从用户行为去理解内容-item2vec及其应用

原创 samuelqiu 腾讯知文 2018-01-30

相关性是对称的

在内容推荐系统里，一个常用的方法是通过理解内容（挖掘内容属性）去挖掘用户的兴趣点来构建推荐模型。从大多数业务的效果来看，这样的模型是有效的，也就是说用户行为与内容是相关的。不过有一点常被忽略的是：相关性是对称的！这意味着如果可以从内容属性去理解用户行为，预测用户行为，那么也可以通过理解用户行为去理解内容，预测内容属性。

利用行为数据生成内容向量

推荐系统里我们一直有基于用户行为去理解内容，典型的例子是基于用户行为构造内容特征，例如内容的点击率、内容的性别倾向，内容的年龄倾向等。这样的理解是浅层的，仅仅是一些简单的统计。我们其实有更好的办法可以构建内容特征，它的第一步是利用用户行为将内容转化为向量，下面会以应用宝业务为例讲解利用用户行为将app转化为向量的思路。

从直觉上来看，用户下载app的先后关系是相关的，以图1的行为数据为例，一个用户之前下载过街头篮球，那么他接下来会下载体育类app的概率会比他接下来下载时尚类app的概率更大。也就是说 P(腾讯体育|街头篮球)>P(唯品会|街头篮球)

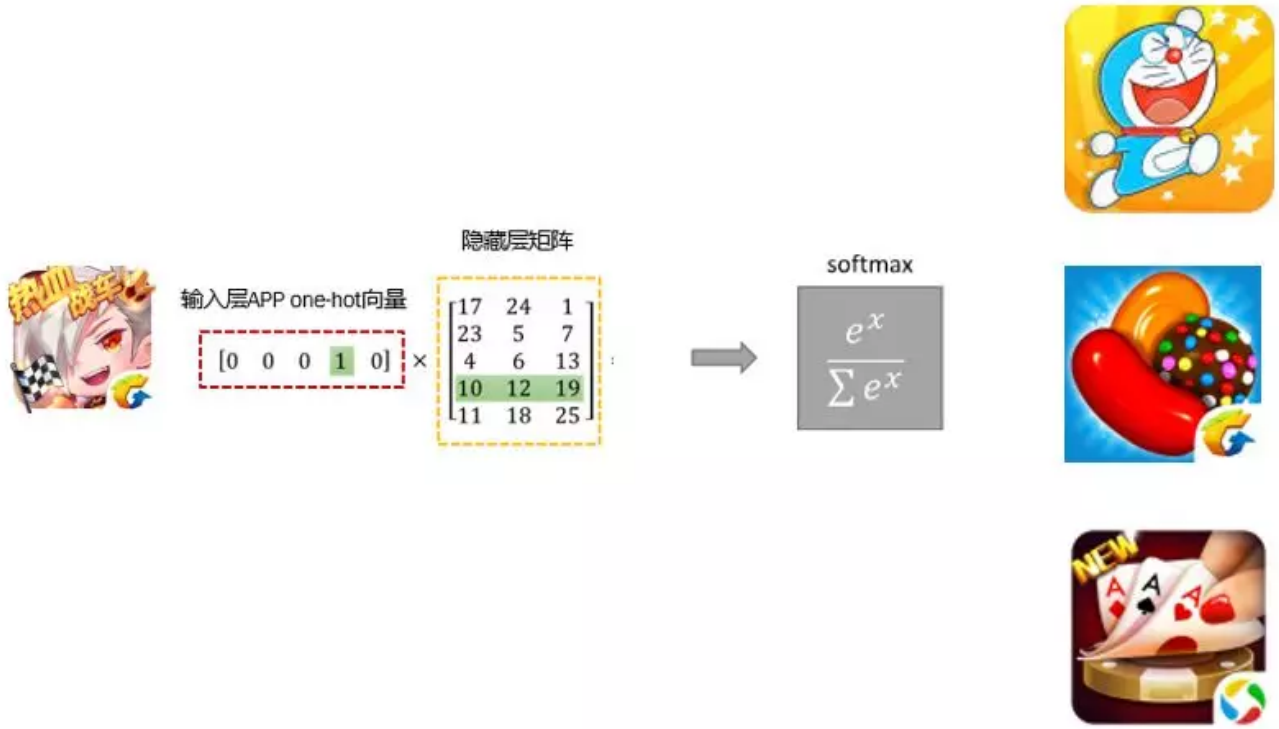
这里我们可以把用户已经下载的app看作是一个N维向量，基于app向量来计算条件概率。 假设用户已经下载的app的向量为 $\mathbf{u}_n$，用户接下来会下载的k个app的向量为 $\{\mathbf{v}_c|c=1,2,\ldots,k\}$（需要注意的是这里把用户已经下载的app向量和以后会下载的app向量分别置于不同的向量空间，所以分别用$\mathbf{u}$和$\mathbf{v}$表示），因此我们可以用softmax函数和app向量来表示条件概率：

$$p(c|n;\mathbf{v}_c,\mathbf{u}_n)=\frac{e^{\mathbf{v}_c^T\mathbf{u}_n}}{\sum_{i=1}^{l}e^{\mathbf{v}_i^T\mathbf{u}_n}}$$

我们的目标是希望模型能尽根据用户的下载行为记录预测用户接下来最可能会下载哪些app，因此我们的优化目标是求得最优的参数$\mathbf{v}_c$, $\mathbf{v}_n$使得上面的条件概率最大。可以得到下面的优化问题：

$$argmax_{\mathbf{v}_c,\mathbf{u}_n}\sum_{(n,c)\in\mathbf{D}}\log p(c|n)=\sum_{(n,c)\in\mathbf{D}}(\log e^{\mathbf{v}_c^T\mathbf{u}_n}-\log\sum_{i=1}^{l}e^{\mathbf{v}_i^T\mathbf{u}_n})$$

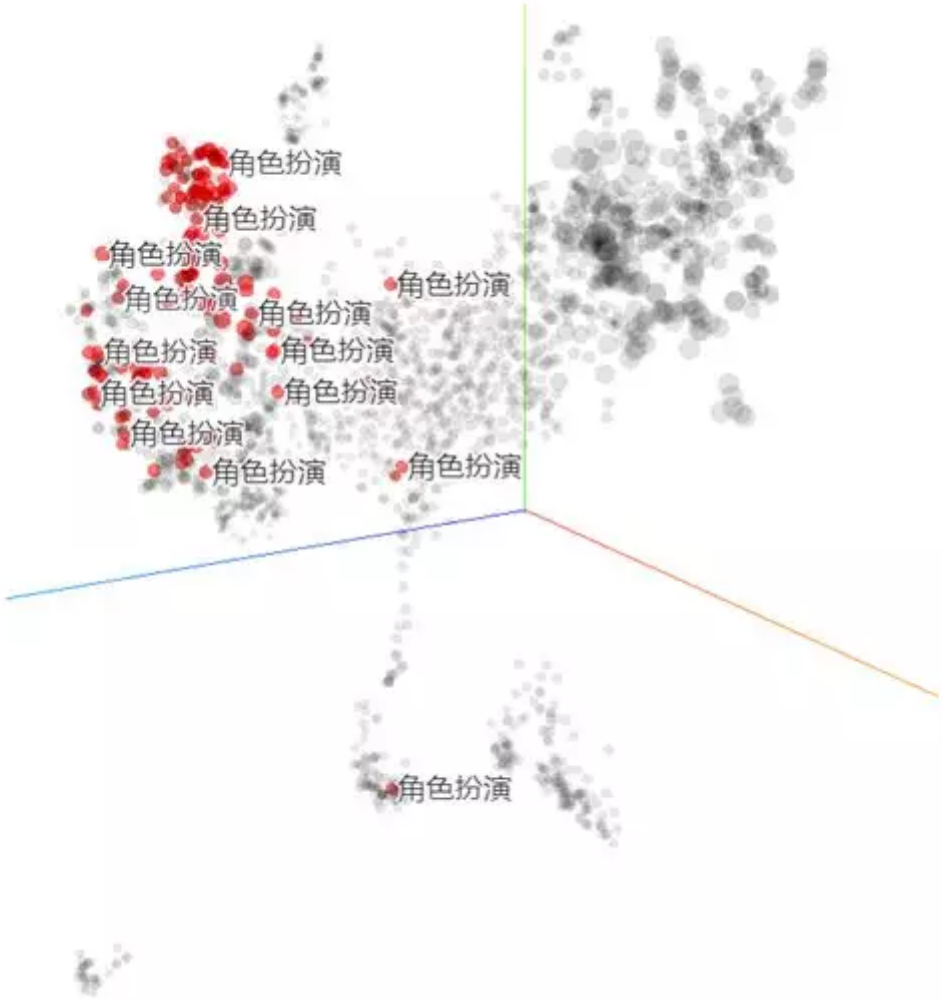
其中$\mathbf{D}$是用户行为序列集。如果公式看着头晕的话可以直接看下面图，图中隐藏层矩阵的每一行对应一个app向量。

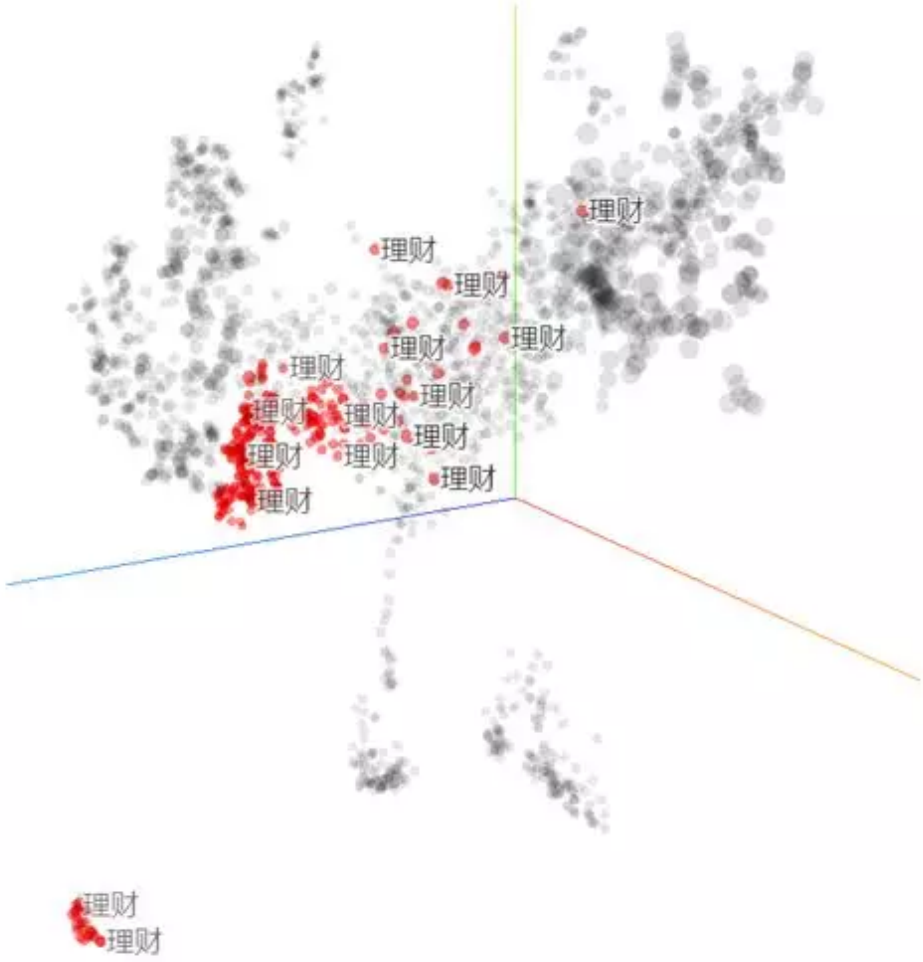


到这里我们已经大致介绍了利用用户行为将内容转化为向量的方法，这里将这种技术称作item2vec。以应用宝为例，它的item是app，它的实际应用也可以称作app2vec。

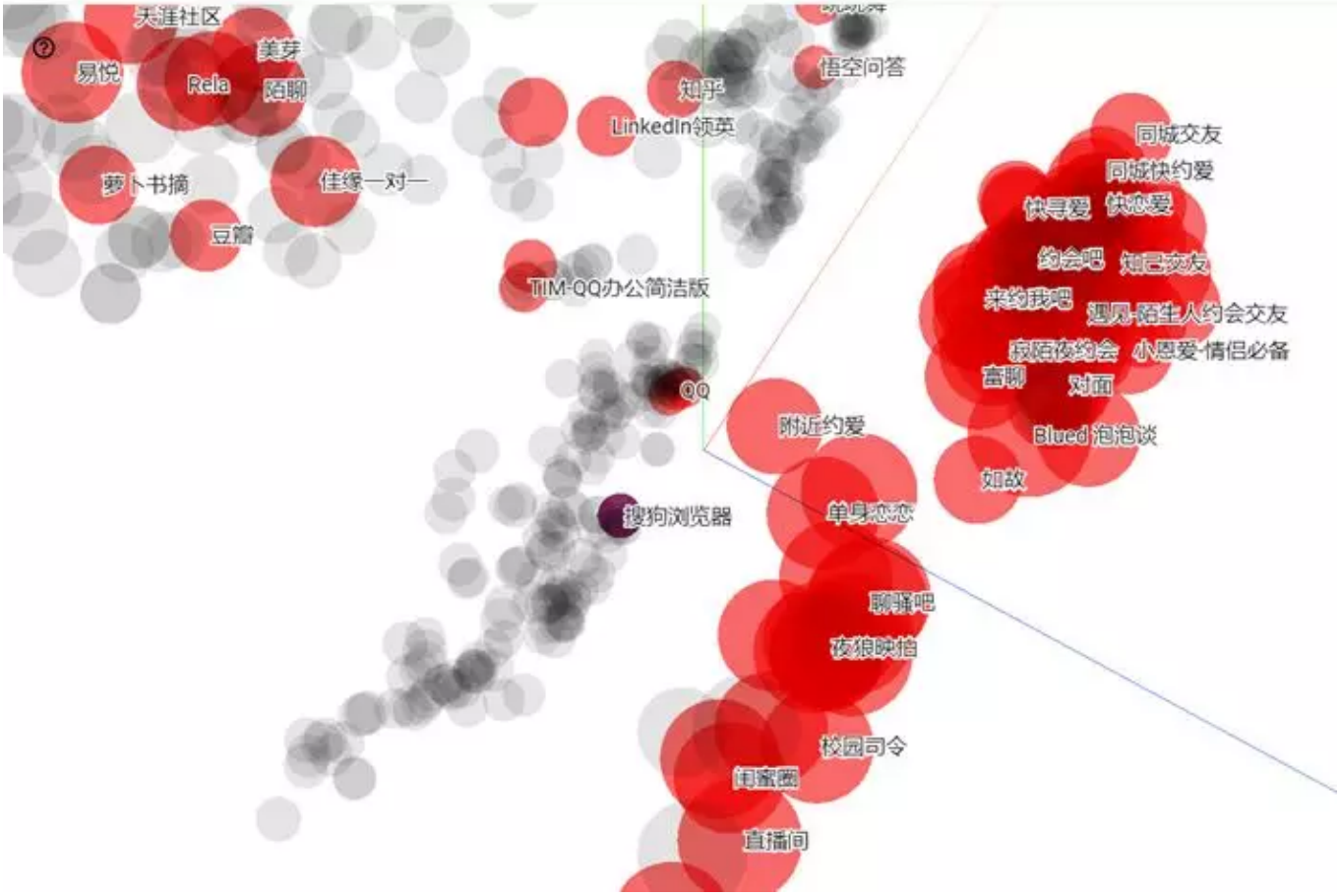
内容向量聚类

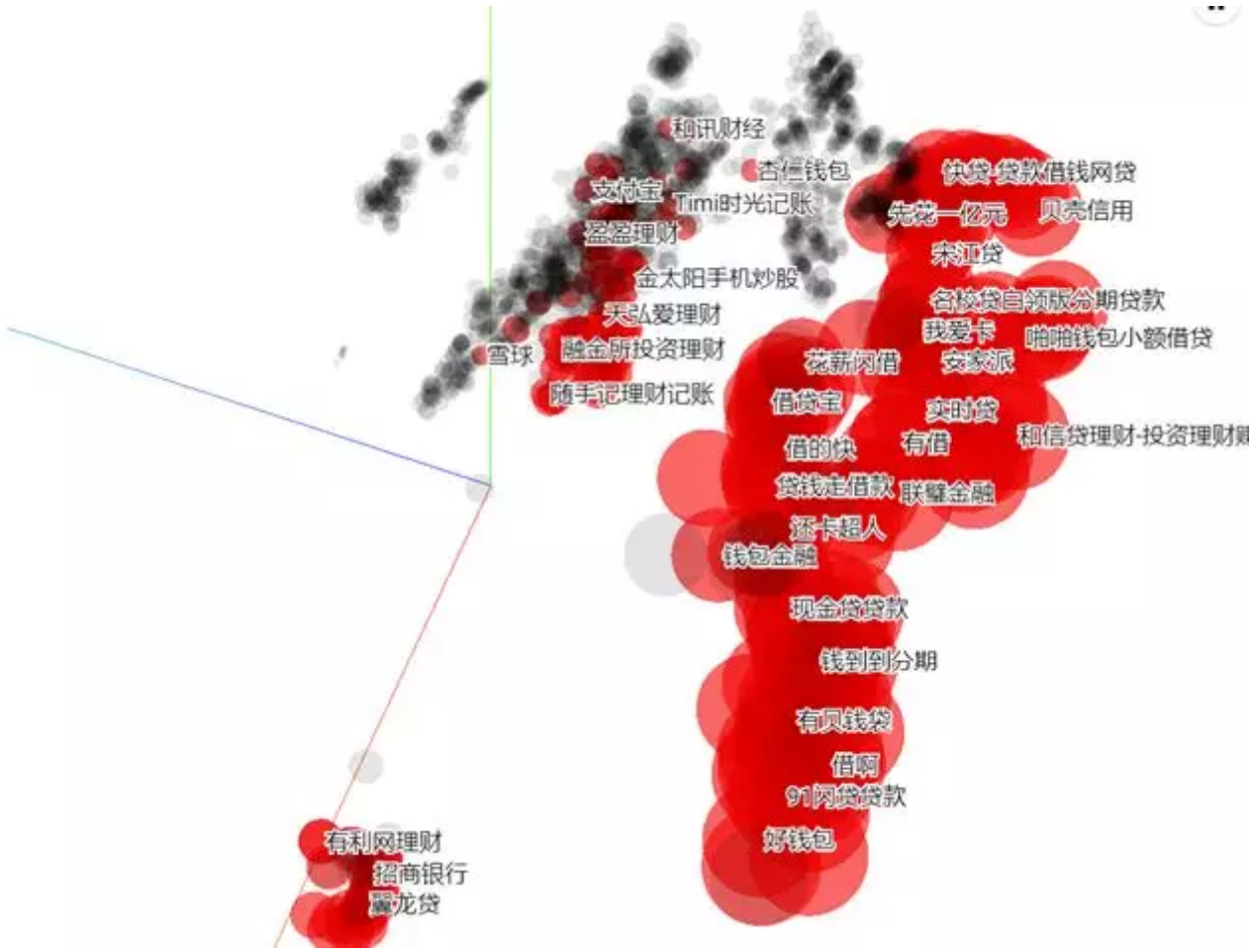
基于应用宝已有的类别体系观察，可以明显区分开角色扮演类游戏app和理财app


角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演
角色扮演



也可以发现一些没有加入类别体系的特殊app群体


天涯社区
美芽
易悦
Rela
陌聊
萝卜书摘
佳缘一对一
豆瓣
悟空问答
知乎
LinkedIn领英
TIM-QQ办公简洁版
QQ
搜狗浏览器
同城交友
同城快约爱
快寻爱
快恋爱
约会吧
知己交友
来约我吧
遇见-陌生人约会交友
寂陌夜约会
小恩爱-情侣必备
畅聊
对面
附近约爱
Blued 泡泡谈
如故
单身恋恋
聊骚吧
夜狼映猎
校园司令
闺蜜圈
直播间


和讯财经
杏仁钱包
支付宝
Timi时光记账
盈盈理财
金太阳手机炒股
天弘爱理财
雪球
融金所投资理财
随手记理财记账
快贷-贷款借钱网贷
先花一亿元
贝壳信用
宋江贷
名校贷白领版分期贷款
我爱卡
啪啪钱包小额借贷
花薪闪借
安家派
借贷宝
实时贷
借的快
有借
和信贷理财-投资理财
贷钱走借款
联鑫金融
还卡超人
钱包金融
现金贷贷款
钱到到分期
有贝钱袋
借啊
91闪贷贷款
好钱包
有利网理财
招商银行
翼龙贷

now直播业务也基于该方法进行了生成了主播向量并对主播进行了聚类，初步结果来看是聚类是可以明显区分开男女主播的，并且也发现了几个有趣的主播类型，例如直播玩王者的主播，直播电影电视剧的主播，直播农村生活的主播，其主播id及描述如下，感兴趣的同学可以下载now直播搜一下看看：

229094658 吃鸡 + 王者荣耀
214471564 王者荣耀
130552418 王者荣耀
130347855 王者荣耀
109760879 王者荣耀
129839027 直播电视剧 胡军版天龙八部
116206227 直播电影 李连杰版黄飞鸿
129842270 直播电视剧 士兵突击
129857689 直播电视剧 亮剑
116202620 直播电影 林青霞
93341095 户外（农村生活）
115982309 户外（农村生活）
71264701 户外（农村生活）
128859451 户外（农村生活）
115056840 户外（农村生活）

基于内容向量的分类模型（打标签模型）

内容向量在应用宝app分类打标签上的应用

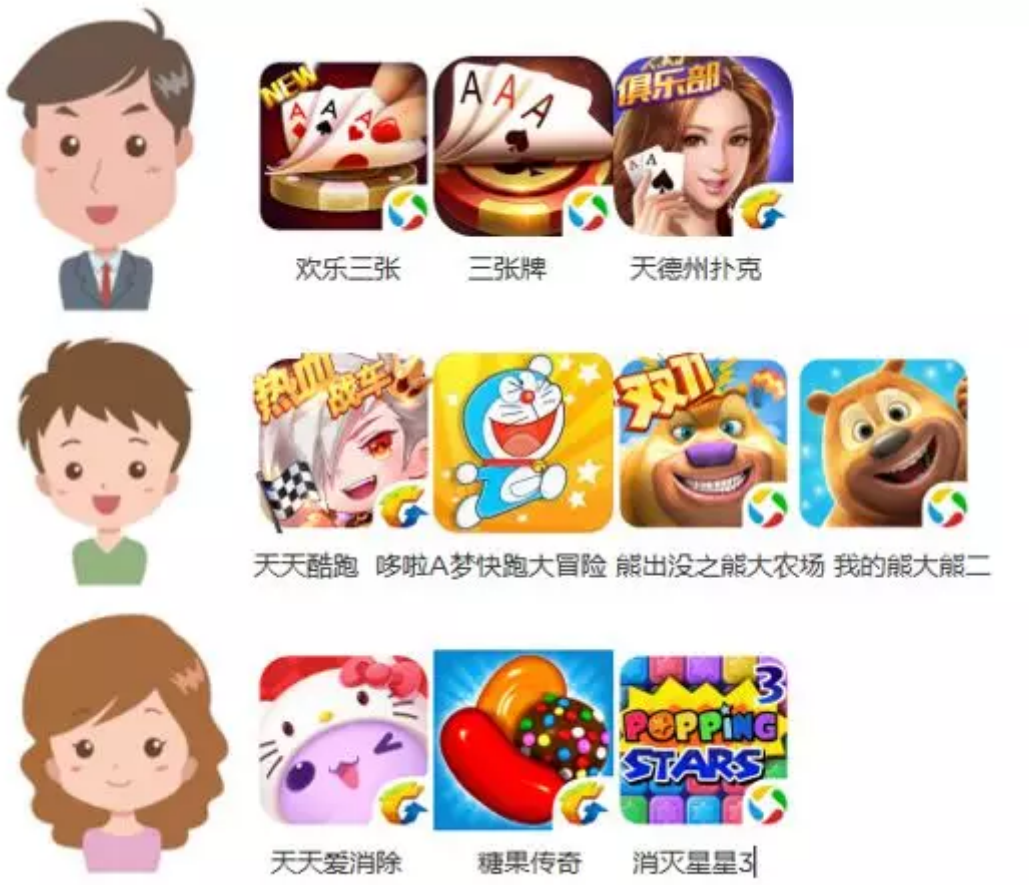
应用宝的app分类（打标签）场景长期以来都存在这样的痛点：

1. 分类体系经常会面临变动
2. app的人工标注成本高，复杂标签体系下app的标注数据很少，大多数标签仅有几个标注数据
3. app属于复杂数据结构的内容，它的内在难以用已有的算法进行挖掘，过去只能通过它的描述和图片来挖掘其信息

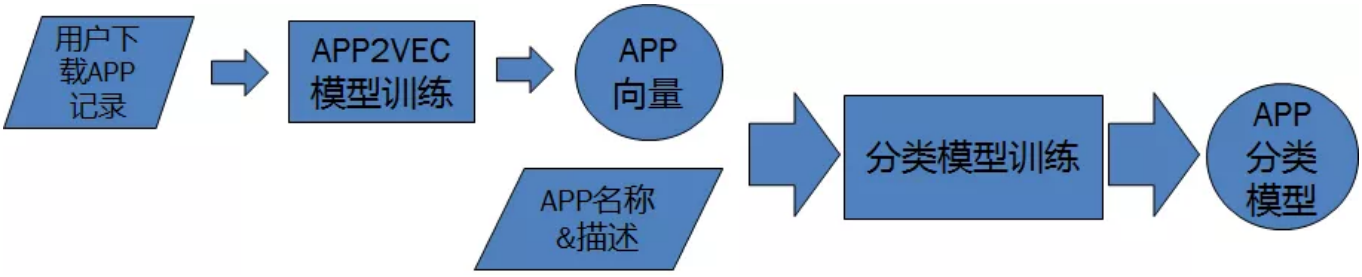
这里我们可以先思考一个问题：为什么要给app做分类和打标签？

答：给app做分类和打标签实际上是为了让用户可以更方便的找到自己想要的app，为了让我们可以更容易地结合用户兴趣给用户推送app。

从问题和答案我们可以得出一个结论：给app做分类和打标签有意义的前提是用户的行为是和app的类别、标签相关的！例如下面的这个例子里，第一位用户喜欢下载纸牌类游戏，第二位用户喜欢下载跑酷类和儿童类游戏，第三位用户喜欢下载休闲类游戏。



上面的分析我们知道用户行为应该可以用于判断app的类别标签。因此在给应用宝的app进行分类和打标签时，我们引入了基于用户行为生成的app向量。具体框架可看下图：



通过增加app向量作为分类模型的特征，可以很大程度上提高app分类的准确度（可以参考聚类中的例子），在实际业务中，部分标签的分类准确度可由40%提高到90%，整体来说准确率和覆盖度都有大幅度提升。

基于京东商品类别数据的item2vec分类模型实验

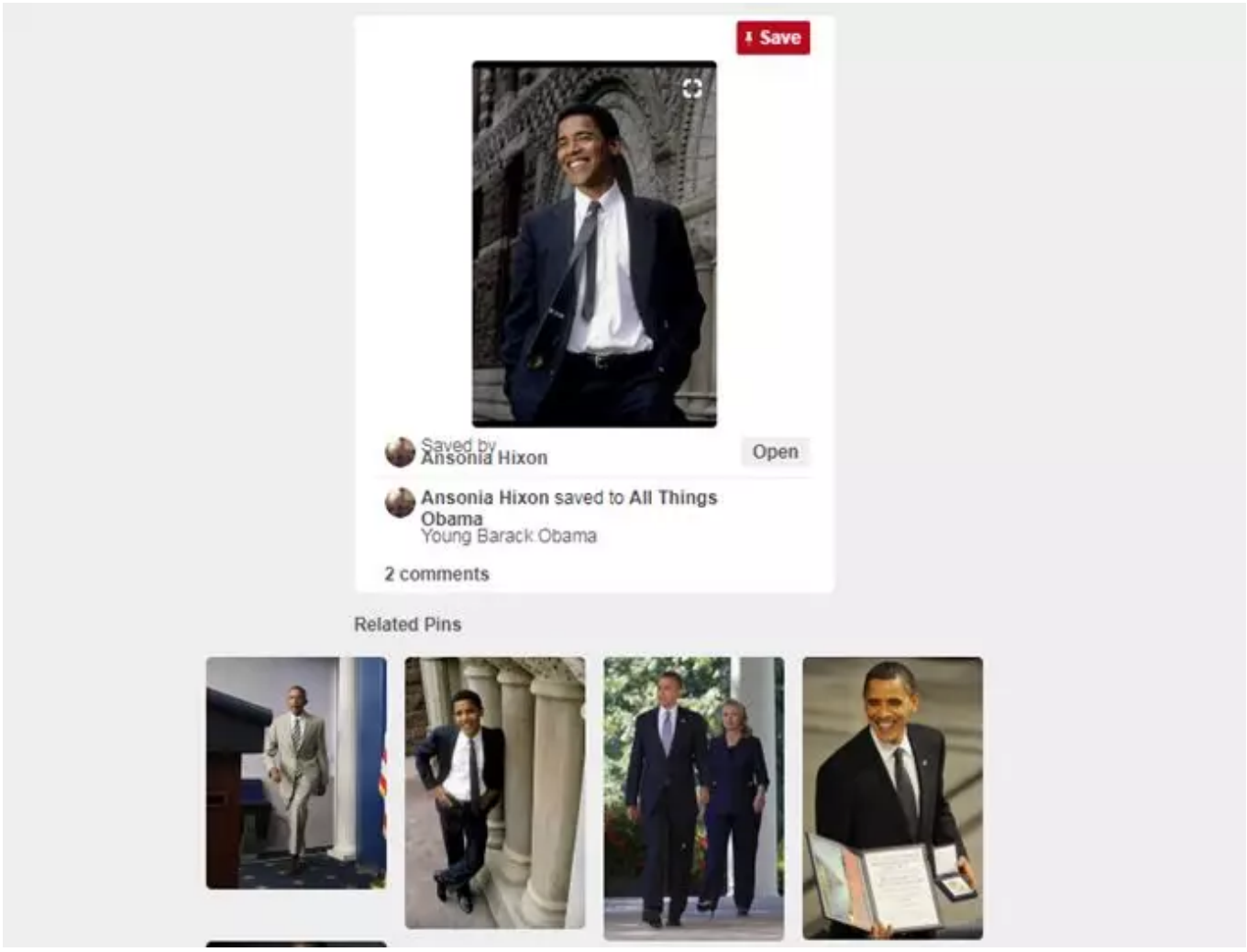
这里贴一下基于京东商品类别数据做的小实验。

1. 实验数据：京东商品一级类别，共33个类别，44776个商品，随机抽取其中80%作为训练，20%作为测试集
2. 模型：one vs rest Logistic Regression
3. 结果： 训练accuracy 0.879 测试accuracy 0.859

下图是将商品向量降维到3维后的商品空间分布图，不同颜色表示不同类别的商品，直观上看不同类别的商品在空间上的位置是不同的。

基于内容行为向量的推荐召回

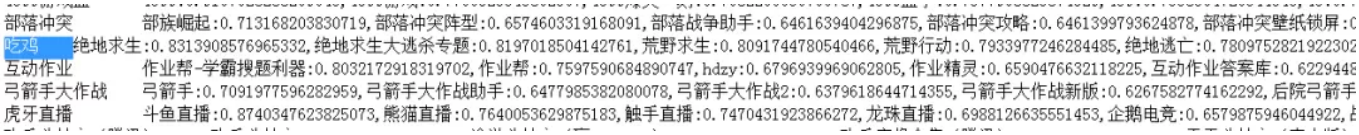
直观的例子是相关推荐，因为这一场景通常不会对召回结果做太多的加工。常见的召回结果生成方法是先计算item与item之间的相似度（一般使用cosine相似度），再取其中的top n相似item。参考文献【2】中Pinterest便使用了这种方法进行了相关推荐，其实际体验如下：



在应用宝两个场景中做了基于item行为向量的召回策略并进行了测试，相对于原模型有明显的效果提升（具体效果不便贴出，请读者谅解）。

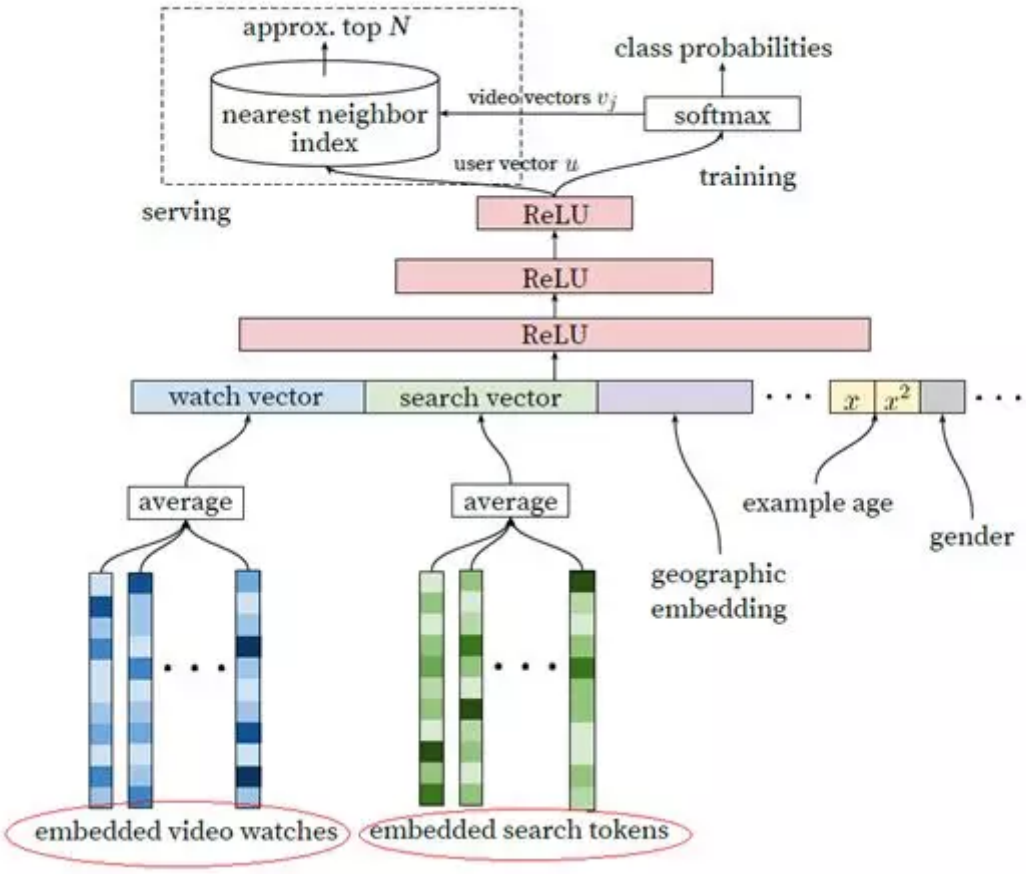
基于内容行为向量的语义召回

在app搜索场景尝试基于行为数据生成搜索词向量来优化了语义召回，明显增强了词的模糊匹配能力。举一个更直观的例子，吃鸡游戏出来的时候，搜索吃鸡出来的都不是吃鸡游戏，但是对此感兴趣的用户后续还是会去找到正确的搜索词，例如之后搜索“荒野行动”，或是下载了“荒野行动”，基于这些行为，可以将“吃鸡”和“荒野行动”关联起来。下图是单纯基于用户行为的搜索词召回：



直接作为深度学习推荐模型的输入特征

大致的思路是通过对用户有过行为item向量求均值得到用户的固定维度user特征，然后作为输入层的输入。YouTube的论文里证明了这种方法的有效性，下图是YouTube的推荐系统方案，详情可看参考文献【3】。



参考文献

【1】《 word2vec Parameter Learning Explained 》

【 2 】《 Related Pins at Pinterest: The Evolution of a Real-World Recommender System 》

【3】《Deep Neural Networks for YouTube Recommendations 》