Amazon推荐系统20年变迁

> *近期，IEEE Internet Computing上发表了一篇名为《亚马逊推荐系统二十年》的文章，提纲挈领地回顾了亚马逊推荐系统二十年来的发展，而这二十年的起点，就是基于物品的协同过滤算法，也就是ItemCF算法的发明时间，而文章的作者，也正是当年ItemCF的发明人。作为靠ItemCF算法养家糊口的从业人员，有必要学习一下“祖师爷”的训导。*

二十年以来，亚马逊一直致力于构建一个千人千面的商店。每个来到亚马逊网站的人看到的都不一样，因为网站针对他们的个人兴趣做了个性化。就如同你走进一个商店，商店架子上的商品开始重新排布，将你可能需要的排在前面，你不太可能喜欢的排在后面。

基于你当前的场景和你过去的行为，亚马逊的推荐系统从一个数以亿计的商品库中，为你挑选出少量你可能感兴趣的物品。背后的算法并不是什么魔法，它只是将其他人已经发现的信息与你共享。一切都由算法自动进行，在计算机的帮助下，人与人之间在隐性、匿名地互相帮助。

亚马逊在1998年上线了基于物品的协同过滤算法（下文简称ItemCF算法），将推荐系统推向服务百万级用户和处理百万级商品这样一个前所未见的规模。自从我们2003年在IEEE Internet Computing上发表关于这一算法的文章之后，该算法在互联网上开始广泛流传，包括YouTube，Netflix和其他很多公司在内都在使用。该算法的成功来源于以下几个方面：

- 简单、可扩展。
- 经常能给出令人惊喜和有用的推荐。
- 可根据用户的新信息立刻更新推荐。
- 可解释性强。

在我们2003年发表的文章中描述的内容这些年来曾经面对很多的挑战，同时也经历了极大的发展。在这里，我们介绍ItemCF算法的一些进展、改进和改良，同时也会阐述我们在协同过滤、推荐系统和个性化未来发展发展的一些看法。

算法

英文中常用the xxx来表示xxx的地位，例如乔丹在98年总决赛中的绝杀被称为the shot，本文这部分的标题叫做the algorithm，虽然作者本意并非如此，但译者觉得用这种方式来表示ItemCF算法在推荐系统中的地位也不为过。

如我们在2003年所描述的，ItemCF算法是很直观的。在90年代中期，协同过滤算法主要还是基于用户的，这意味着算法的第一步是要通过搜索所有的用户来计算某个用户在兴趣方面的相似用户（例如拥有相似的购买模式），之后再看这些相似用户看过哪些这个用户没

有看过的东西。与之相反，我们的算法第一步是计算每个物品的相关物品。这里的“相关”可以表示多种含义，但在这里，我们可以将其模糊地定义为“买了一个物品的人具有超乎寻常的可能性（unusually likely）会买另外一个”所以，对于每个物品i_1，我们希望得到所有购买了i_1的用户会以超乎寻常的频率一起购买的i_2。

其实UserCF也并不是一无是处，从计算形式上来讲它和ItemCF是完全对等的。UserCF适用于用户数的变化频率小于物品数的变化频率的场景，ItemCF则相反。当今的互联网环境下确实是更适合ItemCF发挥，但未来说不好也会有适合UserCF的场景。

一旦这张相关物品的表构建好，我们可以通过一系列的查找来构建推荐系统。对于一个用户当前场景下和历史兴趣中的每个部分，我们寻找到其相关物品，将它们结合起来得到用户最可能感兴趣的物品，过滤掉已经被看过或购买过的，剩下的就是就是待推荐的物品。

> *短短几句话就把推荐系统架构核心点透，祖师爷果然功力深厚。*

这个算法相比于旧的基于用户的协同过滤算法具有很多优势。最重要的是，主要的计算都是在离线发生的——相关物品的批量计算——而推荐的计算过程可以通过实时的一系列查找来完成。推荐结果质量高并且有用，尤其是数据量充足时。虽然二十年来各种新算法在不断被发明，在可观测到的质量方面，ItemCF仍然极具竞争力。该算法可无损地扩展到亿级用户和千万级物品，而不需要抽样或其他会影响推荐质量的手段。该算法在用户兴趣更新时可以立刻随之更新。最后，该算法的结果可以用很直观的方式来解释，因为其来源就是用户记得自己曾经买过的物品列表。

这个算法相比于旧的基于用户的协同过滤算法具有很多优势。最重要的是，主要的计算都是在离线发生的——相关物品的批量计算——而推荐的计算过程可以通过实时的一系列查找来完成。推荐结果质量高并且有用，尤其是数据量充足时。虽然二十年来各种新算法在不

断被发明，在可观测到的质量方面，ItemCF仍然极具竞争力。该算法可无损地扩展到亿级用户和千万级物品，而不需要抽样或其他会影响推荐质量的手段。该算法在用户兴趣更新时可以立刻随之更新。最后，该算法的结果可以用很直观的方式来解释，因为其来源就是用户记得自己曾经买过的物品列表。

2003年：亚马逊，Netflix，YouTube……

截至我们在2003年发表IEEE上的文章时，ItemCF已经在亚马逊广泛使用了。亚马逊在主页非常显眼的位置放置了基于你购买历史和浏览行为的个性化推荐模块。搜索结果页会给出和你搜索相关的推荐。购物车会给你推荐其他可以加入购物车的商品，可能会刺激你在最后一刻完成捆绑购买，或者对你已经打算购买的商品形成补充。在你订单的尾部，会出现更多的推荐，给出建议你之后可以购买的东西。借助电子邮件，列表页，商品详情页以及其他页面，很多亚马逊上的页面多少都会有些推荐模块，开始形成一个千人千面的商店。

在相关技术已经比较成熟的今天，仍然有很多网站大量依靠人工运营来决定网站上内容的排布，可见这种转换中需要克服的阻力和惯性之大。所以我们不仅要佩服亚马逊的技术能力，更要向其敢于革新的勇气和决心看齐。此外，这种"全面推荐化"的产品思路的意义不只是提升技术逼格，更重要的是，它提供了一种规模化的、持续可依赖的效果提升路线。从此销量或点击的提升不再取决于销售或运营人员的灵机一动，也不再受限于运营的人力，而是可以用算法系统化持续取得提升。

但有一个问题就是：开发推荐系统的成本是很高的，是一个链条，比如日志的管理，算法团队，工程师团队，代码落地等等...

很多其他公司和组织也在使用这个算法。在2010年，YouTube宣城他们使用ItemCF来做视频推荐。很多开源工具和第三方厂商都使用了这个算法，这使得该算法在网上零售、旅行、新闻、广告等行业中开始广泛出现。在后面的几年中，根据微软研究院的估计，亚马逊上大约30%的页面浏览来自于推荐系统。类似的，Netflix也在广泛使用推荐系统，他们的首席产品官声称80%以上的电影观看来自于推荐系统，并宣称Netflix推荐系统的价值每年高达十亿美元，还要多。

我们最初发明ItemCF的时候，亚马逊还只是一个网上书店。从那时起，亚马逊的销售额增长了不止一百倍，并且从图书扩展到以非出版物为主，从笔记本电脑到女装。这样的增长挑战着很多算法设计之初的假设，需要适应新的不断改变的大环境。通过一些经验，我们也找到了一些算法的改进方法，来为很多推荐系统新的应用计算出更加相关的推荐。

> *在机器学习技术逐渐兴盛的这个时代，经典算法的调优似乎在逐渐失去人们的重视，但从译者自己的经验来看，ItemCF算法的调优带来的效果提升空间非常大。在此还希望大家能够"重视基础，勿忘初心"。*

定义“相关”物品

推荐的质量很大程度依赖于“相关”的含义。例如，当我们说买了X之后具有“超乎寻常的可能性”会购买Y的时候，究竟是什么意思？当我们观察到用户同时购买了X和Y时，我们会好奇多少买了X的人会随机购买到Y——如果X和Y不相关的话。一个推荐系统说到底是一个统计学的应用系统。用户行为是包含噪音的，而我们面对的挑战就是如何在随机中发现规律。

要估计共同购买X和Y的用户数，N_{xy}，的一种直观的方法，是认为所有购买X的用户都有同样的概率$P(Y)$来购买Y，其中$P(Y)$=购买Y的人数/所有发生购买的用户数，那么购买X的用户数乘以$P(Y)$就可以认为是N_{xy}的一个期望值，记为E_{xy}。在我们2003年的文章中，以及在此之前的很多工作中，使用的都是类似的计算方法。

有趣的是，对于基于任意两个物品X和Y，购买了X的用户总要比整体用户更可能购买Y。这是怎么回事呢？想象一个超级剁手党——一个购买了商店中所有物品的人。当我们在寻找购买了X的用户时，这个用户总是会被选中。类似的，一个购买了1000件商品的用户总要比购买了20件商品的用户的被选中几率高50倍。所以从购买记录中随机采样得到的结果在用户维度上并不是均匀分布的，也就是说我们得到的是有偏的样本。对于任意物品X，购买了X的用户要比整体用户购买量更多。

> *之所以会出现这种现象，原因在于，根据作者的逻辑，买了X的用户中，很大概率会包含一些比普通人更能剁手的剁手党——例如文中那个超级剁手党就肯定会被选中，那么购买X的用户的购买量就被这些人拉高了，要高于整体用户的平均购买量。*

这种用户购买历史的非均匀分布，意味着我们在计算有多少购买了X的用户会随机购买Y时不能忽略是谁买了X。我们发现将用户建模成具有多次购买Y的机会会很有用。例如，对于一个有20次购买的用户，我们视其拥有20次独立的购买Y的机会。

更正式的，对于一个购买了X的用户c，我们可以将c购买Y的概率估计为$1-(1-Py)^{|c|}$，其中$|c|$代表用户c的购买次数减去其对于X的购买次数，
$Py = |Y$的购买次数$|/|$所有的购买次数$|$，代表任意一次购买是对于Y的购买的概率。之后，我们可以通过对所有购买X的用户进行汇总，再加上二项式展开，来计算购买X的用户中购买Y的用户数的期望值E_{xy}。

$$E_{XY} = \sum_{c \in X} [1 - (1 - P_Y)^{|c|}] = \sum_{k=1}^{\infty} \alpha_k(X) P_Y^k \qquad \text{其中} \quad \alpha_k(X) = \sum_{c \in X} (-1)^{k+1} \binom{|c|}{k}$$

我们可以将E_{xy}写作P_y的多项式，其系数只与X有关。实际中，P_y通常都很小，所以可以用一个上界k来做近似。此外，P_y和$\alpha_k(X)$可以事先计算好，所以任意两个X和Y的E_{xy}只需要对事先计算好的值进行简单组合即可到到一个近似值。

有了一个计算E_{xy}的健壮方法之后，我们可以用其来计算观测到的N_{xy}是否明显高于或低于随机。例如，$N_{xy} - E_{xy}$可认为是非随机共现的一个估计，而$(N_{xy} - E_{xy})/E_{xy}$则给出了一个非随机共现相对期望值的比例。这两个例子都可认为是衡量有多少用户会同时购买X和Y的相似度函数$S(X, Y)$。第一种方法，$(N_{xy} - E_{xy})$，会偏向于更流行的Y，例如第一本哈利波特，这会使得推荐结果看上去过于流行或无关。第二种方法，$(N_{xy} - E_{xy})/E_{xy}$，会使得低销量的物品很容易获得高分数，使得推荐结果看着过于奇怪或随机，大量低销量物品的存在使得这个问题尤为严重。所以相关性分数需要在这两者之间找到平衡点。基于$(N_{xy} - E_{xy})/\sqrt{E_{xy}}$的卡方检验就是这样一个平衡的例子。

第一种方法的问题较好理解。第二种方法会偏向低销量商品的原因在于$(N_{xy} - E_{xy})/E_{xy}$这个式子不够稳定，尤其是分母较小时（销量较低时），分子上的小幅度变化就会引起整个式子取值的较大波动。此外分母本身的取值也不稳定。活学活用概率统计的案例：将自己的问题构造成一个统计问题，剩下的交给数学就好了。如果我们能将遇到的度量问题都能成功地进行类似的抽象，很多工作就会简单很多。

除此以外还有一些其他方法和参数可用来衡量相关性，以及从相关物品中做出推荐。我们的经验是，没有那个得分是在所有场景下都最优。最终来讲，只有可观测的质量是推荐系统真正的评价标准，推荐系统只有在人们认为其有用时才是有用的。

机器学习和ABTest可以学习到用户真正的喜好，选择推荐中使用的最优参数。我们不仅可以衡量哪些推荐是有效的，同时我们还可以收集到哪些推荐被用户喜欢、点击和购买，并将这些信息再次输入到算法中，进一步学习那些对用户帮助最大。

例如，兼容性是一种重要的关系。我们可能会观察到购买了某型号数码相机的用户会有很高的几率会购买某特定型号的存库卡，但这并不能保证这张存储卡与这部相机兼容。用户会因为很多与热闹而购买存储卡，我们观测到的相关性可能是随机现象。确实，亚马逊的

商品库中有几十万中存储卡，这里面很多都与这部相机随机相关联 。很多电商网站使用人工编辑的兼容性数据库，而这是很昂贵并且容易出错的，尤其是在有亚马逊这个量级上。我们发现，只要有足够的数据，再加上一个衡量相关性的健壮方法，兼容性可能从人们的行为中学习出来，错误信号逐渐消失，而正确的物品逐渐浮现。

有趣的是，我们发现相关物品的含义会从数据中自己浮现出来，完全依靠用户自己。考虑用户浏览的物品和购买的物品的不同。对于书籍、音乐以及其他低价商品，用户倾向于浏览并购买类似的东西。但对于很多昂贵的物品来说，尤其是非出版物的物品，用户浏览的

和最终购买的会有很大的不同。例如，用户可能会浏览很多电视机，但是最终只会购买一个。他们在浏览这台电视机的同时也在浏览的其他物品，通常会是其他电视机。而他们在购买这台电视机的同时购买的其他物品，则更可能是这台电视机的配套物品，例如一台蓝光播放器或挂墙支架。

这种”自我发现“的数据模式是算法优于人工的另一重要原因，优于每个人固有的局限性，无法穷举所有可能的匹配模式，而且还会存在滞后性，但是机器可以，算法可以。只要有合适的度量方法和计算资源以及数据，所有有用的搭配模式都可以被发现，而且可以在模式刚刚出现时就将其找到。

时间的重要性

充分理解时间扮演的角色对于改进推荐系统质量有着重要的作用。例如，当计算相关物品时，两个物品的相关性很大程度上依赖他们在时间间隔的长短。如果一个用户在买了一本书的五个月之后又买了一本书，那这两本书之间的相关性就要弱于两本在同一天内被购买的书的相关性。时间的方向性也比较有用。例如，用户会在买了相机之后买存储卡，而不是反过来。这告诉我们不应该给购买了存储卡的用户推荐相机。有时物品的购买具有序列性，例如书籍、电影和电视剧，那么推荐就应该给出你下一步想要做的东西。

亚马逊的商品库一直在变化。每天，数千的新商品到来，而很多其他商品则逐渐失效或沉寂。这种循环在某些类别上尤其明显。例如，服饰具有明显的季节性，消费电子更新换代很快。由于没有足够用户行为数据来计算相关性，新物品会有一定的劣势。这种问题被称作冷启动问题，通常需要借助E/E的方式来给予新商品足够的曝光机会。新闻和社交媒体这些易过期的物品在冷启动方面尤其具有挑战性，通常需要融合基于内容的算法（使用题目，主题和文本等）和基于行为的算法（使用购买，浏览和打分等）。

用户维度的生命周期也存在冷启动的问题。在对用户兴趣缺乏足够了解的情况下如何给出

推荐一直都是一个问题。何时利用有限的信息以及何时使用热品来保证策略安全是一个不容易判断正确的复杂转换过程。

冷启动问题的解决不仅是个技术问题，也是个产品问题。例如可以考虑让用户选择一些喜好，但也要考虑用户是否接受这种形式。好的产品引导和设计可以帮助技术更好更快地解决冷启动问题。

即使对于信息完备的用户，正确地使用时间信息对于推荐质量也有着重要影响。随着年龄的变化，之前的购买对于用户当前的兴趣的影响越来越小。更复杂的是，不同类型的物品的减弱效应还各不相同。例如，像”波涛汹涌的大海航行指南“这样的购买记录代表的通常是

可持续的长期兴趣。其他的例如洗碗机修理工具这样的东西在周末的工程之后可能就不再相关了。甚至还有一些像拨浪鼓这样的商品需要随着时间不断变化；四年之后，我们应该推荐的是平衡车而不是奶瓶。还有一些商品，例如书籍，通常只会购买一次；其他的，例如牙膏，经常是被以可预期的周期重复购买。

文中提到的购买模式的问题，直到今天仍然没有得到完全的解决。例如周期性购买的问题，国内多家电商都做过类似尝试，但都未取得预期的效果，其中的复杂度，远比看上去的高。再例如大家经常吐槽的某些电商网站在你买了什么之后立刻就再推一遍什么，都是问题尚未完全解决的代表。

推荐的质量不仅取决于购买的时间，还取决于购买的内容。我们发现一本书的购买信息可以暴露很多用户的兴趣，让我们能够给出很多高度相关的推荐。但是非出版物类的购买，即使次数很多，也不能给我们提供用户的什么信息。我们能从一次订书机的购买记录中收集到什么信息？基于一双袜子的购买记录我们能给出什么样惊奇而有深度的推荐？当前来

说推荐胶带切割机或者更多的内衣或许是有用的，但是长期来说会导致推荐很无聊。所以，我们需要开发一种技术，能够识别哪些购买能提供有用的推荐而哪些应该被忽略。

> *自动识别这两种行为很具有挑战性，但是在自动解决之前，人为地通过策略进行缓解也是不错的选择。*

最后，推荐系统中多样性的重要性也是众所周知的；有时相比一个范围很窄的推荐列表，给出一些更多样的相关物品会更好。亚马逊丰富的商品库以及多样的商品类型，相比例如书店这样的垂类电商，提出了多样性方面更大的挑战。例如，给一个重度阅读爱好者推荐更多的书可能会带动更多的销量，但是从长期来看，让用户发现他们之前从未考虑过的产品线中的新商品可能是更有用的。意图的明确性也是多样性中的一个因素。当用户很明显是在寻找某个具体的商品时，推荐系统应该收窄范围帮助用户快速找到所需。但是当意图并不明确或确定时，探索性和新奇性应该是推荐的目标。想要找到推荐系统中多样性的正确平衡点，不仅需要实验，更需要一颗想要从长期进行优化的心。

这一部分中，作者集中提出了一些推荐系统的核心挑战，其中很多直到现在也没有完全得到解决。而这其中又不全是技术问题，还涉及很多产品设计和利益权衡，例如推荐多样性和E/E问题，其产品决策层面的难度要大于技术实现的难度。但另一方面，其中一些问题都可以通过上面提到的对ItemCF算法的优化来解决或缓解。

未来：推荐无处不在

推荐的未来将通向何方？我们认为未来的机会要比过去的机会更多。我们可以想象一种智能交互，使得购物就像对话一样简单。

这种方式超越了当前基于搜索和浏览的模式。相反，探索过程应该像和一位朋友聊天一样，这位朋友了解你，知道你的爱好，陪伴你的每一步，知道你的需求。

这是一种智能无处不在的愿景。每一次交互都会反映你是谁，你喜欢什么，同时帮助你找到其他和你类似的人已经发现的东西。当你看到和你明显不相关的东西时你会感到空虚和悲哀：难道你现在还不了解我吗？

要想达成这样的 愿景需要从新的角度思考推荐。不应该有推荐特征和推荐引擎。相反，读懂你和其他人，以及当前拥有的资源应该是每次交互都应该拥有的。

推荐和个性化生存在数据的海洋中，我们在穿梭世界的过程中创造了这些数据，包括我们找到的，发现的和喜爱的。我们坚信未来的推荐系统将继续构建在充分利用人类集体智慧的智能计算机算法的基础上。未来将继续是计算机帮助人类互助。

大概二十年前，亚马逊在百万级商品上构建了推荐系统，来帮助百万级用户，帮助人们发现自己无法找到的东西。从那时起，原始的ItemCF算法传播到了互联网的大部分角落，帮助人们寻找可观看的视频，可阅读的资讯，同时也被其他算法和技术挑战着，也被改造以提供更好的多样性、实时性、时间敏感性以及时序性等很多其他问题。由于其简便性、可扩展性、可解释性、可调性以及相对高质量的推荐，ItemCF算法在当今仍然是最为流行的推荐算法之一。

但是该领域仍然充满机会。千人千面的用户体验仍然是一个没有人能够完全做到的愿景。仍然有很多机会可以给系统的每个部分添加智能和个性化，制造一个懂你喜好，懂其他人喜好，同时也知道你有什么选择的老友般的体验。推荐即发现，通过帮助你发现来提供惊喜和快乐。每种交互都应该是推荐。