浅谈基于协同过滤
算法的个性化推荐

姓名：

学号：

班级：

学院：

年　　月　　日

摘要

协同过滤是如今推荐系统中最为成熟的的一个推荐算法系类，　是利用群体的喜好来推测使用者的喜好，　从而向用户产生推荐的算法。　当前协同过滤算法大致可以分为基于用户的协同过滤算法和基于项目的协同过滤算法。　协同过滤为主要算法的推荐系统的应用领域日益广泛，电子商务是其应用的最主要和最成功的领域。但协同过滤算法仍具有很多不足之处，　最突出的不足分别是数据稀疏性问题，　冷启动问题和系统延伸性问题。　在已有的理论和实践研究基础上，　个人提出了协同过滤推荐值得深入研究的方向应包括多维数据的交叉利用，　从而提高协同过滤推荐的精准度。

关键字：　协同过滤推荐，基于用户，基于项目，数据稀疏，冷启动，系统延伸性，多维数据的交叉利用

正文

一、协同过滤推荐的基本定义

（一）协同过滤推荐的概念

协同过滤是如今推荐系统中最为成熟的的一个推荐算法系类，　简单来说是利用某兴趣相投、 拥有共同经验之群体的喜好来推荐使用者感兴趣的资讯，　个人透过合作的机制给予资讯相当程度的回应　(如评分) 并记录下来以达到过滤的目的进而帮助别人筛选资讯，　回应不一定局限于特别感兴趣的，　特别不感兴趣资讯的纪录也相当重要。

（二）协同过滤推荐的主要算法概述

当前协同过滤算法大致可以分为两类，一类是基于用户的协同过滤算法，一类是基于项目的协同过滤算法。

基于用户的协同过滤推荐根据相似用户群的观点来产生对目标用户的推荐。基本思想是如果某些用户对部分项目的评分趋于一致或是很接近，　可以认为他们对其它项目的评分差异就比较小，　进一步，可以使用这些相似用户的项目评分值对目标用户的未评分项目进行估计。　基于用户的协同过滤使用数理统计的方法来寻找与目标用户有相似兴趣偏好的最近邻居用户集合，　再以最近邻居用户对特定项目的评分为基础使用一定的数学方法来预测目标用户对该特定项目的评分，　而预测评分最高的前　N个商品可以看作是用户最有可能感兴趣　top-N 商品返回给目标用户（这就是所谓的　top-N　推荐）。 基于用户的协同过滤推荐算法的核心思想是利用数理统计的方法为目标用户寻找他的最近邻居用户集，　再以最近邻居用户对特定项目的评分为基础使用一定的数学方法来预测目标用户对该特定项目的评分，最终产生最后的推荐结果。　通过最近邻居用户对目标用户未评分项目的评分值进行加权平均来逼近，　这是该算法思想的关键。　基于用户的协同过滤推荐算法的主要工作有：　用户之间相似性的衡量、　最近邻居集的查找和评分预测值的计算。

和基于用户的协同过滤相比，　基于项目的协同过滤推荐算法的思想出发点是完全相反的，　但是计算方法一致。　基于项目的协同过滤推荐算法是根据用户对与

目标项目相似的项目的评分来预测该用户对目标项目的评分。　　基于项目的协同过滤是基于如下两个假设：（1）如果大部分用户对两个项目的评分相似，　　则这两个项目相似（ 2）用户对相似项目的评分也相似。同基于用户的协同过滤推荐算法相似，但是基于项目的协同过滤推荐算法的出发点不同，　　其思想是通过用户对目标项目最近邻居项目的评分来对目标项目进行预测评分进而产生最后的推荐结果，用户对目标项目的预测评分的计算方法同基于用户的协同过滤中使用的数学方法是一样的，就是通过用户对目标项目最近邻居项目的评分的加权平均值来逼近对目标项目的预测评分。

二、　协同过滤推荐的主要应用方向

（一）协同过滤推荐的主要应用方向简述

目前，协同过滤为主要算法的推荐系统的应用领域越来与广泛，包括电子商务，音乐，电影，图片，图书，图片，新闻等等。典型的代表如　　Amazon，淘宝，豆瓣，　Facebook，　Genius 等等。

其中，电子商务是推荐系统应用的最主要的领域，也是目前最成功的领域之一。几乎所有的大型电子商务系统，如　　Amazon，eBay，淘宝等，都不同程度地是有了各种形式的推荐。 就以淘宝为例，淘宝网在不断推出针对不同用户的商品、资讯等资源的推荐活动。 每个用户都有一个能够唯一识别身份的标识，　　通过记录每个人的购买历史信息对用户的行为、　兴趣偏好进行建模，　针对这些数据模型通过协同过滤等推荐技术为用户提供他们有潜在兴趣的商品。

三、协同过滤推荐算法的不足之处

（一）算法的三大基本问题概述

一个算法毕竟不是完美，　对于协同过滤算法来说，　一直存在数据稀疏性问题，　冷启动问题和系统延伸性问题这三大基本问题。

数据稀疏性问题是指由于商品之多，　所以系统的推荐一般只是对个别的几个商品而言，所以大部分商品都很难被推荐上，所形成的矩阵将会很稀疏。举个例子，淘宝上号称有数亿商品，　平均而言一个用户只浏览数百件，　所以矩阵的稀疏度会在百万分之一或以下的量级。　这使得协同过滤算法效果都不好。　但这个问题本质上是无法完全克服的，目前，为了解决这个问题，研究表明找到了很多办法，譬如可以通过扩散的算法，　从原来的一阶关联 （两个用户有多少相似打分或者共同购买的商品） 到二阶甚至更高阶的关联 （假设关联性或者说相似性本身是可以传播的），也可以添加一些缺省的打分，　从而提高相似性的分辨率。 数据规模越大，一般而言越稀疏。

冷启动问题又由新使用者问题和新项目者问题构成。　　新使用者问题是指对于刚开始使用该系统的用户，　于过往的历史记录较少，　所以比较难以判断该用户的兴趣爱好，系统开始时推荐品质较差。　新项目问题是指对于刚刚登入系统的物品，　　由于购买或阅读该物品的用户较少，　　也很难判断刚物品主要受哪类用户所欢迎，　　所以这种情况下，新的商品的被推荐度将大大减小。目前，也有一些解决的方法。一种办法是利用文本信息进行辅助推荐，　或者通过注册以及询问得知一些用户的属性信息，譬如年龄、居住城市、受教育程度、性别、职业等等。最近标签系统的广泛应用提供了解决冷启动问题的可能方案，　　因为标签既可以看作是商品内容

的萃取，同时也反映了用户的个性化喜好。　　当然，利用标签也只能提高有少量行为的用户的推荐准确性，　对于纯粹的冷启动用户，　是没有帮助的，　因为这些人还没有打过任何标签。　最近的研究显示，　新用户更容易选择特别流行的商品，　　说明使用热销榜也能获得不错的结果。　冷启动问题还可以通过多维数据的交叉推荐部分解决，其精确度和多样性又远胜于热销榜。

　　系统延伸性问题是指由于推荐系统的飞速发展，　　系统中的项目数量也在飞速上升。就以电子商务系统来看，　电子商务网站的个性化推荐系统的计算负担在不断的加重。从协同过滤推荐过程来看，算法在线计算复杂度为　　O(m,n)，其中用户相似性度量及最近邻搜寻是最耗时的算法环节。　　同时，从电子商务推荐系统结构来看，全部推荐计算都在服务器端完成，　使得服务器面临庞大的计算量。　因此，如何有效提高协同过滤可扩展性是必须予以研究和解决的重要问题。　　所以，算法的扩展性问题是制约个性化推荐系统发展的一个重要因素。　　提高协同过滤可扩展性的最简单方法是使用功能更强大的站点服务器和增加服务器数量。　　但是这一方面需要网站投入更多成本，　但并不能降低每个推荐的响应时间，　而响应时间对于推荐扩展性非常重要。针对这一问题，目前的研究成果主要分为聚类、　概率方法、数据集缩减等改善算法。

四、个人对协同过滤推荐值得深入研究的方向的思考

（一）多维数据的交叉利用的简介

目前网络科学研究一个广受关注的概念是具有相互作用的网络的结构和动力学。网络与网络之间的相互作用大体可以分成三类：　一类是依存关系，　譬如电力网络和 Internet　，如果发生了大规模停电事故，当地的自主系统和路由器也会受到影响，导致网络局部中断；第二类是合作关系，譬如人的一次出行，可以看作航空网络、铁路网络和公路网络的一次合作；　第三类是交叠关系，　主要针对社会网络。几乎每一个人，　都参与了不止一个大型的社会网络中，　　譬如你可能既有新浪微博的帐号，　又是人人网的注册用户，　还是用手机，　那么你已经同时在三个巨大的社会网络中了。与此同时，你可能还经常在淘宝、京东、麦包包、　　1 号店、库巴网,，这些地方进行网购，那么你也是一张巨大的用户　　- 商品二部分图中的一员。

　　想象如果能够把这些网络数据整合起来，　特别是知道每个节点身份的对应关系(不需要知道你真实身份，　只需要知道不同网络中存在的一些节点是同一个人)，其中有特别巨大的社会经济价值。　交叠社会关系中的数据挖掘，　或称多维数据挖掘，是真正有望解决系统内部冷启动问题的终极法宝——只要用户在系统外部的其他系统有过活动。　单纯从个性化商品推荐来讲，　可以利用用户在其他电商的浏览购买历史为提高在目标电商推荐的精确度——当然，每一个电商既是付出者，也是获利者，　总体而言，　大家能够通过提高用户体验和点击深度实现共赢。　　与此同时，可以利用微博和其他社会网络的活动提高商品推荐的精度，　　还可以反过来利用商品浏览历史提高微博关注对象推荐的精度。　　研究分析了百分点科技服务客户的真实数据，发现有相当比例的用户都具有交叉购物的习惯（在多个独立　　B2C 电商有浏览和购买行为）。

（二）多维数据的交叉利用的算法概述

　　这种跨邻域的推荐，　目前最有望的解决方法是机器学习中的“ 迁移学习 ”算

法。迁移学习的基本思想是应用相关领域的数据，　将相关领域的有用知识“迁移”到目标领域中，用以解决在目标领域的学习任务。　而这正好与上文提出的问题相契合。“迁移学习”可以主动利用用户在不同社交网络和电商平台，将用户所有的信息进行汇总，从而了解用户的喜好，提高对用户定位的精度和推荐的精度。然而迁移学习是利用少量的目标领域数据来判断相关性，　容易出现过度拟合，使泛化误差较大。为了避免以上问题就借鉴半监督思想，　主要通过引入更多的目标领域数据，来解决以上问题。半监督学习是有监督学习与无监督学习结合的一种学习方法，是近年来模式识别和机器学习领域研究的重点问题。　半监督的学习结合给出的少量的标记数据 {(x(L), y)}　，和大量的未标记数据 {x(u)} ，用标记数据的类别信息和未标记数据的分布信息，　两者结合来学习，目的是标记原来未标记的数据。半监督学习有利于减少标记代价，提高学习机器性能。

(三)用户行为模式的挖掘和利用

深入挖掘用户的行为模式有望提高推荐的效果或在更复杂的场景下进行推荐。譬如说，新用户和老用户具有很不一样的选择模式：一般而言，新用户倾向于选择热门的商品，　而老用户对于小众商品关注更多，　新用户所选择的商品相似度更高，老用户所选择的商品多样性较高。　有些混合算法可以通过一个单参数调节推荐结果的多样性和热门程度，　在这种情况下就可以考虑为给不同用户赋予不同参数（从算法结果的个性化到算法本身的个性化）　，甚至允许用户自己移动一个滑钮调节这个参数——当用户想看热门的时候，　算法提供热门推荐；当用户想找点很酷的产品时，　算法也可以提供冷门推荐。　用户行为的时空统计特性也可以用于提高推荐或者设计针对特定场景的应用。　用户的选择可能同时蕴含了长期的兴趣和短期的兴趣，通过将这两种效应分离出来，可以明显提高推荐的精确度。事实上，简单假设用户兴趣随时间按照指数递减，也能够得到改进的推荐效果。利用手机上网现在已经越来越普及，　与此同时，嵌入 GPS的手机越来越多，因此，基于位置的服务成为一个受到学术界和业界广泛关注的问题。　基于位置信息的推荐可能会成为个性化推荐的一个研究热点和重要的应用场景，　而这个问题的解决需要能够对用户的移动模式有深入理解　（包括预测用户的移动轨迹和判断用户在当前位置是否有可能进行餐饮购物活动等）　，同时还要有定量的办法去定义用户之间以及地点之间的相似性。　另外，不同用户打分的模式也很不一样，　用户针对不同商品的行为模式也不一样　（想象你在网上下载一首歌和团购房子时的区别）　，这些都可以用来提高推荐的效果。

(三)多样性与准确性的两难困境

如果要给用户推荐他喜欢的商品，最“保险”的方式就是给他特别流行或者得分特别高的商品，　因为这些商品有更大的可能性被喜欢　（至少贝叶斯会这么想），往坏了说，也很难特别被讨厌。但是，这样的推荐产生的用户体验并不一定好，因为用户很可能已经知道这些热销流行的产品，所以得到的信息量很少，并且用户不会认同这是一种“个性化的”推荐。事实上，　Mcnee等人已经警告大家，盲目崇拜精确性指标可能会伤害推荐系统——因为这样可能会导致用户得到一些信息量为 0 的“精准推荐”并且视野变得越来越狭窄。　让用户视野变得狭窄也是协同过滤算法存在的一个比较主要的缺陷。　与此同时，应用个性化推荐技术的商家，也希望推荐中有更多的品类出现，　从而激发用户新的购物需求。　遗憾的是，推荐多样的商品和新颖的商品与推荐的精确性之间存在矛盾，　因为前者风险

很大——一个没什么人看过或者打分较低的东西推荐出手，很可能被用户憎恶，从而效果更差。 很多时候， 这是一个两难的问题， 只能通过牺牲多样性来提高精确性，或者牺牲精确性来提高多样性。 一种可行之策是直接对推荐列表进行处理，从而提升其多样性。 目前百分点推荐引擎所使用的方法也是类似的。 这种方法固然在应用上是有效的， 但是没有任何理论的基础和优美性可言， 只能算一种野蛮而实用的招数。 我们发现， 通过精巧混合精确性高和多样性好的两种算法， 可以同时提高算法的多样性和精确性， 不需要牺牲任何一方。 遗憾的是， 我们还没有办法就这个结果提供清晰的解读和深刻的见解。 多样性和精确性之间错综复杂的关系和隐匿其后的竞争，到目前为止还是一个很棘手的难题。

五. 总结

协同过滤作为如今推荐系统中最为成熟的的一个推荐算法系类， 应用场景十分广泛。 本文简要地分析了协同过滤算法的基本定义， 主要算法组成及其不足和算法的深入研究点。 这是本人对协同过滤推荐算法的基本认识， 希望日后能够更加深入地学习有关个性化推荐的知识，逐步深入，最终可以形成个人的理念，从而对个性化推荐的某些方面提出一些的改善建议。

参考文献：

[1] 孟祥武，胡勋，王立才，张玉洁 . 移动推荐系统及其应用 [J]. 软件学报，2013, 24（1）：91-108

[2] 赵亮，胡乃静，张守志 . 个性化推荐算法设计 [J]. 计算机研究与发展， 2002(8) ：986-991.

[3] 蒋国瑞， 青海， 黄梯云 . 一种柔性的电子商务推荐系统 [J]. 计算机应用研究 .2009， 3， 930-933.