



详解个性化推荐五大最常用算法

2017年07月08日 12:30:09 量子位

0

允中 若朴 编译自 StatsBots

量子位 出品 | 公众号 QbitAI



推荐系统，是当今互联网背后的无名英雄。



量子位

关注前沿科技资讯，追踪人工智能动态

热文排行

日榜

周榜

月榜

- 1 历史再度重演！阿里和腾讯这一次争抢的..
- 2 最新全国人均住房面积公布，你家达到这..
- 3 BAT华为员工收入比拼，不啃老，三年能...

我们在某宝首页看见的商品，某条上读到的新闻，甚至在各种地方看见的广告，都有赖于它。

昨天，一个名为StatsBots的博客详解了构建推荐系统的五种方法。

量子位编译如下：

现在，许多公司都在用大数据来向用户进行相关推荐，驱动收入增长。推荐算法有很多种，数据科学家需要根据业务的限制和要求选择最好的算法。

为了简化这个任务，Statsbot团队写了一份现有的主要推荐系统算法的概述。

协同过滤

协同过滤(Collaborative filtering, CF)及其变体是最常用的推荐算法之一。即使数据科学的新手也可以用它来构建自己的个人电影推荐系统，起码可以写在简历上。

我们想给用户推荐东西，最合乎逻辑方法是找到具有相似兴趣的人，分析他们的行为，并向用户推荐相同的项目。另一种方法是看看用于以前买的商品，然后给他们推荐相似的。

CF有两种基本方法：基于用户的协同过滤和基于项目的协同过滤。

无论哪种方法，推荐引擎有两个步骤：

了解数据库中有多少用户/项目与给定的用户/项目相似。


- 4 来自日本的视角：中国科技实力究竟怎样
- 5 寒门“代沟”，比阶层固化更令人揪心
- 6 炒房客竟玩房价腰斩游戏！揭秘背后鲜为..
- 7 GDP增速达7% 印度会是下一个中国吗？
- 8 看到这些城市房价和收入比，买房真的不..
- 9 楼市降价潮袭来！房价腰斩或许一切只是..
- 10 神奇少女改卖保健品：公号被封，南京农..



考虑到与它类似的用户/项目的总权重，评估其他用户/项目，来预测你会给该产品用户的打分。

“最相似”在算法中是什么意思？

我们有每个用户的偏好向量(矩阵R的行)，和每个产品的用户评分向量(矩阵R的列)，如下图所示。

User / Item	Batman	Star Wars	Titanic
Bill	3	3	
Jane		2	4
Tom		5	

首先，我们只留下两个向量的值都已知的元素。

例如我们想比较Bill和Jane，已知比尔没有看泰坦尼克号，Jane没看过蝙蝠侠，于是，我们只能通过星战来衡量他们的相似度了。谁没看过星球大战呢是吧？

测量相似度的最流行方法是余弦相似性或用户/项目向量之间的相关性。最后一步，是根据相似度用加权算术平均值填充表中的空单元格。

矩阵分解

这是一个非常优雅的推荐算法，因为当涉及到矩阵分解时，我们通常不会太多地去思考哪些项目将停留在所得矩阵的列和行中。但是使用这个推荐引擎，我们清楚地看到， u 是第 i 个用户的兴趣向量， v 是第 j 个电影的参数向量。

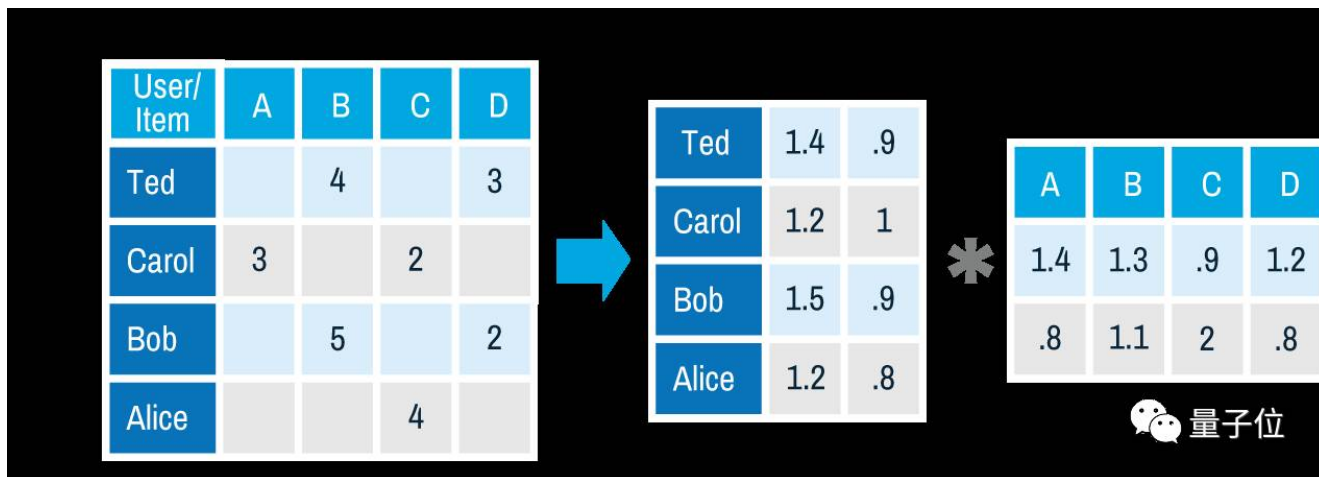
$$x_{ij} \approx \langle u_i, v_j \rangle$$

$$\sum_{i,j} (\langle u_i, v_j \rangle - x_{ij})^2 \rightarrow \min$$

量子位

所以我们可以用 u 和 v 的点积来估算 x (第 i 个用户对第 j 个电影的评分)。我们用已知的分数构建这些向量，并使用它们来预测未知的得分。

例如，在矩阵分解之后，Ted的向量是(1.4; .8)，电影A的向量是(1.4; .9)，现在，我们可以通过计算(1.4; .8)和(1.4; .9)的点积，来还原电影A-Ted的得分。结果，我们得到2.68分。



聚类

上面两种算法都极其简单，适用于小型系统。在这两种方法中，我们把推荐问题当做一个有监督机器学习任务来解决。

现在，该开始用无监督学习来解决问题了。

假设我们正在建立一个大型推荐系统，这时协同过滤和矩阵分解花费的时间更长了。第一个浮现在脑海里的解决之道，就是聚类。

业务开展之初，缺乏之前的用户数据，聚类将是最好的方法。

不过，聚类是一种比较弱的个性化推荐，因为这种方法的本质是识别用户组，并对这个组内的用户推荐相同的内容。

当我们有足够数据时，最好使用聚类作为第一步，来缩减协同过滤算法中相关邻居的选择范围。这个方法还能挺高复杂推荐系统的性能。

每个聚类都会根据其中用户的偏好，来分配一组典型的偏好。每个聚类中的用户，都会收到为这个聚类计算出的推荐内容。

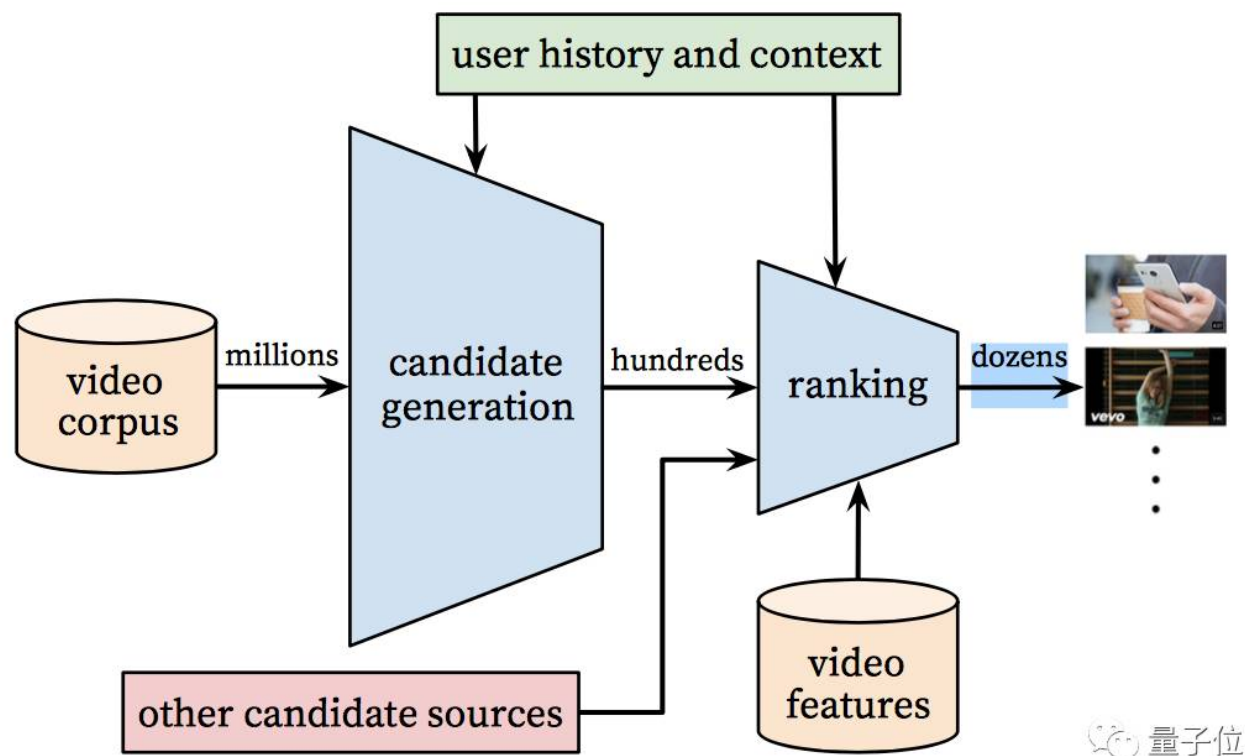
深度学习

在过去的十年中，神经网络已经取得了巨大的飞跃。如今，神经网络已经得以广泛应用，并逐渐取代传统的机器学习方法。

接下来，我要介绍一下YouTube如何使用深度学习方法来做个性化推荐。

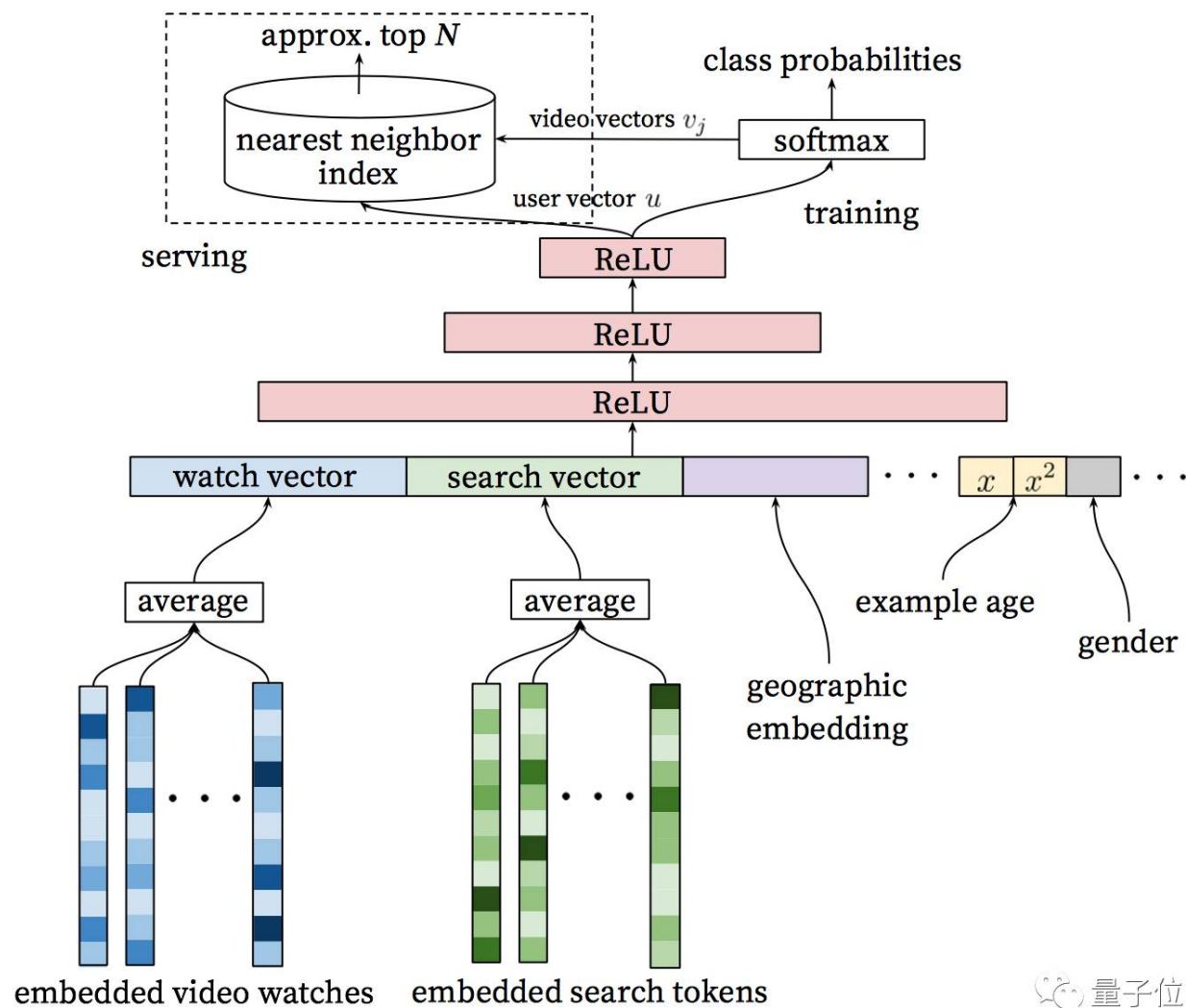
毫无疑问，由于体量庞大、动态库和各种观察不到的外部因素，为YouTube用户提供推荐内容是一项非常具有挑战性的任务。

根据《Deep Neural Networks for YouTube Recommendations》
(<https://static.googleusercontent.com/media/research.google.com/ru//pubs/archive/45530.pdf>)，YouTube的推荐系统算法由两个神经网络组成：一个用于候选生成，一个用于排序。如果你没时间仔细研究论文，可以看看我们下面给出的简短总结。



以用户的浏览历史为输入，候选生成网络可以显著减小可推荐的视频数量，从庞大的库中选出一组最相关的视频。这样生成的候选视频与用户的相关性最高，然后我们会对用户评分进行预测。

这个网络的目标，只是通过协同过滤提供更广泛的个性化。

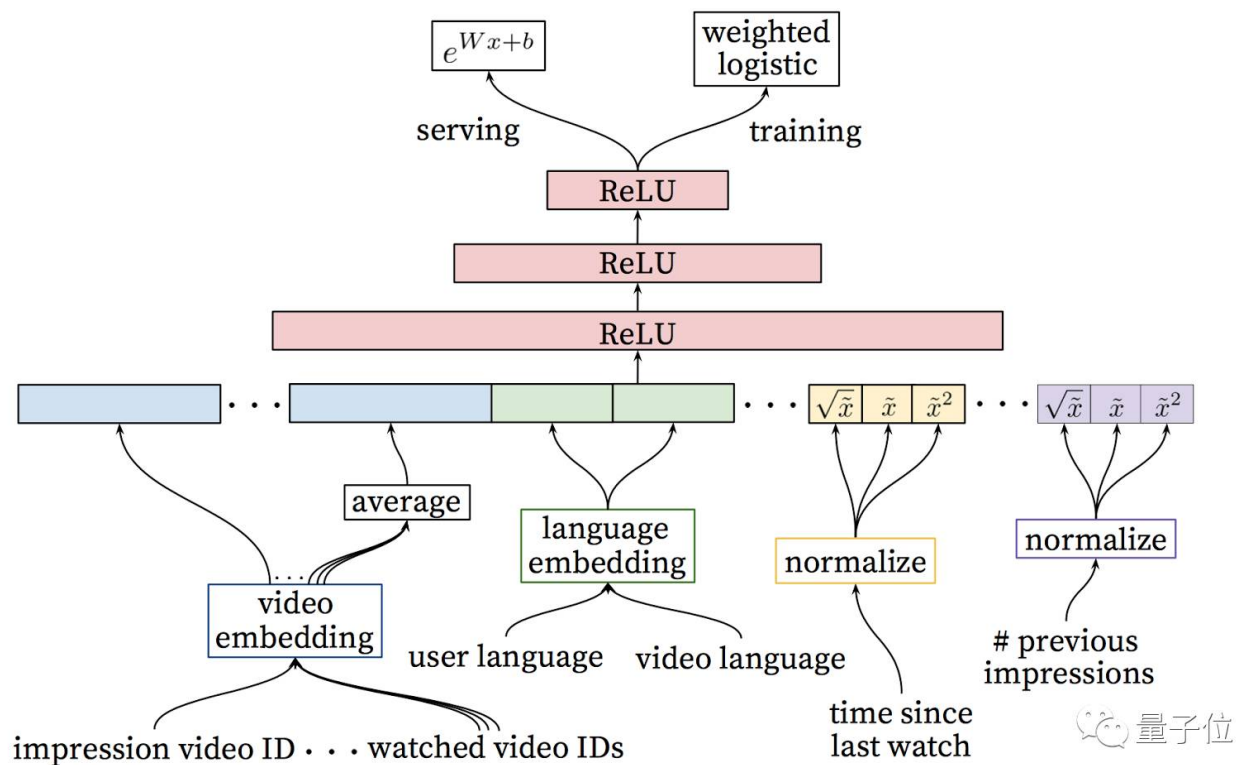


量子位

进行到这一步，我们得到一组规模更小但相关性更高的内容。我们的目标是仔细分析这些候选内容，以便做出最佳的选择。

这个任务由排序网络完成。

所谓排序就是根据视频描述数据和用户行为信息，使用设计好的目标函数为每个视频打分，得分最高的视频会呈献给用户。



通过这两步，我们可以从非常庞大的视频库中选择视频，并面向用户进行有针对性的推荐。这个方法还能让我们把其他来源的内容也容纳进来。

$$P(w_t = i|U, C) = \frac{e^{v_i u}}{\sum_{j \in V} e^{v_j u}}$$

量子位

推荐任务是一个极端的多类分类问题。这个预测问题的实质，是基于用户(U)和语境(C)，在给定的时间t精确地从库(V)中上百万的视频类(i)中，对特定的视频观看(Wt)情况进行分类。

建立推荐系统前该知道的要点

如果你有一个庞大的数据库，而且准备提供在线的推荐，最好把这个任务拆分成两个子问题：

选择Top N个候选；

排序。

如衡量推荐模型的质量？

除了标准质量指标之外，还有一些针对推荐问题的指标：比如说召回率与准确率(https://en.wikipedia.org/wiki/Information_retrieval#Precision_at_K)。还有一些其他的指标，见《软件工程中的推荐系统》一书第12章(<http://www.ict.swin.edu.au/personal/jgrundy/papers/rsse2014.pdf>)。

如果你正在使用分类算法解决推荐问题，应该考虑生成负例样本。如果用户购买了推荐的商品，你应该将其添加为正例样本，而其他列为负例样本。

要从在线得分和离线得分两个方面考察算法质量。一个只基于历史数据的训练模型，可能会导致低水平的推荐，因为算法没办法与时俱进。

推荐阅读

个性化推荐在产品里都能用在哪呢？

量子位曾报道过知乎、Quora、Airbnb是如何使用机器学习技术的，推荐系统是其中的重头戏：

详解：知乎如何使用机器学习，未来还有哪些想象空间

详解：估值18亿美元的新晋独角兽Quora，如何使用机器学习？

搞日租房的Airbnb，如何用机器学习对接上百万的房东和租客？

【完】

一则通知

量子位读者5群开放申请，对人工智能感兴趣的朋友，可以添加量子位小助手的微信qbitbot2，申请入群，一起研讨人工智能。

另外，量子位大咖云集的自动驾驶技术群，仅接纳研究自动驾驶相关领域的在校学生或一线工程师。申请方式：添加qbitbot2为好友，备注“自动驾驶”申请加入~

招聘

量子位正在招募编辑/记者等岗位，工作地点在北京中关村。相关细节，请在公众号对话界面，回复：“招聘”。

△ 扫码强行关注『量子位』

追踪人工智能领域最劲内容

■

0

作者历史文章

科学家正让AI自己做实验，想要机器摆脱人类的直觉



李杉 编译自 Science量子位 报道 | 公众号 QbitAI如果说这是未来的生物实验室，它似乎与现在的实验室没有多大差别。里面有身穿白大褂的科学家，还有许[详细]

2017年 07月08日 12:30

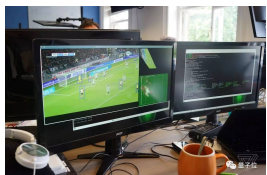
特斯拉市值一周蒸发120亿美元，通用汽车“躺赢”重回第一



问耕 允中 发自 凹非寺量子位 报道 | 公众号 QbitAI对伊隆·马斯克（Elon Musk）来说，这是感觉复杂的一周。这位硅谷钢铁侠旗下的SpaceX经历[详细]

2017年 07月07日 14:00

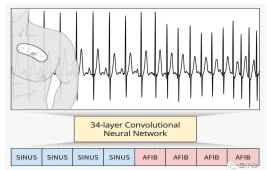
博彩公司正用人工智能来赌球：足球比赛简直是最容易预测的事



唐旭 编译自The Verge量子位出品 | 公众号 QbitAI本周一，2017年联合会杯决赛在俄罗斯圣彼得堡十字架体育场进行。最终，凭借施廷德尔在第20分钟[详细]

2017年 07月07日 14:00

吴恩达带斯坦福ML组发了新论文：深度学习攻克心律不齐难题



李林 编译整理量子位 报道 | 公众号 QbitAI吴恩达带领的斯坦福机器学习组 (Stanford ML Group)最近开发了一种深度学习新算法，能诊断14类[详细]

2017年 07月07日 14:00

机器学习中的“哲学”



作者：阿萨姆普华永道|数据科学家量子位 已授权编辑发布0. 前言我更喜欢把“思想”认为是一种“道”，而“模型”是一种“术”，也可类比为“外功”和“内功”。本文有[详细]

2017年 07月07日 14:00

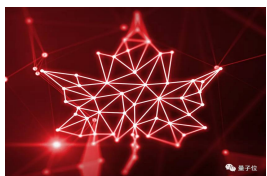
全面剖析无人车三大基本技术：计算、动力和电传线控



陈桦 编译自 Voyage官方博客量子位 报道 | 公众号 QbitAI打造一辆无人车，究竟需要哪些软件和硬件？无人车创业公司Voyage今天在官方博客上发文，[详细]

2017年 07月07日 14:00

DeepMind首次走出英国根据地，在埃德蒙顿创建实验室



安妮 编译整理量子位出品 | 公众号 QbitAI昨天，DeepMind在加拿大埃德蒙顿建立了首个国际AI实验室。这是DeepMind自2014年被谷歌收购以来[详细]

2017年 07月06日 14:00

创新工场深度学习训练营DeeCamp六大逗趣项目预告大放送



本文转自创新工场（chuangxin2009）DeeCamp创新工场深度学习训练营开营在即，究竟有哪些既逗趣又充满挑战性的项目呢？小编带你提前探秘一下~Chap[详细]

2017年 07月06日 14:00

阿里售价499的智能音箱背后，终极目标还是开放平台、生态系统



允中 发自 鼓楼量子位 报道 | 公众号 QbitAI果不其然，阿里首款智能音箱发布。就在百度首届AI开发者大会召开的同时，阿里巴巴人工智能实验室昨天下午也在北[详细]

2017年 07月06日 14:00



关于头条 | 如何入驻 | 发稿平台 | 奖励机制 | 版权声明
用户协议 | 帮助中心 © 1996-2015 SINA Corporation, All Rights Reserved