51 | 综合应用篇答疑和总结：如何进行个性化用户画像的设计？

黄申 2019-04-12

00:00 12:01

讲述：黄申　　大小：11.01M

你好，我是黄申。今天是综合应用篇的答疑和总结。

在这个模块中，我们讲述了不同数学思想在系统设计和实现中的综合运用。相对于前面几个模块，综合应用的内容更注重实践，也更加有趣。大家对这些内容也提出了很多值得思考的问题。今天，我会讲解其中一个问题，如何进行个性化用户画像的设计？。最后，我也会照例对整个应用篇进行一个总结。

个性化用户画像的设计

如今是个性化的时代，互联网和人工智能技术正在把这点推向极致。无论是主动搜索还是进行浏览，用户都希望看到针对自己的结果。

举个例子，A 品牌的奶瓶在全网是非常畅销的，可是对于一位 5 岁儿子的妈妈来说，儿子早已过了喝奶瓶的阶段，所以在她输入 A 品牌后，返回“奶瓶”肯定不合适。同时，如果她一直在购买 A 品牌的儿童洗衣液，那么返回 A 品牌的洗衣液就更合理，顾客体验也会更好，这就是**品类的个性化**。

从另一个场景来看，这位妈妈没有输入 A 品牌，而是输入了“儿童洗衣液”，如果是 A 品牌的洗衣液产品排在首页，而不是她所陌生的其他品牌，用户体验也会更好，这就是**品牌的个性化**。

在进行个性化设计之前，最关键的问题是，如何收集和运用顾客的行为数据。

第 48 节，我在讲解查询分类的时候，介绍了如何利用用户的搜索行为。而实践中，用户个人的

行为涉及面更为广泛，需要更多细致的分析。通常我们将相应的工程称为“用户画像”。为了让你更好地理解，这里我给出一个较为全面的设计概述。

如何通过数据生成用户标签？

开发用户画像，首先要解决的问题是：哪些用户数据可以收集，以及如何通过这些数据生成用户标签。

最基本的原始数据包括网站浏览、购物、位置、气候、设备等信息。除了这些原始的数据，我们还可以结合人工的运营，生成一些包含语义的用户标签。这里的用户标签，或者说属性标签，是一个具有语义的标签，用于描述一组用户的行为特征。例如，“美食达人”“数码玩家”“白领丽人”“理财专家”等。对于标签的定义，按照概率统计篇和线性代数篇所介绍的机器学习方法论，既可以考虑采用监督式的分类方法，也可以采用非监督式的聚类方法。

分类的好处在于，可以让人工运营向计算机系统输入更多的先验知识，也可以让标签的制定和归类更为精准。从操作的层面考虑，又可以细分为基于人工规则和基于标注数据。人工规则是指由运营人员指定分类的主要规则。

例如，运营人员指定最近 1 个月，至少购买过 2 次以上母婴产品，消费额在 500 元以上的为“辣妈” 标签。这里规则就相当于直接产生类似决策树的分类模型，它的优势在于具有很强的可读性，便于人们的理解和沟通。但是，如果用户的行为特征过于繁多，运营人员往往很难甄别出哪些具有代表性。这时如果仍然使用规则，那么就不容易确定规则的覆盖面或者是精准度。

另一种方法是使用标注数据，通过训练样本来构建分类器。例如，通过运营人员挑选一些有代表性的用户，对他们的特征进行人工标注，然后输入给系统。之后，让系统根据分类技术来学习，模型可以使用决策树、朴素贝叶斯 NB（Naive Bayes）或支持向量机 SVM（Support Vector Machine）等等。

不过，除了决策树的模型，其余模型产生的人群分组可能会缺乏可读性内容，很难向业务方解释其结果。一种缓解的办法是让系统根据数据挖掘中的特征选择技术，包括我们之前讲解的信息增益 IG（Information Gain）、开方检验 CHI 等，来确定这组人群应该有怎样的特征，并将其作为标签。

除了分类，我们也可以使用非监督式的聚类。这种方法中，运营人员参与最少，完全利用用户之间的相似度来确定，相似度同样可以基于各种用户的特征和向量空间模型来衡量。其问题也在于结果缺乏解释性，只能通过特征选择等技术来挑选具有代表性的标签。

如果我们比较一下分类和聚类的方法，会发现分类的技术比较适合业务需求明确、运营人员充足、针对少量高端顾客的管理，其精准性可以提升 VIP 顾客服务的品质。而聚类更适合大规模用户群体的管理，甚至是进行在线的 AB 测试，其对精准性要求不高，但是数据的规模比较大，对系统的数据处理能力有一定要求。

无论是哪种方法，只要我们能获取比较准确的用户标签，那么我们就可以给出用户的画像，刻画他们的主要行为特征。下面我们来看看基于用户画像，可以进行哪些个性化的服务。首先是在搜索中增加个性化因素，相比普通的搜索，个性化的搜索可以投用户之所好，增加搜索结果的点击率、商品的购买转化率等等。具体来说，我们可以在下面这几点下功夫：

第一点，个性化的排序，根据用户经常浏览的品类和属性，对搜索结果中的项目进行个性化的排序，开头提到的 5 岁儿子妈妈的案例体现了这点的核心思想。

第二点，个性化的搜索词推荐。例如，一位体育迷搜索“足球”的时候，我们可以给出“足球新

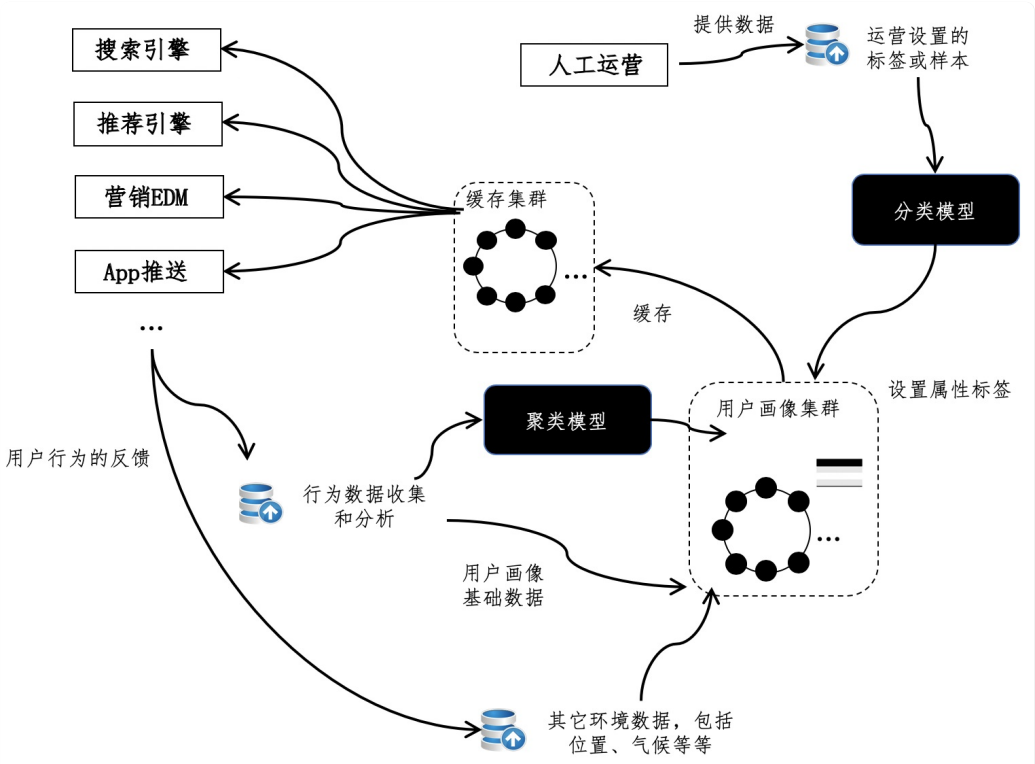
闻”“冠军杯”等相关搜索。而在一位彩票用户搜索“足球“的时候，我们可以给出”足球彩票“等相关搜索。

第三点，个性化的搜索下拉提示。例如，经常购买儿童洗衣液的用户，输入儿童用品的品牌后，在搜索下拉框中优先提示该品牌的儿童洗衣液。

除了搜索，个性化还可以运用在推荐系统、电子邮件营销 EDM（Email Direct Marketing）、移动 App 的推送等等。对于推荐系统来说，在用户画像完善的前提下，我们能更准确地找到相似的用户和物品，从而进行效果更好的基于用户或基于物品的协同过滤。相对于传统的线下营销，电子邮件营销不再受限于印刷和人力成本，完全可以做到因人而异的精准化定向投放。

比如，系统根据品类、品牌、节日或时令，分为不同的主题进行推送。运营人员甚至只用制定模板和规则，然后让系统根据用户画像的特征，自动的填充模板并最终生成电子邮件的内容。另外，随着移动端逐渐占据互联网市场的主导地位，掌上设备的 App 推送变成了另一个重要的营销渠道。从技术层面上看，它可以采用和电子邮件营销类似的解决方案。不过，内容的运营要考虑到移动设备屏幕尺寸和交互方式的特性，并进行有针对性的优化。

有了上述这些设计理念和模块，我们需要一个整体的框架来整合它们。我在这里画了一张框架图，供你参考。



这种架构包括行为数据的收集和分析、聚类、分类、构建画像、缓存等几个主要模块。随着数据规模的不断扩大，我们可以选择一些分布式系统来存储用户画像数据，并使用缓存系统来升数据查询的效率，为前端的搜索、推荐、EDM 和 App 推送等应用提供服务。当然，我们还可以利用行为数据的跟踪，进一步分析这套画像系统的质量和效果，形成一个螺旋式上升的优化闭环。

综合来看，用户画像也许概念上并不复杂，可是一旦落实到技术实施，我们需要综合很多不同领域的知识。从用户标签的角度来说，可能涉及的领域包括监督式和非监督式的机器学习算法，以及相关的特征选择。从系统集成的角度来说，可能涉及的领域包括分布式、缓存、信息检索和推荐系统。这些内容我们在之前的各个模块都有介绍，今天我通过用户画像的设计进行了知识的串联。当然，我这里讲解的方案也只是一种参考，你可以结合自身的需求来进一步的设计和实现。相信经过一定量的项目实践和经验积累，你对这些内容的综合性运用会更加得心应手。

综合应用篇总结

在综合应用篇之前，我们分别从基础模块、概率统计模块和线性代数模块出发，详细阐述了不同编程技术背后的数学的知识。在综合应用这个模块，我们又从几个非常实用的案例出发，讲解了如何结合不同的编程技术，设计并架构大型的系统，最终为商业需求提供解决方案。

如今的数据系统越来越庞大，系统设计时常常会用到缓存系统来提升记录查找的效率。对缓存系统的强烈需求也催生了很多开源的项目，例如 Memcached 和 Redis 等等，这些系统都已经相当成熟。而在这个模块，我们同时使用了哈希函数和队列，实现了一个最简单的缓存系统。哈希函数确保了查找的高效率，而队列则实现了 LRU 的淘汰策略。通过这两点，你就能理解缓存设计的基本原理和方法。

和缓存类似，搜索引擎的倒排索引也使用了哈希表结构来提高查询效率。当然，倒排索引的功能不仅限于数据对象的快速定位。它本身还能存放很多额外的信息，包括词频 tf、tfidf、关键词出现的位置等等。在这个模块中，我展示了如何利用这些信息，实现更为复杂的相关性模型，例如向量空间模型、概率语言模型等等。另外，倒排索引可以帮助我们过滤掉完全无关的数据，大大降低这些模型的计算量。

除了基本的及时性和相关性，搜索引擎还应该按照不同应用的需求进行优化。例如，电商平台的搜索，就和通用型的搜索不一样，对于电商搜索来说，用户更加关注的是商品的品类。我讲解了如何根据商品目录和用户行为反馈，构建查询的分类器。这样，当用户进行搜索的时候，系统首先对用户输入的关键词进行分类，弄清楚用户最感兴趣的品类是哪些，然后再优化商品的排序，最终增加商品搜索结果的相关性。

和搜索引擎同样重要的是推荐引擎。有的时候，用户自己不会输入想要查询的关键词，而是喜好不断地浏览网页。这个时候推荐技术起到了很关键的作用，它可以主动地为用户提供他们可能感兴趣的内容。在这个领域，协同过滤是非常经典的算法。我通过代码的实践，给你讲解了如何通过矩阵操作实现基于用户和基于物品的过滤。除此之外，我们还探讨了如何使用 SVD，对用户和物品之间的关系进行分解，帮助我们找到隐藏在用户和物品之间的潜在因素，比如电影的主题。

无论是设计搜索还是推荐系统，我们都可以加入个性化的元素，而这点往往是提升业务的关键。今天，我讲解了用户画像的原理、用户标签的设计和实现、以及如何使用用户画像来给搜索和推荐系统加入个性化。而这个整体方案涉及的技术面是相当广的，你可以结合之前的各期专栏，对每一个环节进行消化和理解。

思考题

对各种知识的综合应用对个人能力要求很高，却也是最重要的。我想听你说说，在平时的开发项目中，你有没有结合使用本专栏所讲的不同知识点的经历？能不能和我们说说你在这方面的心得体会？

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。

Ctrl + Enter 发表　　0/2000字　　提交留言

精选留言(1)

Dont we all

你好，最近在做遗传算法，请问可以介绍一下吗

2019-04-12