时尚产业应用中的可复用自注意力推荐系统

Marjan Celikik
marjan.celikik
@zalando.de Zalando SE 柏林，德国

Jacek Wasilewski
jacek.wasilewski@zalando.de
Zalando SE 柏林，德国

Ana Peleteiro Ramallo
ana.peleteiro.ra
mallo@zalando.de Zalando SE 柏林，德国

arXiv:2301.06777v1 [cs.IR] 17 Jan 2023

摘要

大量关于在推荐系统领域应用自注意力模型的实证研究，都是基于标准化数据集进行离线评估并计算指标。此外，其中许多研究没有考虑侧信息，例如商品和客户元数据，然而深度学习推荐系统只有纳入大量异构特征，才能充分发挥其潜力。通常，模型也只用于单一应用场景。由于这些不足，即使以往研究具有参考价值，也未必能体现其在真实产业应用中的实际效果。本次演讲将通过展示实时实验结果，说明用户留存率最高可提升 30%，以助于弥合这一差距。此外，我们还将分享为时尚产业的多种应用构建可复用、可配置推荐系统的经验与挑战。具体而言，我们将聚焦于时尚灵感相关的应用场景，例如穿搭排序、服装搭配推荐和实时个性化穿搭生成。

CCS 概念

• **计算方法** → **神经网络**；• **信息系统** → **推荐系统**。

关键词

推荐系统、Transformer、时尚产业

ACM 参考格式：

MarjanCelikik、JacekWasilewski 和AnaPeleteiroRamallo。2022 年。《时尚产业应用中的可复用自注意力推荐系统》。载于第十六届 *ACM* 推荐系统会议（*RecSys '22*），*2022*年*9*月*18*日至*23*日，美国华盛顿州西雅图。ACM，美国纽约州纽约市，4 页。https: //doi.org/10.1145/3523227.3547377

1 引言

Zalando 的主要目标之一是成为时尚灵感的起点。实现这一目标的一种方式是提供编辑内容，其中创作者展示新颖时尚的穿搭创意，并展示算法生成的穿搭，为顾客已购买、浏览或加入愿望清单的商品进行搭配。除了内容质量之外，提升参与度和客户留存率的主要驱动力，是在客户旅程中的不同客户触点上提供个性化的穿搭体验，见图1。从历史上看，这些客户触点分别由针对不同用例的端到端系统提供服务（例如，针对长期兴趣的排序和推荐、会话内推荐、相关商品推荐等），导致系统复杂度增加，维护成本显著上升。我们的场景还有另一层复杂性：客户在客户旅程中不仅会与穿搭互动，也会与其他时尚实体互动，例如时尚商品或创作者。此外，每种互动都可能传递不同的信号（例如点击、加入愿望清单或购买）。最后，由于平台上只有部分客户会与穿搭互动，我们还面临冷启动问题。

允许为个人或课堂使用而免费制作本作品的部分或全部内容的数字副本或纸质副本，但不得为盈利或商业利益而制作或分发副本，且副本首页必须包含此声明及完整引文。必须遵守本作品中第三方组件的版权规定。如需用于其他用途，请联系所有者/作者。*RecSys '22*，*2022*年*9*月*18*日至*23*日，美国华盛顿州西雅图 © 2022 版权所有者/作者所有。ACM ISBN 978-1-4503-9278-5/22/09。
https://doi.org/10.1145/3523227.3547377

2 模型架构

与“不同的推荐系统适用于不同用例”这一普遍看法相反， [9]，我们展示了，单个基于Transformer的推荐系统可以在来自不同来源的多种交互数据上进行训练，并在许多相关用例中重复使用，例如针对短期兴趣的基于会话的推荐，以及基于长期用户偏好的个性化排序。这有两项优势：第一，显著扩大训练数据集；第二，减弱反馈循环，即推荐系统使用前一天同一推荐栏的数据进行训练。反馈循环可能会放大流行度和展示偏差等偏见 [1, 4, 9]。同一个模型还可利用会话中的其他类型交互，为部分冷启动用户提供服务；对于没有交互数据的完全冷启动用户，则可利用有关用户的“静态”上下文信息。

我们将每位用户建模为一个交互序列，其中每次交互都可以涉及不同的实体（例如，商品、穿搭或创作者），交互类型也各不相同（例如，点击、购买、加入愿望清单）。我们在序列中的每个商品上进行训练，但只对其中一种实体（例如套装）进行预测；得益于引入的目标布尔掩码，只有这些实体会计入损失。对于我们的**推荐用例**，我们采用标准的Transformer编码器 [3, 5, 11]，并使用因果语言模型（CLM）方法进行训练，不过该方法不依赖于训练逻辑，也可以使用掩码语言建模（MLM）进行训练 [3, 10]。对于**个性化穿搭生成用例**，我们采用 [2, 6] 中所示的标准编码器-解码器（序列到序列）Transformer架构的变体 [11] 。模型使用相同来源的交互序列进行训练。然而，由于后者模型的一项主要任务是学习不同商品之间的时尚搭配兼容性，因此馈入解码器的目标是

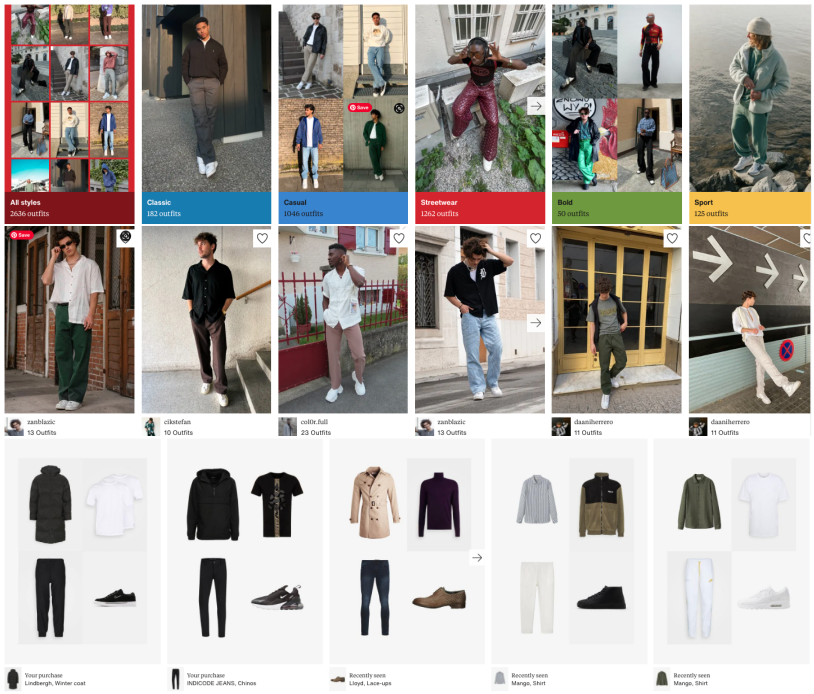


图1：“穿搭目录中的个性化排序（左上）和风格预览轮播，展示不同风格的穿搭（右上）；‘Get-the-look’轮播根据顾客近期的交互展示会话内穿搭推荐（中间）；算法时尚助手（AFC）展示根据顾客近期购买和浏览记录生成的算法穿搭（底部）。

用户互动过的穿搭中的商品，而穿搭交互之前的交互则馈入编码器。

2.1 输入嵌入

输入序列中的每个实体都表示为其异构输入的串联，例如与类别特征对应的学习嵌入。缺失的特征用 0 填充，例如，穿搭可能有创作者，而单件物品则没有。由于一套穿搭由多个物品组成，我们采用一种简单的表示方法：对来自不同物品的同一类别特征的嵌入取平均值。序列中的单件物品表示为仅含一个物品的穿搭，而创作者则视为一组穿搭。我们将交互特征的独热编码与实体嵌入串联起来。用户上下文编码为一个嵌入，其维度与物品嵌入相同，并置于 Transformer 的第一个位置（或多个位置）。

2.2 长期与短期兴趣的会话建模

交互序列由不同的会话组成。会话是在给定时间范围（例如一天）内发生的一系列交互，此时用户的意图明确；而用户的兴趣在不同会话之间可能发生剧烈变化。忽略这种结构而直接对序列建模会降低性能，我们在实验中也观察到了这一点。我们通过引入时间输入来对会话建模，具体形式为交互新近度，其定义为

操作时间戳与模型训练/服务时间戳之间的天数。我们对新近度进行离散化，仅考虑时间戳的整数部分。每次交互都有各自的新近度，并将其与其他物品特征拼接。因此，注意力机制能够从序列中选出与预测最相关的交互。图 2 总结了建模方案以及训练所用的不同数据来源。

3 在线实验

表1和表2展示了在线实验的结果：在图1所示的不同场景下，我们将推荐算法和穿搭生成算法与多种现有算法（LTR [7], 基于CNN嵌入的kNN、孪生网络 [12]）进行了比较。我们的新算法以较大优势优于所有基线方法。

4 实用设计选择与可扩展性

我们设计了一种灵活且可配置的推荐系统，得益于后端嵌入的可配置过滤逻辑，它可以在不同用例中重复使用。我们的设计被证明尤其务实且可扩展，并遵循 **输入 → 评分 → 过滤 → 重排序** 框架。我们的系统既能处理实时交互数据，也能处理按日提供的交互数据。系统将 Transformer 最后一个位置的最终输出输入 softmax层，为各项目评分。其他排序函数，例如

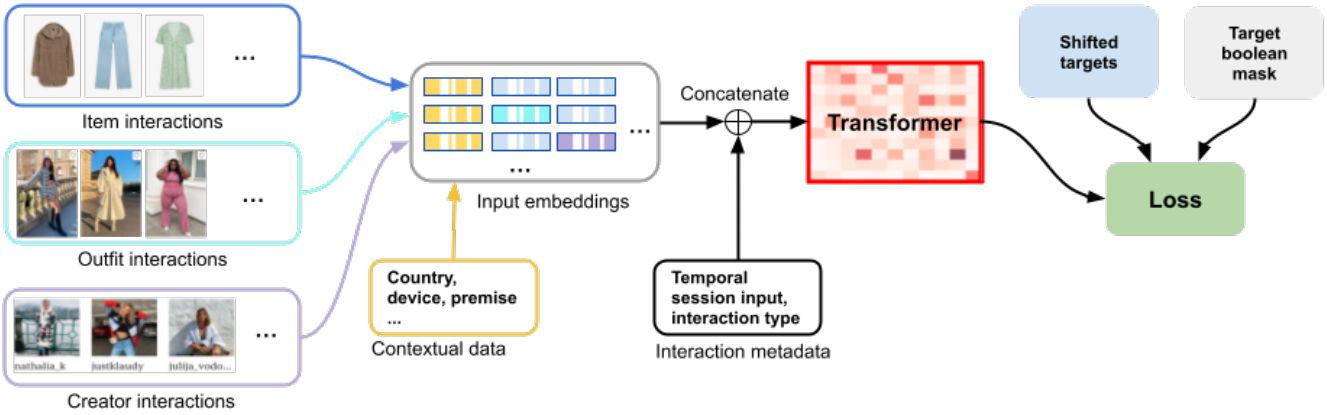


图2：在不同交互实体和实体类型上训练的模型概览，这些实体和类型被转换为结构相同的嵌入。时序会话和交互数据直接与学习嵌入拼接。使用掩码，仅考虑特定实体类型（例如套装）的目标输出。

表1：A/B测试结果显示，与图1所示不同场景下的现有算法相比，我们的基于Transformer的推荐系统提升了用户参与度（和留存率）。

细分群体	套装目录（LTR）	风格预览轮播（LTR）	Get-the-look（卷积神经网络-近邻算法）
冷启动	+42.0%	+39.7%	+109.4%
现有方案	+27.0%	+23.1%	+130.6%
All	+28.5%	+23.9%	+130.4%

表2：AFC 上的 A/B测试结果（图1，下方），比较以下方案的用户参与度：（1）用于个性化穿搭生成的 Transformer算法与现有的孪生网络 [12]。（2）带备用模型的同一 Transformer 算法（用于应对业务规则过滤导致的供应不足）：与孪生网络相比，该版本效率更高，能在输出层预测单个嵌入。

	(1) Transformer 与孪生网络	(2) Transformer +带嵌入层的 Transformer 与 Transformer + 孪生网络
参与度	+5.79%	+1.54%
留存率	+11.20%	+4.9%

例如 BPR、top1 [8] 和二元交叉熵 [5] 在我们的场景中效果较差。过滤和重排序逻辑会针对不同用例分别应用。重排序阶段包括应用新鲜度规则和多样化。新鲜度规则的示例包括通过物品年龄特征开关进行重排序、采用年龄指数衰减等。在我们的案例中，后者尤其有效，因为它在不损害相关性的同时，大幅增加了 Top-k 推荐中的新鲜内容。用于丰富结果列表的示例包括基于相似度的多样化启发式方法，以及通过引入多臂老虎机框架进行采样来平衡探索。

在典型的微服务后端架构中，传递给服务端点所需的数据必须从内存数据库中获取，并转换为特征，再通过与训练期间相同的逻辑输入模型。由于交互数据可以是穿搭或创作者，而二者都被视为物品集合，因此调用次数以及需要通过网络传输的数据量都会大幅增加

我们的场景。在服装搭配生成场景中，这一问题尤为严重：由于服装搭配生成过程具有自回归特性，后端服务必须等待套装中的下一个单品。这显著增加了后端复杂度和网络开销。因此，为避免服务与训练逻辑重复、降低网络延迟并简化后端，我们将输入数据处理与实际模型耦合起来。这意味着实际实体数据以映射的形式存储在与模型相同的服务中，因此训练和推理使用相同的逻辑。请注意，在我们的场景中，这种方法没有导致数据缺失或数据陈旧问题，因为我们每天都会重新训练模型并更新实体数据。服务端点只接收原始交互 ID。得益于这些设计选择，与基于 TensorFlow Ranking [7] 的现有排序系统相比，我们的系统 P99 延迟大约降低了一半，扩展能力也显著更强（在流量相同的情况下，约提升 3 倍）；后者由于扩展性问题，每次请求只能对 1000 套穿搭重新排序。

5 演讲者简介

Marjan Celikik 是 Zalando 的首席应用科学家，致力于个性化、推荐系统以及大规模机器学习项目产品化等相关问题。他拥有弗赖堡大学和马克斯·普朗克计算机科学研究所的信息检索与搜索引擎博士学位，以及萨尔大学的计算机科学硕士学位。

参考文献

[1] Allison J. B. Chaney、Brandon M. Stewart 和 Barbara E. Engelhardt。2018年。推荐系统中的算法混杂如何增加同质性并降低效用（*RecSys '18*）。计算机协会，美国纽约州纽约市。[2] Qiwei Chen、Huan Zhao、Wei Li、Pipei Huang 和 Wenwu Ou。2019年。阿里巴巴电子商务推荐中的行为序列 Transformer。载于第一届高维稀疏数据深度学习实践国际研讨会（阿拉斯加州安克雷奇）（*DLP-KDD '19*）。计算机协会，美国纽约州纽约市。[3] Gabriel de Souza Pereira Moreira、Sara Rabhi、Jeong Min Lee、Ronay Ak 和 Even Oldridge。2021年。Transformers4Rec：弥合自然语言处理与序列／基于会话的推荐之间的鸿沟。载于第十五届 *ACM* 推荐系统会议（*RecSys '21*）。计算机协会，美国纽约州纽约市。

[4] Pedro Nogueira 等。2021年。针对在线结果的离线评估决策批判性分析：实时推荐案例研究。（2021年）。[5] Wang-Cheng Kang 和 Julian McAuley。2018年。自注意力序列推荐（*ICDM '18*）。[6] Alexander Lorbert、David Neiman、Arik Poznanski、Eduard Oks 和 Larry Davis。2021年。可扩展且可解释的服装搭配生成。载于*CVPR 2021* 时尚、艺术与设计计算机视觉第四届研讨会。[7] Rama Kumar Pasumarthi、Sebastian Bruch、Xuanhui Wang、Cheng Li、Michael Bendersky、Marc Najork、Jan Pfeifer、Nadav Golbandi、Rohan Anil 和 Stephan Wolf。2019年。TF-Ranking：用于学习排序的可扩展 TensorFlow 库。[8] Massimo Quadrana、Alexandros Karatzoglou、Balázs Hidasi 和 Paolo Cremonesi。2017年。利用分层循环神经网络个性化基于会话的推荐。载于第十一届 *ACM* 推荐系统会议论文集（意大利科莫）（*RecSys '17*）。计算机协会，美国纽约州纽约市。[9] Harald Steck、Linas Baltrunas、Ehtsham Elahi、Dawen Liang、Yves Raimond 和 Justin Basilico。2021年。推荐系统深度学习：Netflix 案例研究。《*AI Magazine*》42（2021年）。[10] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou 和 Peng Jiang。2019年。BERT4Rec：基于 Transformer 双向编码器表征的序列推荐（*CIKM '19*）。计算机协会，美国纽约州纽约市。[11] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。《Attention Is All You Need》（*NIPS '17*）。美国纽约州雷德胡克。[12] Ruomei Wang、Jianfeng Wang 和 Zhuo Su。2022年。通过互补服装匹配学习服装搭配推荐的兼容性知识。《计算机通信》（2022年）。