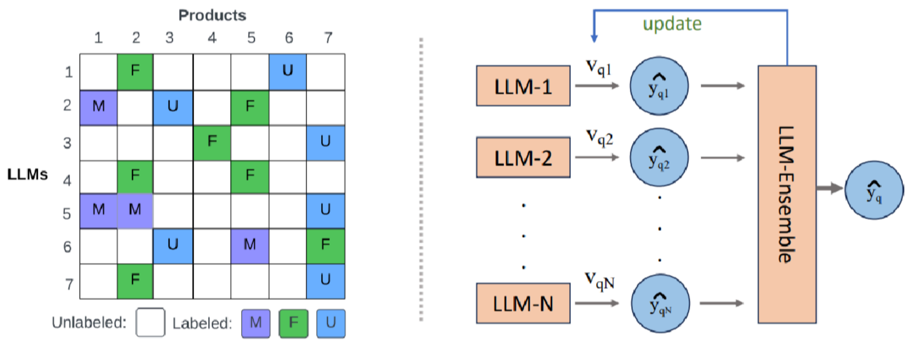
The provision of precise product attribute values is fundamental in ensuring high-quality recommendations and enhancing customer satisfaction. Recently LLMs have demonstrated state-of-the-art performance in numerous attribute extraction tasks. However, varying strengths and weaknesses are exhibited by different LLMs due to the diversity in data, architectures, and hyperparameters, making them complementary to each other, with no single LLM dominating all others. In this paper, they propose a novel algorithm called LLM-ensemble to ensemble different LLMs' outputs for attribute value extraction. They iteratively learn the weights for different LLMs to aggregate the labels with weights to predict the final attribute value. Not only can the method be proven theoretically optimal, but it also ensures efficient computation, fast convergence, and safe deployment.

II. Dataset

The datasets are WalmartAge and Walmart-Gender, which contain products sensitive to the ages or genders of customers. Each dataset has 20K items for offline evaluation. The ground-truth label is created by the crowdsourcing results.

III. Data Mining Workflow

This paper introduces a novel algorithm called LLM-ensemble designed to ensemble the outputs of various LLMs for the purpose of attribute extraction. At its core, the approach is based on the Dawid-Skene Model, a structured latent variable model, to iteratively learn and assign weights to different LLM outputs.



LLMs	1	2	3	4	5	6	7
1		F				U	
2	M		U		F		
3				F			U
4		F			F		
5	M	M					U
6			U		M		F
7		F					U

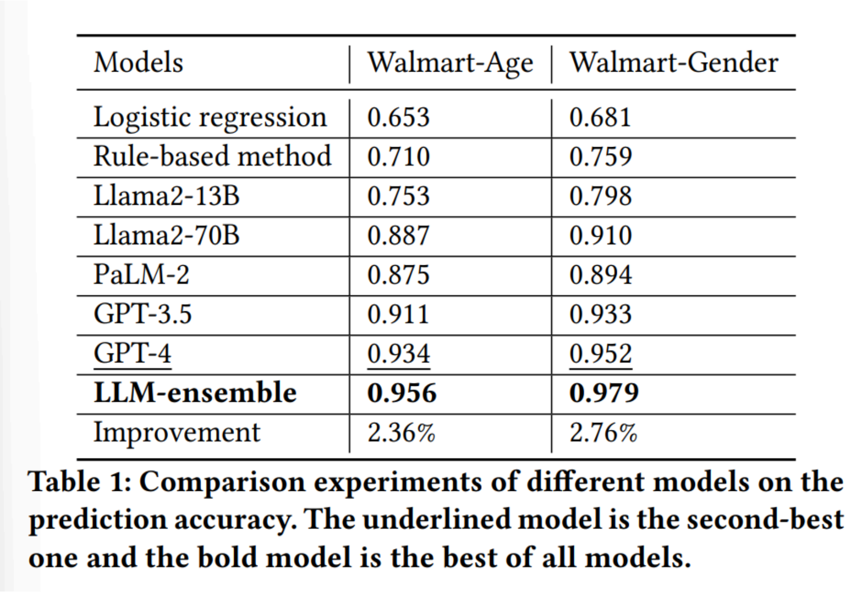
(a) Input data matrix W

(b) LLM-ensemble algorithm

Figure 1: (a) The input data matrix W. We take the attribute "gender" as an example, and its labels are "Male" (M), "Female" (F), and "Unisex" (U). (b) The illustration of LLM-ensemble procedures. To learn the label of a product for attribute q, we have N LLMs as inputs to the LLM-Ensemble algorithm. After several rounds of iteration, the algorithm generates the weights for each LLM and aggregates the labels with weights to predict the final label $\hat{y}_q$.

IV. Results

The proposed LLM-ensemble is compared with its base LLMs: Llama2-13B, Llama2-70B, PaLM-2, GPT-3.5, GPT-4, and also includes logistic regression. From the below table of the experiment results, it can be seen that the proposed LLM-ensemble achieves the best performance (highest accuracy) compared to all other baseline models, which demonstrates the effectiveness of the LLM ensemble algorithm. The method is applied to Walmart's internal data and launched in several production models, leading to improved Gross Merchandise Volume (GMV), Click-Through Rate (CTR), Conversion Rate (CVR), and Add-to-Cart Rate (ATC).



Models	Walmart-Age	Walmart-Gender
Logistic regression	0.653	0.681
Rule-based method	0.710	0.759
Llama2-13B	0.753	0.798
Llama2-70B	0.887	0.910
PaLM-2	0.875	0.894
GPT-3.5	0.911	0.933
GPT-4	0.934	0.952
LLM-ensemble	**0.956**	**0.979**
Improvement	2.36%	2.76%

Table 1: Comparison experiments of different models on the prediction accuracy. The underlined model is the second-best one and the bold model is the best of all models.

[Paper #13]

E-commerce Search via Content Collaborative Graph Neural Network

I. Target Problem

Recently, many E-commerce search models are based on Graph Neural Networks (GNNs). Despite their promising performances, they are (1) lacking proper semantic representation of product contents; (2) less efficient for industry-scale graphs; and (3) less accurate on long-tail queries and cold-start products. To address these problems, this paper proposes CC-GNN, a novel Content Collaborative Graph Neural Network. Firstly, CC-GNN enables content phrases to participate explicitly in graph propagation to capture the proper meaning of phrases and semantic drifts. Secondly, CC-GNN presents several efforts towards a more scalable graph learning framework, including efficient graph construction, MetaPath-guided Message Passing, and Difficulty-aware Representation Perturbation for graph contrastive learning. Furthermore, CC-GNN adopts Counterfactual Data Supplement at both supervised and contrastive learning to resolve the long-tail/cold-start problems.

II. Dataset

1. Industry-scale Product Query Search dataset

The dataset, named IPQS, is constructed by collecting 91 days of log data in a real-world E-commerce platform. Each record in the IPQS dataset corresponds to a product query and a clicked item and contains the necessary information about the queries and items, including product IDs, categories, prices, sales, images, titles, etc. They randomly sample 35 million queries, 87 million items, and 709.7 million interactions from the log data in the first 90 days as training data and use the log data of the last day for testing.

2. Amazon Sport Dataset

The dataset contains 35,598 users, 18,357 items, and 296,337 interactions, which is a widely used recommendation data.

III. Data Mining Workflow

The method focuses on a particular type of E-commerce search where users use products to search products. The goal of E-commerce search with product queries is to return a list of relevant items for a given query. Although a query is also an item that is drawn from the universe of products, queries and items are modeled separately. In addition, they define another type of entity, i.e., phrases which appear in the titles of the queries and items. It is natural to model the three types of entities and their relationships as a graph.

The overall framework of CC-GNN is shown in Figure 2. They construct a Content Collaborative Graph and use MetaPath-guided Message Passing to get the node embeddings. CC-GNN combines supervised learning and contrastive learning, where the contrastive learning is based on Difficulty Aware Representation Perturbation and Counterfactual Data Supplement at contrastive learning.

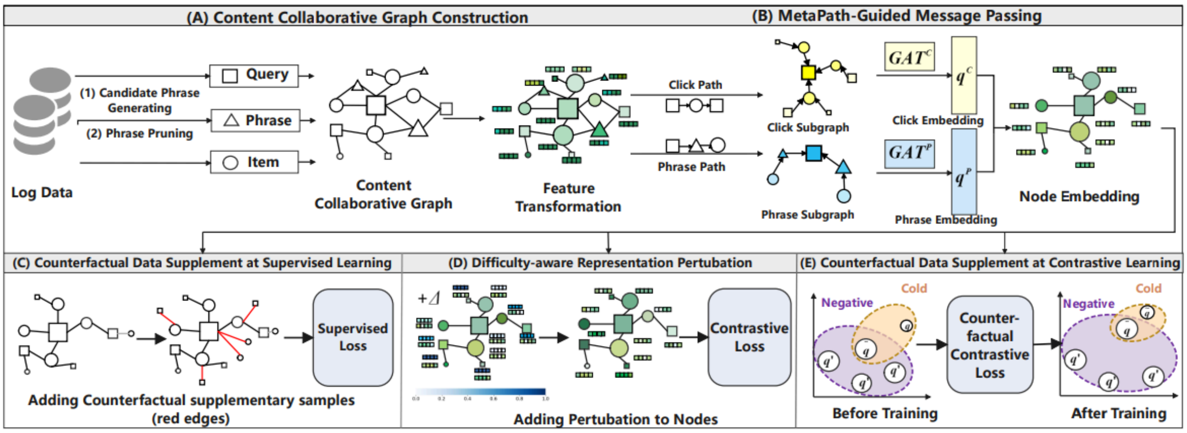

(A) Content Collaborative Graph Construction
(B) MetaPath-Guided Message Passing
(1) Candidate Phrase Generating
(2) Phrase Pruning
Log Data
Query
Phrase
Item
Content Collaborative Graph
Feature Transformation
Click Path
Phrase Path
Click Subgraph
Phrase Subgraph
GAT^c
GAT^p
q^c
q^p
Click Embedding
Phrase Embedding
Node Embedding
(C) Counterfactual Data Supplement at Supervised Learning
(D) Difficulty-aware Representation Pertubation
(E) Counterfactual Data Supplement at Contrastive Learning
Supervised Loss
Adding Counterfactual supplementary samples (red edges)
$+\Delta$
Contrastive Loss
Adding Pertubation to Nodes
Negative
Cold
Counter-factual Contrastive Loss
Before Training
After Training

Figure 2: Overall framework of CC-GNN

IV. Results

CC-GNN is compared on the IPQS dataset with different GNNs in the encoding phase of retrieval models. (1) GraphSAGE samples and aggregates messages from a local neighborhood. GraphSAGE has been widely utilized in retrieval models. (2) AdaptiveGCN samples neighbors and uses the sampled subgraph and learned residual graph to generate node embeddings. (3) LasGNN samples the neighbors layer-wise along the metapaths and constructs a subgraph to aggregate massage.

1. Overall performance Analysis

   CC-GNN improves over the best baseline (i.e., LasGNN) in terms of overall *Recall*@100, *MRR*@100, *NDCG*@100 by 11.8%, 14.5%, and 16.3%, respectively.

2. Head queries and long-tail queries

   CC-GNN significantly improves long-tail query performance. The *Recall*@100, *MRR*@100, *NDCG*@100 on long-tail queries are increased by 13.7%, 16.7%, 14.2% comparing with the best baseline.

3. Cold-start items

   CCGNN greatly enhances cold-start item performance. The *Recall*@100, *MRR*@100, *NDCG*@100 on cold-start items are increased by 11.1%, 13.5%, 9.8%, respectively.

[Paper #14]

EXTR: Click-Through Rate Prediction with Externalities in E-Commerce Sponsored Search

I. Target Problem

Click-Through Rate (CTR) prediction, estimating the probability of a user clicking on items, plays a key fundamental role in sponsored search. E-commerce platforms display organic search results and advertisements (ads) together as a mixed list. The items displayed around the predicted ad, i.e. external items, may affect the user clicking on the predicted. Previous CTR models assume the user click only relies on the ad itself, which overlooks the effects of external items, referred to as external effects. During the advertising prediction, the organic results have been generated by the organic system, while the final displayed ads on multiple ad slots have not been figured out, which leads to two challenges: 1) the predicted (target) ad may win any ad slot, bringing about diverse externalities. 2) external ads are undetermined, resulting in incomplete externalities. Facing the above challenges, inspired by the Transformer, they propose EXternality TRansformer (EXTR) which regards target ads with all slots as query and external items as key-value pairs to model externalities in all exposure situations in parallel.

II. Dataset

Since there is no publically available click dataset with context information for CTR prediction, the authors constructed a real-world dataset by collecting one-week impression logs and user clicks in November, 2021 from Taobao. The dataset contains large-scale records including 35.4 million users, 277.8 million queries, and nearly one million sellers, covering more than twelve thousand item categories such as clothing, electronic products, and fresh products. A total of 13.9 million ads and 67,901 million organic items are involved in the dataset.

III. Data Mining Workflow

The authors propose an efficient model to exploit personalized externalities in CTR prediction for the practical ecommerce sponsored search system. Inspired by the Transformer, which employs an attention-based querykey mechanism to allow for parallelization on input sequences, they design a new deep neural network, called EXternality TRansformer or EXTR, to jointly learn the diverse externalities for all ad exposure situations in a parallel way. The network is composed of two kinds of Transformer layers: self-attention Transformer layers focus on the interactions of external items and heterogeneous Transformer layers are responsible for the externality extraction. The self-attention layer is a classic structure which has been widely used in many fields, while the heterogeneous layer designs a querykey attention mechanism for different objects to learn the diverse externalities.

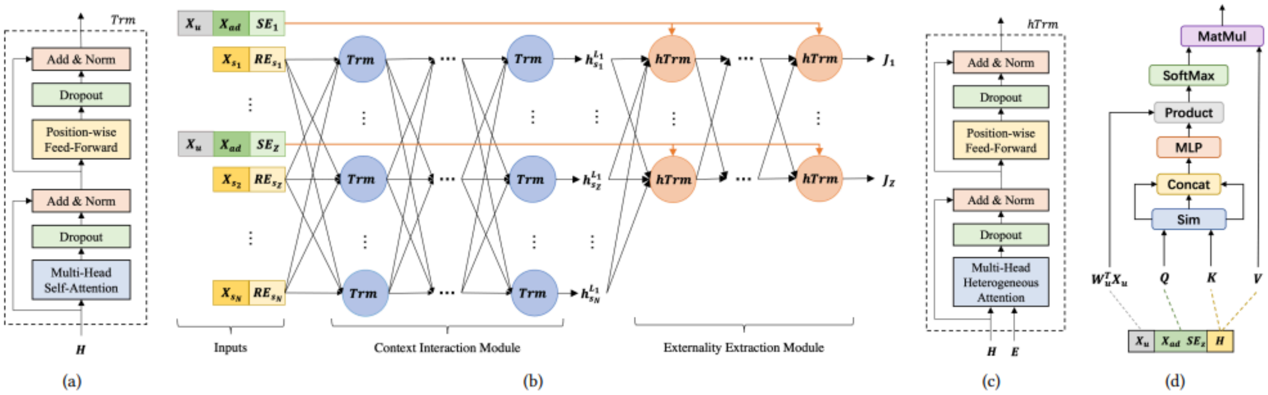


Figure 2: The whole model architecture of EXTR is shown in (b) which consists of two kinds of Transformer layers. (a) displays the self-attention Transformer layer Trm applied by Context Interaction Module. (c) shows the heterogeneous attention Transformer layer $hTrm$ employed by Externality Extraction Module. (d) depicts the heterogeneous attention of $hTrm$.

IV. Results

EXTR brings 0.0148 absolute AUC gain and 0.0263 absolute COPC gain over the SOTA independent baseline which is a significant improvement to our system. In summary, all these results suggest that external items can help CTR models make more accurate predictions, moreover, it is necessary to model the externalities on different ad slots.

[Paper #15]

ReprBERT: Distilling BERT to an Efficient Representation-Based Relevance Model for E-Commerce

I. Target Problem

Text relevance or text matching of query and product is an essential technique for e-commerce search engines, which helps users find the desirable products and is also crucial to ensuring user experience. A major difficulty for e-commerce text relevance is the severe vocabulary gap between query and product. Some neural networks have been proposed to solve the text matching task. However, for representation-based architecture, the encoding vectors may lose the fine-grained matching information, which causes degraded performance; for interaction-based models, they are mostly time-consuming and hard to be deployed online. Recently BERT has achieved significant progress on many NLP tasks including text matching, but it is a big challenge to deploy BERT to the e-commerce relevance task. To realize the goal, the authors propose ReprBERT, which has the advantages of both excellent performance and low latency, by distilling the interaction-based BERT model to a representation-based architecture.

II. Dataset

The large unlabeled dataset used for knowledge distillation is collected by randomly sampling from the search logs of Taobao within a year, which contains about 50 million query-product pairs. The samples are then annotated by the teacher model to generate soft labels for knowledge distillation. For the finetuning and evaluation of the proposed model, a large-scale human-annotated dataset is used. The dataset contains query-product pairs also sampled from the search logs, and then labeled Good (relevant) or Bad (irrelevant) by experienced human annotators. This is a daily task running in Taobao, which has accumulated more than one million labeled samples. The average length of the query and title is 7.3 and 32.6 Chinese characters on the whole annotated data, respectively.

III. Data Mining Workflow

1. ReprBERT Encoder
   - Uses BERT as the encoder to produce query and product embeddings.
   - Introduces a context-guided attention mechanism to better capture token-level relevance and collision, enhancing representations.
2. Late Interaction
   - Introduces query-product interaction after embeddings are generated.
   - Combines embeddings via addition, subtraction, and max pooling for finer-grained matching.
3. Intermediate Interaction
   - Performs fine-grained interaction at each layer, then applies weighted pooling to generate a comprehensive intermediate representation.

- Combines intermediate and late interaction results for more accurate relevance scoring.

4. Knowledge Distillation
    - Employs StructBERT as the teacher model, further trained on e-commerce data for domain-specific tasks.
    - Uses the teacher model to generate soft labels on large-scale unlabeled data for distillation training of the student model.
    - Fine-tunes the student model on annotated data using both soft and hard labels to narrow the performance gap with the teacher model.

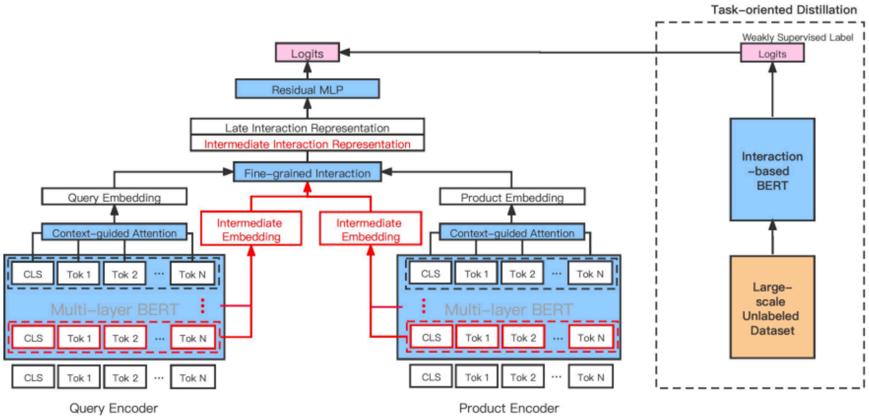


Figure 1: The illustration of ReprBERT. ReprBERT is a Siamese architecture with shared BERT encoders to encode query and product title. Their final embeddings and intermediate embeddings are extracted from the corresponding layers through attention mechanism. These embeddings are fed into the interaction module to get the interaction representations. Finally the residual MLP is used to predict the distribution of target classes.

IV. Results

ReprBERT can achieve only about 1.5% AUC loss from the interaction-based BERT, but has more than 10% AUC improvement compared to previous state-of-the-art representation-based models. In the online evaluation, ReprBERT improves the number of transactions by about 0.6% on average, and daily human annotations indicate a 0.5% improvement in relevance rate. Now, ReprBERT has already been deployed on the search engine of Taobao and serves the entire search traffic, achieving significant gain of user experience and business profit.

The Summarized Table

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
1	2024	About 10% of all queries have spelling errors.	The authors refined pre-trained models(BART, T5, mT5). They employed no-teacher distillation to prune and fine-tune models. Deployment used a CPU/GPU-based Java framework with load balancing for testing and production.	1. Twitter Dataset: https://luululu.com/tweet/#cr 2. Webis Dataset: https://dl.acm.org/doi/pdf/10.1145/1146847.1146848 (AOL search logs) 3. https://aclanthology.org/2021.bea-1.4/	BART and T5 have achieved a 4% enhancement in F1 score compared to the baseline. The model achieved a 100% successful request service rate within real-time scenarios
2	2024	Current methods introduce pooling bias by mistakenly sampling false negatives, diminishing performance and business impact.	The paper adopts BHNS, which consists of two modules, sampling regularization and pseudo label generation, which are used to mitigate the issue of pooling bias and give each query product pair a pseudo label to train the cross-encoder model.	1. The benchmark dataset. (http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark) 2. Training and evaluating the cross-encoder model: from Instacart's ongoing data collection.	The experiment confirmed BHNS as effective for practical e-commerce use.

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
3	2023	No existing research work can completely exploit the information within the transaction networks of e-commerce platforms for group-based fraud detection.	They proposed Group-based Fraud Detection Network(GFDN) to detect fraud in real-world applications. The model consists of two parts: structural feature generation module and a community-aware fraud detection network.	1. "Ride Item's Coattails" attack detection: TC: https://tianchi.aliyun.com/dataset/123862 TB: on the platform Taobao 2. Dataset for STARS attack detection: made by https://ieeexplore.ieee.org/document/7837846	Experiments showed the superior effectiveness and efficiency of GFDN for group-based fraud detection.
4	2024	The ever-growing volume of products bombards online shoppers, making it difficult to identify items of interest.	The paper introduces C-STAR. They include shopping trajectory represented by PR-Graph, inter-trajectory distribution similarity, intra-trajectory semantic correlation, pre-training strategy, and downstream task support.	The dataset consists of 28 days of anonymized customer engagement data, processed for three tasks, and they collected 1M, 5M, and 5M data, respectively.	The extensive evaluation demonstrated the effectiveness of C-STAR, which enhanced personalized shopping experiences.
5	2021	Fashion articles ordered online do not always fit the customers.	The paper presents SizeFlags, a Bayesian model using customer return data and expert priors, human feedback, and computer vision to predict size issues and reduce fashion e-commerce returns.	The dataset used in this paper is large-scale, weakly annotated data derived from customer returns in the fashion e-commerce platforms.	The model showed a strong impact in robustly reducing size-related returns in online fashion in over 14 countries.

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
6	2023	Much of the recent work faced the issues of a cold-start problem, poor quality and low diversity of attributes in candidates.	The paper adopts the CADENCE framework to tackle key issues in Query AutoComplete systems.	The final query dataset has about 6 million unique queries, and the catalog dataset contains about 4 million unique titles.	CADENCE generated about 700K new offline queries, resulting in significant improvement in recall.
7	2022	With insufficient data, business decisions are often based on intuition without a detailed evaluation of the decision.	The paper uses a causal inference framework, ASPIRE, with machine learning, combining propensity score matching, doubly robust estimation, and optimization techniques to allocate air-shipping capacity efficiently, maximizing revenue and conversion rates.	They used more than 40 million product page views from an emerging marketplace over a period of three months in 2019.	ASPIRE showed a lift of +79 base points of revenue as measured through A/B testing in Amazon.
8	2024	To improve large-scale product search in Taobao by enhancing indexing efficiency, reducing information loss in pruning, and aligning search results with user preferences.	The paper adopts a hybrid approach combining static index pruning and semantic retrieval to improve product search on Taobao. It incorporates user search preferences to minimize information loss during pruning, using term-level analysis and the TermRank algorithm to capture higher-order term dependencies.	1.5 million interactions were sampled, with an inverted index containing 2 billion candidate items and a semantic index of 100 million.	SMIND shows an improvement over state-of-are methods. It reaches an improvement of 1.34% in Pay Order Count and 1.50% in Gross Merchandise Value.

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
9	2023	The inefficiency and limited accuracy in the two-stage re-ranking process, caused by inadequate context modeling and the evaluation-before-reranking issue	- Permutation-Level InterestBased End-to-End Re-ranking (PIER) - Fine-Grained Selection Module (FPSM) - Omnidirectional Context-Aware Prediction Module (OCPM)	1. Avito dataset (public) 2. Meituan dataset (industrial)	PIER gets CTR and GMV increase by 5.46% and 5.83% respectively.
10	2022	Optimizing markdown pricing strategies to improve profitability while addressing incomplete data and reliance on rule-based methods challenges	Promotheus comprises multiple components: - Ithax - Randomizer - Depth Optimizer - Feasible Region Construction	historical data (The paper didn't mention the source.)	Promotheus adopting Ithax successfully hit stock depth and stock value targets for every single markdown event.
11	2023	Detecting fraudulent transactions using user behavior sequences while addressing the scarcity of labeled data	Sequence As GEnes (SAGE) - Input Data Paradigm: Snapshots - Two-Stage Transformer Architecture - Two-Stage Pre-Training Technique (Sequential Mutation Task & Sequential Recombination Task)	1. Pre-training: online e-commerce scenarios data in Alibaba 2. Evaluation: e-commerce scenarios data in Tmall Supermarket, Tmall Global, Taobao, and Ali Health	At a precision of 0.9, SAGE achieved a recall rate increase of approximately +4.93% compared to other methods.

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
12	2024	Optimizing attribute value extraction by effectively combining outputs from multiple LLMs with varying strengths and weaknesses	- LLM-ensemble - Use Dawid-Skene Model to learn the weights	WalmartAge & WalmartGender datasets	LLM-ensemble achieves the highest accuracy compared to all other baseline models.
13	2023	Improving e-commerce search models by addressing semantic representation, scalability for large graphs, and accuracy on long-tail queries and cold-start products	CC-GNN: - Content Collaborative Graph - MetaPath-guided Message Passing - Supervised learning and Contrastive learning combination - Difficulty Aware Representation Perturbation - Counterfactual Data Supplement at contrastive learning	1. Industry-scale Product Query Search dataset 2. Amazon Sport Dataset	CC-GNN improves over the best baseline (i.e., LasGNN) in terms of overall *Recall*@100, *MRR*@100, *NDCG*@100 by 11.8%, 14.5%, and 16.3%, respectively.
14	2022	accurately predicting Click-Through Rates (CTR) for ads by modeling the external effects on e-commerce platforms	EXternality TRansformer(EXTR) - self-attention Transformer layers - heterogeneous Transformer layers	Real-world dataset: by collecting one-week impression logs and user clicks in November, 2021 from Taobao.	EXTR brings 0.0148 absolute AUC gain and 0.0263 absolute COPC gain over the SOTA independent baseline.

#	Year	Target Problem	Adopted Approach/ Techniques	Datasets	Results & Contributions
15	2022	Improving e-commerce text relevance by addressing the vocabulary gap between queries and products, while balancing performance and deployment efficiency.	- ReprBERT Encoder - Late Interaction - Intermediate Interaction - Knowledge Distillation	The search logs of Taobao in a year: contains query-product pairs	ReprBERT can achieve only about 1.5% AUC loss from the interaction-based BERT, but has more than 10% AUC improvement.

Comparative Study & Discussions

- **Pros & cons of the surveyed papers**
    - Paper #01

      The authors refined LSTM-based classification models, which achieved a significant improvement boost. It also balances accuracy with real-time latency constraints using techniques like model pruning. They also extended utility through a multilingual model, which improves cross-regional consistency. However, they only focused on spelling correction, which limits broader applicability in query information. Also, it depends on high-quality pre-trained models, which may not be universally accessible.

    - Paper #02

      This paper proposed an innovative strategy, Bias-mitigating Hard Negative Sampling, that effectively reduces false negatives. It demonstrates improvements in search model performance through experimental validation. In addition, the domain-agnosic design facilitates application across different e-commerce platforms. However, extensive fine-tuning may be required to adapt to specific datasets.

    - Paper #03

      The authors provided a robust GFDN framework to detect group-based fraud on bipartite graphs. It effectively combines structural and community-aware information for improved fraud detection. Also, it’s validated through large-scale experiments with significant performance improvements. Nevertheless, The framework is limited to detecting group-based fraud, with less emphasis on individual-level anomalies. Plus, it’s computationally intensive, which poses potential challenges for smaller-scale implementations.

    - Paper #04

      This paper introduced a versatile C-STAR framework for learning customer shopping trajectories. The framework is effective across multiple downstream tasks like recommendation and profiling. Also, the pre-training strategy enhances representation quality, which yields consistent performance improvement. However, embedding and pre-training processes are highly complex. In addition, there are also challenges in adapting to platforms with limited customer interaction data. In future work, the authors plan to investigate two significant directions. The first one is to enhance their model by incorporating multimodal information to enrich the object features, where the difficulty is in proposing an effective knowledge fusion methodology. Secondly, it’s worth exploring streaming methods through Continual Learning instead of re-training the model, considering that trajectory data undergoes continuous evolution.

- Paper #05
  The authors proposed SizeFlags, an innovative use of Bayesian model, integrating human expert feedback and computer vision. It demonstrates a measurable reduction in return rates and environmental impact through extensive experimentation across 14 countries. Also, it enhances customer satisfaction and platform profitability by addressing a significant e-commerce pain point. However, it limited application scope. It focuses primarily on fashion e-commerce and size-related returns. There are challenges in generalizing the model to other industries with less structured feedback data. The limitations of this work include the coarse definition of size and fit issues. Future work can explore extending the Bayesian model to a hierarchical model with a multinomial likelihood and Dirichlet before including finer-grained size and fit problems and more high-fidelity data such as article measurements.

- Paper #06
  The authors proposed CADENCE, which addressed the challenges of cold-start, concept drift, and low diversity in autoSuggest system. Plus, they employed neural language models and customized beam searches to improve query relevance and coverage. Nevertheless, extensive training data and computational resources may be required for optimal performance. Moreover, there is a risk of overfitting to specific categories due to the constrained generation process. Also, the next step should use the model to generate queries on the fly as the user types in and improve the AutoSuggest coverage further. The authors said that they would like to try open-source LLMs like OPT in the offline setting.

- Paper #07
  This paper proposed ASPIRE, which implements causal inference to optimize the trade-off between revenue and delivery costs. Also, it provided quantifiable benefits validated through extensive online A/B testing. What's more, the doubly-robust estimation improves decision-making accuracy. However, the paper is limited to air-shipping scenarios, which reduces applicability to broader logistics challenges. In addition, the computational complexity may hinder implementation in smaller markets or platforms.

- Paper #08
  The authors proposed SMIND to address term dependency challenges using "user-query-item" hypergraphs in this paper. The framework enhances semantic matching by addressing vocabulary mismatches, significantly improving metrics like Pay Order Count and Gross Merchandise Value. Also, it's scalable for billions of queries, which proves its utility in large-scale applications. However, it has high computational costs associated with building and maintaining hypergraphs. In this work, the semantic indexer was trained using only text corpora, whereas they also have rich image data. Therefore, in the future, they believe that building a multi-model semantic indexer can further improve the performance of their system.

- Paper #09
  The paper introduces PIER, an end-to-end ranking framework with two core modules: FPSM and OCPM. FPSM leverages SimHash to select top-K candidates from full permutations based on user behavior modeling; OCPM incorporates a novel omnidirectional attention mechanism to capture permutation context. Jointly trained with a comparative learning loss, PIER outperforms existing methods in offline and online tests and is deployed on the Meituan platform. However, its complexity and reliance on comparative learning may cause scalability and adaptation challenges in diverse applications.

- Paper #10
  The paper proposes Promotheus, an end-to-end machine learning framework for optimizing markdown. It adopted two markdown management solutions: a supply-side markdown algorithm for rational pricing without demand estimation, and a full price elasticity-based framework optimizing for profit. Both markdown systems achieve superior profitability, with improvements of 86% (Promotheus) and 79% (Ithax) relative to manual strategies. However, the dependency on historical data for price elasticity modeling may not generalize well to novel products and rapidly changing market conditions. Also, the dual-system approach may require great integration effort for effective deployment.

- Paper #11
  The paper introduces SAGE, a general framework for user behavior modeling in fraud transaction detection. By applying the concept of "Sequence As Genes," the authors focus on intercepting key segments of long-term user behavior sequences to handle varying sequence lengths. They propose the sequential mutation and sequential recombination tasks to generate robust long-term behavior representations. The method has demonstrated effectiveness in fraud detection tasks. Future work could further improve performance by incorporating GNNs to model complex relationships between users, which may improve fraud detection accuracy.

- Paper #12
  The paper proposes the LLM-ensemble, a method designed to combine the outputs of multiple large language models for e-commerce product attribute extraction. It improves accuracy and efficiency by iteratively learning the weights for different models and aggregating their predictions. The potential defects may originate from the computational complexity of combining multiple LLMs, especially for large-scale e-commerce systems with high throughput requirements. The possible improvement includes optimizing model selection by including only the most diverse and complementary LLMs to reduce redundancy, and using knowledge distillation to train a smaller, single model that approximates the ensemble’s performance for deployment.

- Paper #13
  The paper introduces CC-GNN, an efficient graph learning method that demonstrates superior performance in industry-scale product search, addressing challenges like long-tail queries and cold-start items. Key components such as the Content Collaborative Graph and Counterfactual Data Supplement significantly enhance baseline GNN models and recommendation systems across various datasets. However, while CC-GNN is claimed to handle industry-scale graphs efficiently, detailed evaluations of computational overhead and scalability for larger datasets remain unclear.

- Paper #14
  The paper proposes an efficient framework EXTR for exploiting personalized externalities in e-commerce sponsored search systems for CTR prediction. By designing a Transformer-based model to jointly learn diverse externalities across all ad exposure situations and introducing the Potential Allocation Generator to optimize ad slot allocation, the model outperforms state-of-the-art baselines. One potential limitation of EXTR is its reliance on large-scale real-world datasets for training, which may limit its applicability in smaller platforms or scenarios with less data.

- Paper #15
  The paper introduces ReprBERT, a model for measuring semantic relevance between queries and products in e-commerce. By applying knowledge distillation from the original BERT model and integrating context-guided attention with interaction techniques, ReprBERT achieves performance close to the original BERT while enabling low-latency deployment for high-traffic environments. A potential limitation of ReprBERT is its reliance on knowledge distillation, which may lead to some performance loss compared to the original BERT model. Future improvements could focus on enhancing the teacher model and exploring more advanced distillation techniques to boost model performance.

- **Possible improvements and extensions of e-commerce**

These papers highlight significant enhancements in different e-commerce domains, including **logistic optimization, search accuracy, customer behavior analysis, product attribute extraction, and fraud detection**. Building on these innovations, there are opportunities to further transform e-commerce platforms into more efficient, customer-centric, and secure ecosystems. For us, e-commerce is a common field that most people are exposed to. Therefore, it's necessary to improve user experience, facilitate its convenience, and prevent anything harmful to the customers.

In **logistics**, integrating casual inference frameworks like ASPIRE with predictive analytics could streamline decisions across multiple delivery modes, such as ground or drone delivery, thereby expanding the scope beyond air shipping, which would also lift the profitability of online shopping platforms. Advanced modeling approaches could also incorporate dynamic real-time data, enabling more agile responses to fluctuating demands and constraints, such as sudden surges in order during sales events. By doing so, the platform manager's decision-making might be based on something other than intuition.

What's more, semantic and structural insights from models like SMIND, ReprBERT, and CADENCE can be extended for **search and recommendation systems** to create highly personalized experiences that account for user intent and product diversity. PIER's modular design also allows potential integration with real-time feedback mechanisms or multi-modal data like images and text for richer context modeling on recommendation systems. Future researchers can blend fine-tuned transformer models to reach spelling correction with sophisticated indexing systems. In this way, the platforms could ensure seamless query handling, even for ambiguous or incomplete user queries. Also, the user experience would be better than now. Users would not spend more time on retype products again and again.

Moreover, in terms of **profitability** on e-commerce platforms, representation learning frameworks such as C-STAR could deepen customer insights, improve recommendations, and, therefore, lift the profitability. Graph-based learning frameworks such as CC-GNN could also enhance search relevance, improve recommendations for long-tail queries and cold-start items, therefore boost the overall performance and profitability of e-commerce platforms. We can see that data mining techniques enable e-commerce platforms to gain deeper insights into customer behavior, leading to more personalized recommendations and optimized search results. Thus significantly enhance user engagement and overall profitability.

In another direction, **fraud detection**, leveraging group-based detection frameworks like GFDN could be enhanced with cross-platform collaboration, creating models that identify fraud patterns. This effectively prevents customers from being fraudulent. An innovative user behavior modeling framework like SAGE has also improved fraud transaction detection by capturing key segments of long-term behavior sequences. For us, fraud detection can also reduce unnecessary returns or packages, which leads to less waste.

More importantly, **sustainability** is a critical issue in our lives. Take SizeFlags as an instance, it could significantly reduce the return rates, which reduces carbon footprint and the waste of packages. In the future, researchers can develop a framework that is not limited to size-based return issues, which would advancely lower the carbon footprint. In my opinion, researchers may be able to build a framework to plan the delivery routes to minimize carbon emissions.

Last but not least, **a framework of an ML model** can be built to recommend optimal packaging configurations for items based on size, weight, and fragility. Researchers may also create a framework for recommending green choices first to encourage eco-friendly purchasing. Also, when users go on a shopping platform, researchers could build a model-driven platform to promote resale, recycling, and refurbishing to reduce the waste of products.