061 | WSDM 2018论文精读：看京东团队如何挖掘商品的替代信息和互补信息

2018-02-21 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 07:57 大小 3.65M

本周我们来精读 WSDM 的几篇论文，周一我们分享了一篇来自谷歌团队的文章，其核心是利用点击模型来对位置偏差进行更加有效的估计，从而能够学习到更好的排序算法。

今天，我们来介绍 WSDM 2018 的最佳学生论文《电子商务中可替代和互补产品的路径约束框架》（A Path-constrained Framework for Discriminating Substitutable and Complementary Products in E-commerce），这篇文章来自于京东的数据科学实验室。

作者群信息介绍

这篇论文的所有作者都来自京东大数据实验室，我们这里对几位主要作者做一个简单介绍。

第三作者任昭春（Zhaochun Ren）目前在京东数据科学实验室担任高级研发经理。他于 2016 年毕业于荷兰阿姆斯特丹大学，获得计算机博士学位，师从著名的信息检索权威马丁·德里杰克（Maarten de Rijke）。任昭春已经在多个国际会议和期刊上发表了多篇关于信息检索、文字归纳总结、推荐系统等多方面的论文。

第四作者汤继良（Jiliang Tang）目前是密歇根州立大学的助理教授。汤继良于 2015 年从亚利桑那州立大学毕业，获得计算机博士学位，师从著名的数据挖掘专家刘欢（Huan Liu）教授。他于 2016 年加入密歇根州立大学，这之前是雅虎研究院的科学家。汤继良是最近数据挖掘领域升起的一颗华人学术新星，目前他已经发表了 70 多篇论文，并且有四千多次的引用。

最后一位作者殷大伟（Dawei Yin）目前是京东数据科学实验室的高级总监。2016 年加入京东，之前在雅虎研究院工作，历任研究科学家和高级经理等职务。殷大伟 2013 年从里海大学（Lehigh University）获得计算机博士学位，师从信息检索领域的专家戴维森（Davison）教授。目前已经有很多高质量的研究工作发表。殷大伟和笔者是博士期间的实验室同学以及在雅虎研究院期间的同事。

论文的主要贡献

我们首先来看一下这篇文章的主要贡献，梳理清楚文章主要解决了一个什么场景下的问题。

对于工业级商品推荐系统而言，一般通过两个步骤来产生推荐结果。第一步，产生候选集合，这里主要是从海量的物品中选择出几百到几千款用户可能会购买的商品；第二步，利用复杂的机器学习模型来对所有候选集中的产品进行排序。

这篇文章主要探讨了如何能够更好地产生候选集产品，即如何更好地产生“替代品”（Substitutes）和“互补品”（Complements）来丰富用户的购买体验。

那么，什么是替代品和互补品呢？

根据这篇文章的定义，替代品就是用户觉得这些商品可以互相被替换的；而互补品则是用户会一起购买的。挖掘这些商品不仅对于产生候选集具有很重要的意义，也对于某些场景下的推荐结果有很好的帮助，比如当用户已经购买了某一件商品之后，给用户推荐其他的互补品。

虽然替代品和互补品对于互联网电商来说是很重要的推荐源，但并没有多少文献和已知方法来对这两类商品进行有效挖掘。而且这里面一个很大的问题是数据的“稀缺”（Sparse）问题。因为替代品或者互补品都牵扯至少两个商品，而对于巨型的商品库来说，绝大多数的商品都不是两个商品一起被同时考虑和购买过，因此如何解决数据的稀缺问题是一大难点。

另一方面，商品的属性是复杂的。同一款商品有可能在某些情况下是替代品，而在另外的情况下是互补品。因此，如何在一个复杂的用户行为链路中挖掘出商品的属性，就成为了一个难题。很多传统方法都是静态地看待这个问题，并不能很好地挖掘出所有商品的潜力。

归纳起来，这篇文章有两个重要贡献。**第一，作者们提出了一种“多关系”（Multi-Relation）学习的框架来挖掘替代品和互补品。第二，为了解决数据的稀缺问题，两种“路径约束”（Path Constraints）被用于区别替代品和互补品。**作者们在实际的数据中验证了这两个新想法的作用。

论文的核心方法

文章提出方法的第一步是通过关系来学习商品的表征（Representation）。这里文章并没有要区分替代品和互补品。**表征的学习主要是用一个类似 Word2Vec 的方式来达到的。**

也就是说，商品之间如果有联系，不管是替代关系还是互补关系，都认为是正相关，而其他的所有商品都认为是负相关。于是，我们就可以通过 Word2Vec 的思想来学习商品的表征向量，使得所有正相关的商品之间的向量点积结果较高，而负相关的向量点积结果较低。这一步基本上是 Word2Vec 在商品集合上的一个应用。

通过第一步得到的每个商品的表征，是一个比较笼统的**综合的表征**。而我们之前已经提到了，那就是不同的情况下，商品可能呈现出不同的属性。因此，我们就需要根据不同的场景来刻画产品的不同表征。**文章采用的方法是，对于不同类型的关系，每个商品都有一个对应的表征**。这个关系特定的表征是从刚才我们学到的全局表征“投影”（Project）到特定关系上的，这里需要学习的就是**一个投影的向量**。

第三个步骤就是挖掘替代关系和互补关系了。这篇文章使用了一个不太常见的技术，用“**模糊逻辑**”（Fuzzy Logic）来表达商品之间的约束关系。在这里我们并不需要对模糊逻辑有完整的理解，只需要知道这是一种把“硬逻辑关系”（Hard Constraints）转换成为通过概率方法表达的“软逻辑关系”（Soft Constraints）的技术。

在这篇文章里，作者们重点介绍的是如何利用一系列的规则来解决数据稀缺的问题。具体来说，那就是利用一些人们对于替代关系或者互补关系的观察。

比如，商品 A 是商品 B 的替代品，那很可能商品 A 所在的类别就是商品 B 所在类别的替代品。再比如，商品 B 是商品 A 的替代品，而商品 C 又是商品 B 的替代品，而如果 A、B 和 C 都属于一个类别，那么我们也可以认为商品 C 是 A 的替代品。

总之，作者们人工地提出了这样一系列的规则，或者叫做约束关系，希望能够使用这样的约束关系来尽可能地最大化现有数据的影响力。当然，我们可以看到，这样的约束并不是百分之百正确的，这也就是作者们希望用“软逻辑关系”来进行约束的原因，因为这其实也是一个概率的问题。

整个提出的模型最终是一个集大成的优化目标函数，也就是最开始的物品的综合表征，在特定的关系下的投影的学习，以及最后的软逻辑关系的学习，这三个组件共同组成了最后的优化目标。

方法的实验效果

这篇文章使用了京东商城的五大类商品来做实验，商品的综述大大超过之前亚马逊的一个公开数据的数量。作者重点比较了之前的一个来自加州大学圣地亚哥团队的模型，以及几个矩阵分解的经典模型，还比较了一个基于协同过滤的模型。

从总的效果上来看，这篇文章提出的模型不管是在关系预测的子任务上，还是在最后的排序任务上均要大幅度地好于其他模型。同时，作者们也展示了逻辑关系的确能够帮助目标函数把替代关系和互补关系的商品区分开来。

小结

今天我为你讲了 WSDM 2018 年的一篇来自京东数据科学团队的文章，这篇文章介绍了如何利用多关系学习以及模糊逻辑来挖掘商品的替代信息和互补信息，然后训练出更加有效的排序算法。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息；第二，我们详细介绍了这篇文章要解决的问题以及贡献；第三，我们简要地介绍了文章提出方法的核心内容以及实验的结果。

最后，给你留一个思考题，互补商品或者替代商品是双向关系还是单向关系，为什么呢？

欢迎你给我留言，和我一起讨论。

上一篇　060 | WSDM 2018论文精读：看谷歌团队如何做位置偏差估计

下一篇　062 | WSDM 2018论文精读：深度学习模型中如何使用上下文信息？

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。