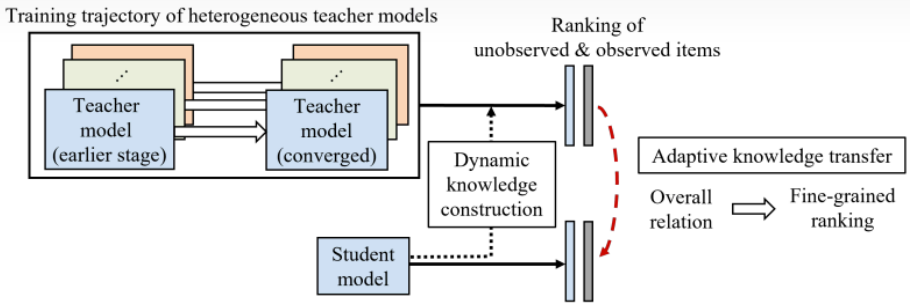
知乎



微软-2023：异构图模型蒸馏方法引领推荐系统新革命

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

3 人赞同了该文章

Introduction

RS使用个性化排序列表来辅助决策制定。过去的几十年里，研究人员探索了矩阵分解和图神经网络[+]等异构模型以生成高质量的列表。这些模型有不同的归纳偏置[+]，可能导致它们更倾向于特定假设而非其他假设，从而更好地捕获某些用户/物品偏好。组合多个模型可显著提高准确性，但可能因为计算延迟比单个模型高而难以应用于实时服务。 本文主要贡献如下：

- 通过动态知识构造监督学生模型，该模型采用由易到难的顺序排列。其首先识别每个教师路径中的适合知识，随后动态构建指导学生模型的目标知识。
- 如何转移：使用自适应知识转移，该方法根据学生的学习状态调整蒸馏目标。首先训练学生模型关注整体关系，然后学习精细级别排序。此外，还引入了一种新的转移策略，利用了未观察到的用户-物品交互知识。
- 解决异构模型中的知识转移排名难题：易到难的新视角。
- 提出新框架HETCOMP，通过有效压缩异构模型实现紧凑化，显著降低模型集合计算负担且保持高精度。
- 我们使用实证研究验证了提出的优越性，并进行了全面的方法分析。

Preliminaries

Problem Formulation

定义用户集合$\mathcal{U}$和物品集合$\mathcal{I}$。根据隐含的用户-物品交互（例如点击）历史，学习每个用户-物品对的排序得分的推荐模型是

$$f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}。$$

基于这些预测分数，推荐系统为每个用户提供一个他们未观察到的前K个物品的排名列表，称为前K推荐。我们提出了一种方法，可以将一组复杂的教师模型

$$\mathcal{F} = \{f^1, f^2, \ldots, f^M\}$$

压缩成轻量级的学生模型f。这个学生模型在推理上需要较少的计算，适合于实时服务和资源有限的环境。我们的目标是提供一种无模型依赖的方法，使得服务提供商可以根据他们的环境选择他们喜欢的模型来使用。

Ranking Matching Distillation

使用排名匹配蒸馏训练学生模型来模拟教师的排序。主要策略是使用Plackett-Luce模型为每个排序分配一个概率，并训练学生使他们的排序最大程度地接近教师的排序。近期研究只关注匹配顶级项目的排名，而不考虑其余项目的排名。列表级KL损失定义为对$[P; N]$的排序概率的负对数似然[+]。

给定项目集合 P 和剩余项目集合 N，学生模型学习在 P 中保持精确排序，同时惩罚在 N 中排名低于 P 中最低项的策略。在 N 中的项顺序可能不受保留。

Study on Ranking Knowledge Distillation

研究发现，在压缩异质教师模型时，提取其ensemble知识是一个挑战。为了解决这个问题，我们提出一种从教师中间训练状态获取线索的方法。该方法使用嵌入大小为6的学生模型MF，所有教师模型均使用嵌入大小为64的嵌入。这种方法在其他类型的学生成绩上也有类似的效果。此外，我们还通过蒸馏训练学生（公式）来实现这一目标。有关详细设置，请参阅相关章节。

Discrepancy.

$D(\pi, \pi^t) = \frac{1}{N}\sum_{i=1}^{N}(pi_i - pi_i^t)^2$

$D@K(\pi, \pi^t) = 1 - NDCG@K(\pi, \pi^t),$

$D@K(\pi, \pi^t) = 0$ 表示 学生模型完全保留 π^t 的前 K 排名；$NDCG$ 是一种广泛应用的列表排序评估指标，在此我们假设 π^t 是最优排序。

$$NDCG@K(\pi, \pi^t) = \frac{DCG@K(\pi)}{DCG@K(\pi^t)}, DCG@K(\pi) = \sum_{k=1}^{K}\frac{2^{y_{\pi_k}} - 1}{\log(k+1)}$$

定义每个元素的相关性（y_i）为评分（对于明确反馈）或二进制值（对于隐式反馈）。为了强调顶级项在π^t中的位置，我们使用参数几何分布：若i位于π^t的前K个，则$y_i = e^{-r(\pi^t, i)}/\lambda$，否则为0。其中$\lambda > 0$是控制分布尖锐性的超参数[+]。

Observations and analyses.

我们通过蒸馏方式训练学生模型，然后比较学生与监督的差异。具体来说，我们在表中展示了三种情况：(a)使用与学生相同的模型类型的教师；(b)使用六种初始化不同的均匀教师；(c)使用六种不同类型教师。对于每一种情况，我们使用给定的监督来计算学生模型的负对数似然（NLL），NLL越低，学生越能成功地模仿给定的监督。我们的实验结果显示，"压缩(c)比压缩(a)和(b)显著地增加"，这说明将异构教师的知识组合转移到学生的知识时，蒸馏的效果会明显下降。这个现象表明，学习过程中的排列顺序对于项目的排名非常重要。

Table 1: Discrepancy to the given supervision after KD.

Dataset	Supervision (Teacher)		Discrepancy		
	Type	Recall@50	D@10	D@50	NLL
Amusic	(a) Single-teacher (MF)	0.2202	0.6640	0.5167	0.5805
	(b) Ensemble (MF)	0.2396	0.6699	0.5162	0.6048
	(c) Ensemble (Het)	0.2719	0.7417	0.5958	0.7206
CiteULike	(a) Single-teacher (MF)	0.2604	0.5101	0.3716	0.5962
	(b) Ensemble (MF)	0.2763	0.5373	0.3910	0.5977
	(c) Ensemble (Het)	0.3144	0.6983	0.5269	0.6906

(c)中分析了每个教师学习情况。我们将每个收敛后的教师和其中间训练状态训练学生模型(MF)，中间训练状态是每个教师的四个训练状态(E1-E4)，每个状态对应于在25%,50%,75%,100%的收敛期后。图展示了差异。与使用相同模型(E4)的教师相比，使用不同模型(E4)的教师蒸馏可能会导致更大的差异。这是因为异构模型具有不同的归纳偏置，使模型更倾向于选择某些假设而非其他假设。因此，从异构教师学习可能具有挑战性，因为它需要学习那些不符合学生模型的关系。有趣的是，教师准确度高不一定导致更大的差异。例如，在CiteULike数据集上，GNN和AE的准确度相当，但从AE学习时，差异更大。这表明教师知识在训练过程中变得更加复杂，后期预测包含更多的多样性及个性化用户排名。这使我们提出使用教师学习轨迹作为学生的自然课程的动机。

METHODOLOGY

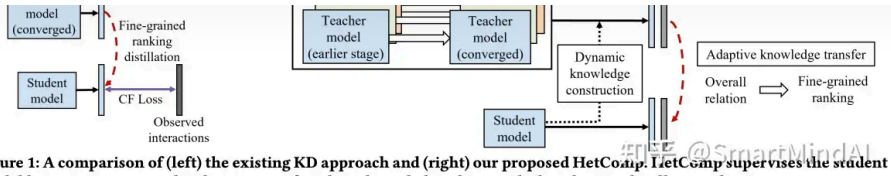


Figure 1: A comparison of (left) the existing KD approach and (right) our proposed HetComp. HetComp supervises the student model by using an easy-to-hard sequence of ranking knowledge along with the adaptive distillation objective.

Dynamic Knowledge Construction

1. 每个教师模型和用户的学习困难不同，需要考虑这种差异并进行排序。
2. 教师/学生模型无关，可以应用于各种模型类型，无需使用针对特定模型的统计或额外的保持数据性能指标。$\mathcal{T}$是教师训练轨迹的表示，用于表示每个教师在不同训练状态下对未观察到物品的排名。每个教师使用其在不同训练状态下的预测序列，即

$$T^x = [\pi^{x,1}, \ldots, \pi^{x,E}]。$$

最终排序$\pi^{x,E}$对应于收敛后的预测结果。我们的研究表明，教师的知识在训练过程中变得越来越复杂，因此模拟较简单的序列更困难。因此，我们从最简单的序列$\pi^{x,1}$开始，然后逐渐移动到复杂的序列$\pi^{x,E}$。我们使用v来表示M维的选择向量，其中每个元素$v[x]$表示教师当前选择哪个训练状态来引导学生模型。在训练过程中，我们根据v构建动态监督π^{d}。 算法中的知识构造部分可以详细说明。

通过学生模型f预测π 生成

$$\pi^{\mathrm{d}} = g(\{\pi^{x,v[x]}\}_{x\in\{1,\ldots,M\}})$$

$$\gamma^x = \frac{d[x]}{D@K(\pi, \pi^{x,(v[x]+1)})},$$

当开始学习$\pi^{x,v[x]+1}$时，我们需要注意到初始差异$d[x]$被视为常数。另外，随着学生训练过程的发展，$D@K(\pi, \pi^{x,(v[x]+1)})$会逐渐减小。$\gamma^x$衡量了学习难度下降的速度。之后，我们采用贪心策略[+]，即如果γ^x大于阈值α，就切换到下一个教师状态。阈值α控制着过渡速度，其值越高，过渡速度越慢。基于差异的贪婪策略考虑了学生学习状态和教师教学难度的变化，并且由于差异可以测量任何模型，所以该策略具有模型无关性。

$$\pi^{\mathrm{d}} = g(\{\pi^{x,v[x]}\}_{x\in\{1,\ldots,M\}})$$

在这里，可以灵活地使用各种聚合技术，从简单的平均值到具有可学习重要性的更复杂的聚合技术。

Adaptive Knowledge Transfer

Adaptive distillation objective

为了解决在学生训练期间目标排序可能会发生变化的问题，学习目标排序的详细信息既不必要也不现实。因此，我们首先使用$\mathcal{L}_O$训练学生模型来学习目标排序的整体关系。然后，一旦构建了最终收敛的预测中的目标排序（即对于所有的x，$v[x]$等于E），我们将转向学习精细级别排序，使用$\mathcal{L}_F$。为了实现这一目标，我们需要修改公式。蒸馏目标可以定义如下：

$$\mathcal{L}_O(P,N) = -\log \prod_{k=1}^{|P|} \frac{\exp\ f(u,P_k)}{\exp\ f(u,P_k)+\sum_{i=1}^{|N|}\exp\ f(u,N_i)},$$

仅考虑P中项的排名高于N中的项。

Transferring knowledge of observed/unobserved items

P^+是教师对学生观察到项的平均值。P^+在学生训练前产生一次，因观察到项集固定且$\mathcal{L}_O$不能传递详细排名。Distillation Loss定义如下：

$$\mathcal{L} = \mathcal{L}_{KD}(P^+, N) + \mathcal{L}_{KD}(P^-, N),$$

推离顶级排名。相反，它允许一些与观察到的项目高度相关的未观察到的项目自然地混入排名列表的顶端。 与CF损失不同的是，$\mathcal{L}_{KD}(P^+, N)$考虑了观察到的项目的排名，并且它不对顶级未观察到的项目进行惩罚。

随机初始化学生模型f 初始化选择变量$v_u[x] = 1$和 $d_u[x]$ $\forall x, \forall u$ 通过已收敛的教师集合获取 P_u^+ $\forall u$

The Overall Training Process

对每个用户进行知识构建。同时，直到学生从最终收敛的教师（即 $v[x]$ 等于 E，$\forall x$）学习为止，都会持续的知识构建和 $\mathcal{L}_O$ 进行。每经过10个时期，我们会执行一次知识构建，因为改变目标排列会耗费大量时间和资源，而这是不必要的。在附件中，我们详细分析了此算法的离线训练成本。

Experiments

实验设置如下： 使用 Amazon 音乐、CiteULike 和 Foursquare 数据集。将每个用户的交互随机划分为训练集、验证集✦和测试集✦，比例为 60%：20%：20%。 使用两种顶级排名指标：R\@K 召回率和 N\@K 奖励指标。 对于学生模型，使用 MF（基于邻接）、ML（基于流行）和 DNN（深度神经网络✦）算法。 对于所有教师模型，将用户/物品嵌入维度设为 64，而所有学生模型则设为 6，以使学生拥有老师使用的参数的一半或更少。这是为了保持与之前研究相同的参数配置。基线方法：我们将，除了MTD之外，所有方法都使用了由等同于（以**Ensemble**表示）生成的排名。

第1组基准包括一种基于点对齐的KD方法。

- RD 表示将顶级项目的权重转移。权重定义为每个项目在排名中的位置。

第2组：排名匹配的KD方法

- RRD 是推荐系统中的排名匹配方法，其损失函数为最高的排名物品。
- MTD：多教师知识蒸馏，该方法将知识传递给每个用户排名上的每个教师的可训练重要性。

第3组：高级方案提升蒸馏质量的最新排名知识蒸馏方法

- CL-DRD: 是目前最先进的文档检索知识图谱✦方法。它采用课程学习，通过定义绝对排名位置来预先设定学习难度。
- 在推荐系统✦领域，DCD是一种先进的知识图谱强化学习方法，它通过双矫正损失来修正学生未准确预测部分，并与RRD结合使用。

预测得分的KD方法不可行，因为教师和学生的得分分布不同。

Distillation Effects Comparison

Table 2: The recommendation performance comparison. *Imp* **denotes the improvement of HetComp over the best baseline.**

	Method	Amusic				CiteULike				Foursquare			
		R@10	N@10	R@50	N@50	R@10	N@10	R@50	N@50	R@10	N@10	R@50	N@50
	Best Teacher	0.0972	0.0706	0.2475	0.1139	0.1337	0.0994	0.2844	0.1392	0.1147	0.1085	0.2723	0.1635
	Ensemble	0.1096	0.0820	0.2719	0.1273	0.1550	0.1156	0.3144	0.1571	0.1265	0.1213	0.2910	0.1786
Student Model: MF	w/o KD	0.0449	0.0303	0.1451	0.0594	0.0568	0.0422	0.1372	0.0634	0.0726	0.0666	0.1806	0.1047
	RD	0.0522	0.0387	0.1602	0.0693	0.0610	0.0472	0.1514	0.0725	0.0778	0.0703	0.1921	0.1153
	RRD	0.0890	0.0659	0.2353	0.1077	0.0973	0.0740	0.2422	0.1113	0.0982	0.0905	0.2539	0.1446
	MTD	0.0901	0.0649	0.2279	0.1043	0.0993	0.0749	0.2425	0.1118	0.0955	0.0890	0.2402	0.1394
	CL-DRD	0.0883	0.0648	0.2375	0.1071	0.1033	0.0794	0.2512	0.1175	0.1001	0.0933	0.2528	0.1464
	DCD	0.0956	0.0675	0.2380	0.1079	0.1106	0.0851	0.2640	0.1246	0.1034	0.0965	0.2547	0.1491
	HetComp	**0.1036***	**0.0747***	**0.2469***	**0.1157***	**0.1379***	**0.1031***	**0.2814***	**0.1396***	**0.1118***	**0.1036***	**0.2722***	**0.1594***
	Imp	8.37%	10.67%	3.74%	7.23%	24.68%	21.15%	6.59%	12.04%	8.12%	7.36%	6.87%	6.91%
Student Model: ML	w/o KD	0.0447	0.0310	0.1522	0.0623	0.0210	0.0148	0.0859	0.0323	0.0184	0.0139	0.0804	0.0356
	RD	0.0706	0.0507	0.1874	0.0840	0.0835	0.0615	0.1914	0.0890	0.0729	0.0677	0.1811	0.1059
	RRD	0.0903	0.0643	0.2422	0.1074	0.0981	0.0701	0.2529	0.1116	0.0925	0.0813	0.2505	0.1366
	MTD	0.0843	0.0590	0.2293	0.1003	0.0944	0.0690	0.2519	0.1098	0.0909	0.0811	0.2440	0.1347
	CL-DRD	0.0866	0.0621	0.2409	0.1061	0.0989	0.0718	0.2583	0.1131	0.0931	0.0825	0.2541	0.1387
	DCD	0.0928	0.0653	0.2466	0.1086	0.1003	0.0724	0.2592	0.1144	0.0943	0.0845	0.2530	0.1399
	HetComp	**0.1020***	**0.0751***	**0.2470**	**0.1156***	**0.1251***	**0.0916***	**0.2686***	**0.1287***	**0.1039***	**0.0962***	**0.2645***	**0.1521***
	Imp	9.91%	15.01%	0.16%	6.45%	24.73%	26.52%	3.63%	12.50%	10.18%	13.85%	4.09%	8.72%
Student Model: DNN	w/o KD	0.0460	0.0324	0.1396	0.0597	0.0414	0.0339	0.1095	0.0518	0.0693	0.0665	0.1608	0.0987
	RD	0.0531	0.0378	0.1545	0.0670	0.0584	0.0445	0.1440	0.0671	0.0746	0.0683	0.1820	0.1060
	RRD	0.0851	0.0613	0.2255	0.1016	0.1034	0.0792	0.2552	0.1186	0.1016	0.0939	0.2584	0.1484
	MTD	0.0802	0.0563	0.2210	0.0958	0.0982	0.0710	0.2322	0.1058	0.0888	0.0797	0.2321	0.1305
	CL-DRD	0.0889	0.0623	0.2365	0.1047	0.1083	0.0816	0.2575	0.1183	0.1039	0.0983	0.2635	0.1536
	DCD	0.0919	0.0646	0.2404	0.1071	0.1114	0.0838	0.2668	0.1245	0.1066	0.1017	0.2671	0.1576
	HetComp	**0.1045***	**0.0768***	**0.2534***	**0.1190***	**0.1381***	**0.1050** *	**0.2864***	**0.1413***	**0.1136***	**0.1079** *	**0.2759***	**0.1642***
	Imp	13.71%	18.89%	5.41%	11.11%	23.97%	25.30%	7.35%	13.95%	7.17%	6.10%	3.29%	4.19%

表1展示了不同知识蒸馏方法训练的学生模型的推荐性能。此外，表2总结了最佳基线（即DCD）和提出的差异值。'最好老师'表示在每个数据集上所有异构教师中表现最好的教师模型。

服务是有益的。实验结果显示，该方法明显优于最优基准。

- 相较于点对点的KD方法（RD），其他直接转移排名顺序的KD方法表现出了更高的性能。这再次显示了排名知识在Top-K推荐中的重要性。一方面，MTD的表现相较于RRD有限，因为用户交互数据高度稀疏，导致用户特定的学习权重容易过拟合。
- 先进方案提高KD效果，但CL-DRD与学生模型间有显著差异，可能因其不考虑学生状态，使用预定义规则定义难度。DCD优于RRD，直接转移未预测内容实现稳定表现。
- 表1显示了ensemble和proposed的参数数和推断延迟。我们将mf的大小逐步增大，直到它达到与ensemble相似的性能。相比于ensemble，需要多次模型前向传播从而产生高昂推断成本的ensemble，proposed可以通过将知识压缩到紧凑的学生模型中显著降低推断成本。

Conclusion

我们提出了一种新的框架，通过将多模态推荐模型中的群智知识压缩到轻量级模型中，降低了推理成本并保持了高精度。我们将异质教师的精馏视为一项具有挑战性的任务，并发现教师的训练轨迹可以缓解学习难度。我们使用动态的知识构造和适应性知识转移，逐步提供更难的排名知识和更精细的排名信息。通过大量实验证明，我们的方法显著提高了学生的精馏效率和泛化能力。此外，由于该框架与现有的推荐系统具有良好的兼容性，我们期待它能成为解决推荐系统中准确性和效率权衡问题的一种解决方案。

发布于 2023-12-05 15:53 · IP 属地北京

微软（Microsoft） 工业级推荐系统 图神经网络（GNN）