基于 Spark 的菜品推荐算法研究

赵　娟

（河南省南阳市南阳理工学院，河南 南阳 473004）

摘要：基于排行榜模式的热门菜品推荐方式难以满足目前的用户需要。为了满足大量餐饮用户的需求，本论文依据餐饮外卖平台的海量历史订单数据，研究出每个用户的兴趣爱好，口味特点，发现其兴趣点，从而准确发现各个用户的需求，尤其是将长尾产品（指餐饮外卖平台中热销菜品以外的菜品，它们总体数量大，但单位销量却少）准确地推荐给用户，使得在最短时间内为用户推荐最适合的菜品，提供定制化服务。

一、背景

都市生活紧张忙碌，不少上班族已经习惯于在外卖餐饮平台上订餐，外卖平台的菜品种类丰富多彩参差不齐，提供各式的美食服务。W 餐饮外卖平台向广大用户提供网上订餐服务，其市场占有率在近几年不断增加。当用户在 W 餐饮外卖平台订餐完成订餐后，平台会引导用户对品尝过的菜品进行评价打分，最高 5.0 分，最低 1.0 分。但最近运营方发现老用户的下单率呈现下降态势，来自市场部的调查报告表明，此平台的老顾客在经过一段时间的订餐后，对一些热门菜品及偏爱菜品不再产生新鲜感及满足感，因此减少了下单消费。因此，市场部门建议，针对老用户进行个性化的菜品推荐，包括用户偏爱口味菜品推荐及新菜品推荐等。

要对用户进行针对的推荐，就可以考虑在原有订餐系统的基础上建立订餐推荐模块。与搜索引擎不同，推荐系统并不需要用户输入任何内容，后台通过分析用户的访问行为，主动为用户推荐满足他们需求的菜品及新出的菜品。

在经过初步讨论与评估后，根据近期用户对菜品的评分历史数据，建立菜品推荐模型，向用户们提供菜品推荐。菜品智能推荐系统，作为餐饮外卖平台系统的扩展与补充，主要负责对用户的历史评分数据进行整理，并在此基础上生成推荐结果集。

二、用户数据获取

任何解决方案都是针对需求并建立在基础数据上的。在本节，将从分析任务需求开始，再梳理、评估与目标相关的数据资源，选择适合的推荐算法来建立推荐模型，最终形成可行的推荐方案。

以热门菜品进行推荐，也是一种常用的推荐方法，但这种推荐方式更大程度是迎合了用户的选择，就是只推荐用户都喜欢的热门菜品，但不能满足用户的个性化或差异性的口味要求。另外，热门菜品通常只占全部菜品数量的很少比例，其他大部分菜品难以得到被推荐。W 餐饮平台不但要找出用户可能喜欢的菜品，而且尽可能将菜品列表中的长尾菜品（数量多但订单量少）精准推荐给其他用户。

（二）用户的评分数据集

数据是指导解决方案的基础，因此需要先来梳理一下已有的数据资源。既然要向用户进行推荐，就要从与用户行为相关的数据分析开始。当用户对菜品进行评论打分后，网站的后台服务器就会以 JSON 格式保存这些用户评分数据。结果如下所示

- [{“UserID”：”A2WOH395IHGS0T”，
- “Rating”：5.0，
- “ReviewTime”：1483202656，
- “Review”：”风味独特，真的不错！”，
- “MealID”：”B0040HNZTW”}，
- {“UserID”：”A1YQ4Z5U9NIGP”，
- “Rating”：5.0，
- “ReviewTime”：1483202876，
- “Review”：”家常美味，推荐！”，
- “MealID”：”B00CDBTQCW”}，

（二）用户的评分数据集

因为业务数据的安全原因，用户评分数据集的数据已做了脱敏处理，只保留部分重要属性。菜品的数据集在网站的后台数据库（MySQL）中保存着菜品的数据集。

三、推荐算法研究

（一）常用推荐算法比较

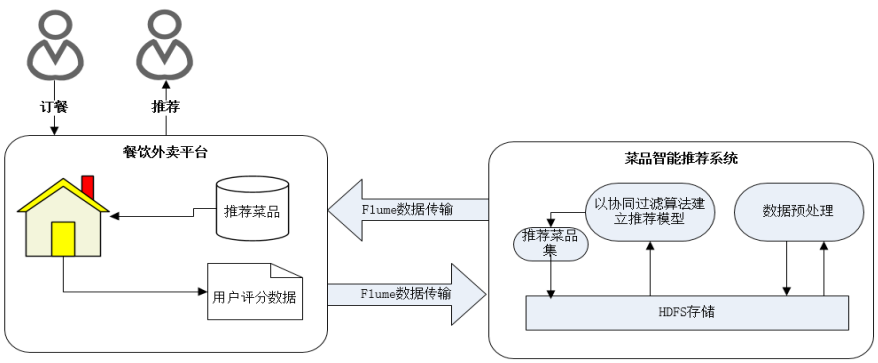
推荐算法是整个推荐系统中最核心、最关键的部分，很大程度上决定了推荐系统性能的优劣。目前，主要的推荐算法包括基于内容的推荐、协同过滤推荐、基于关联规则的推荐、基于效用推荐、基于知识推荐等。他们各自的优缺点如下表所示：

推荐方法	优点	缺点
基于内容推荐	推荐结果直观，容易解释 不需要领域知识	稀疏问题；新用户问题；复杂属性不好处理；要有足够的数据构造分类器
协同过滤推荐	新兴趣点发现，不需要领域知识 随着时间推移，推荐效果提高 推荐个性化，自动化程度高 能处理复杂的非结构化数据	稀疏问题；可扩展性问题；新用户问题；质量取决于历史数据集；系统开始时推荐质量差
基于规则推荐	能发现新兴趣点 不要领域知识	规则抽取难、耗时；产品名同义性问题，个性化程度低
基于效用推荐	无冷启动和稀疏问题 对用户偏好变化敏感 能考虑非产品特性	用户必须输入效用函数；推荐是静态的，灵活性差；属性迭代问题
基于知识推荐	能把用户需求映射到产品上 能考虑非产品属性	知识难获得 推荐是静态的

综合以上的各种推荐算法，考虑到新兴趣点发现与推荐个性化方面的表现，因此本案例将选择协同过滤推荐算法。

（二）推荐流程设计

在选定了用户数据与合适的推荐算法后，结合原餐饮外卖平台系统，设计出一份如下图所示的推荐流程方案。



上图列出了推荐系统的流程方案的相关过程，按系统开发顺序说明如下：

1. 用户评分数据的生成。用户登录餐饮平台，订餐并评分后生成用户评分记录。这些评分数据以 JSON 文件的格式存储在平台服务器上。

2. 用户评分数据的传输。当用户评分数据生成后，通过 Flume 管道传输到推荐系统所需要的分布式文件系统上。

3. 数据预处理。对用户评分数据进行探索、统计、异常值处理及数据转换，最终构造出可供建立推荐模型的数据集。

4. 以协同过滤算法建立推荐模型。使用协同过滤的多种算法对用户评分数据集进行建模，并调整相关的模型参数，以获取最优的模型。

5. 推荐菜品数据集。使用最优推荐模型导出推荐结果，并将结果存到 HDFS 上。

6. 推荐结果上传。将推荐菜品数据集通过 Flume 管道传输到网站平台的存储服务器，或者导出到 mysql 数据库中。

7. 向用户进行推荐。当某用户登录餐饮平台时，系统从数据库中读取与该用户相关的推荐结果，推送给用户。

上述推荐流程中的数据传输环节（2）和（6），通常采用以 Oozie 配置工作流的方式来实现。Oozie 是 Apache Hadoop 生态系统中的一个关键组件，它可以把多个任务组成一个工作流，自动完成任务的调用。下面主要对推荐系统中的菜品智能推荐系统进行逐步的推导与实现。

四、数据预处理

本节将从原始的用户评分数据入手，首先需要进行数据探索分析，然后根据探索分析的结果判定是否存在异常数据。如果有异常数据，则对异常数据先进行处理。

（一）加载评分数据

原始数据是以 JSON 格式存储，数据结构是固定的，每条记录是由 5 个属性构成，分别是用户 ID、菜品 ID、用户评分、用户评论、评论时间戳。因此它非常适合以 Spark SQL 方式来加载，生成 DataFrame 后进行数据查询。将原始数据加载到 DataFrame 后进行评分数据的查询，比如只显示前 5 条记录的 scala 代码如下：

```
scala> // 用户评分数据，读取前5条

scala> mealResults.take(5).toList.foreach(println)
[A2WOH395IHGS0T,B0040HNZTW,5.0,风味独特，真的不错！,1496177056]
[A32KHS0VN0N0HB,B006Z48TZS,3.0,有特色，也比较卫生,1496177108]
[A1YQ4Z5U9NIGP,B00CDBTQCW,5.0,家常美味，推荐！,1496177276]
[A3E5V5TSTAY3R9,B00751IYQ4,4.0,好吃,1496179256]
[A1V50CTTDJ73ZM,B00C00LT6S,5.0,不得不赞,1496180009]
```

（二）评分数据的探索与统计

在 Spark SQL 的处理框架下，非常方便地使用 SQL 语句对数据进行查询，下面为对数据的分布及其他属性的统计结果。

总纪录数	总用户数	总菜品数	最高评分	最低评分
38384	5130	1685	5.0	1.0

（三）评分数据的分组统计

下面对评分项进行分组统计。其实现代码与统计结果如下图所示。结果表明大部分用户的评分为 4.0 分与 5.0 分，占总体数量的 79%。这说明用户对于餐饮平台的提供的菜品，总体评价还是比较正面的。

```
scala> // 各级评分的分组统计

scala> hiveCtx.sql("select Rating,count(*) AS Num from mealData group by Rating order by Rat
ing").show()
+------+-----+
|Rating|  Num|
+------+-----+
|   1.0| 1782|
|   2.0| 1960|
|   3.0| 4475|
|   4.0| 9142|
|   5.0|21025|
+------+-----+
```

五、实现菜品推荐

经过上节的推荐模型评估，综合考虑后选择了基于物品与基于 Spark ALS 算法建立的推荐模型。本节将采用这两种模型实现向用户推荐新菜品。

（一）向用户推荐 10 个新菜品

任意选择一个用户，例如，为测试用户（用户编码 =1000）推荐 10 份预测评分最高的菜品。这里的菜品将引入真实的菜品名称，因此需要从外部数据库中加载菜品详细信息数据。另外，由于推荐模型是以用户编码和菜品编码来建立的，因此产生的推荐结果集也是由用户编码与菜品编码组成，所以还需要加入一个反编码的处理过程，即把用户编码转换为用户 ID，把菜品编码转换为菜品 ID。

向测试用户推荐 10 份预测评分最高的菜品。这里的菜品将引入真实的菜品名称，因此需要从外部数据库中加载菜品详细信息数据。加载用户与菜品的编码数据集。加载外部数据库中的菜品数据。生成推荐数据集。

```
scala> // 引用真实的菜品名称,最终的推荐结果集（UserID,MealID,Meal,Rate）

scala> val realRecommendationRecords = recommendationRecords.map{
     | case (user,meal,rating) => (user,meal,mealsMap.get(meal).get,rating)}.foreach(pri
ntln)
(A1PTBXZG04GT29,B00627P604,蒜蓉荷兰豆,5.665500826093317)
(A1PTBXZG04GT29,B00CDBQ9B8,当归红枣蛋,5.1925772045089715)
(A1PTBXZG04GT29,B00EM6MGHI,干煸苦瓜,5.156393439895823)
(A1PTBXZG04GT29,B000UU4IVM,虾仁西兰花,5.134808791464488)
(A1PTBXZG04GT29,B0084RZPPQ,柠檬海蜇头,5.109945250678736)
(A1PTBXZG04GT29,B009F7QE0E,炸猪排升级版,5.094849874175011)
(A1PTBXZG04GT29,B000I62JFU,锅塌豆腐,5.093429890054092)
(A1PTBXZG04GT29,B0047YZISE,萝卜烧肉,5.065520644335779)
(A1PTBXZG04GT29,B009ZQ7M7O,自制蕃茄酱,5.057789120795025)
(A1PTBXZG04GT29,B000ZU1KMW,家传红烧肉,5.047784261958275)
realRecommendationRecords: Unit = ()
```

推荐结果评价。

检查测试用户（用户编码 =1000）在训练数据集中的记录，它共有 5 条菜品记录，这些记录基本反映了该用户的口味及爱好。为便于比较，对这些菜品记录按类别进行整理，如下表所示。

测试用户的训练数据

序号	菜品	类别
1	干煸豆角	素菜
2	妈妈牌红焖肉	猪肉
3	海鲜炖蛋	海鲜/蛋
4	橙汁鸡球	鸡肉
5	台湾泡菜	佐餐

数字媒体时代下职业院校美术教改思考

孙光宇
（江苏省无锡交通高等职业技术学校，江苏无锡 214000）

摘要：科学技术是第一生产力，当今时代下的中国，科学技术的发展日新月异，数字媒体技术蓬勃发展，在社会各领域得到了普遍关注与应用，并促进了社会的全面发展，同时为我国高等院校的教育事业的发展提供了难得的契机。关于高等院校的美术教学，数字媒体技术对于美术教学模式的创新与发展起到了至关重要的作用，很大程度上弥补了以往教学的不足。本文就当前我国职业院校教学的实际情况进行阐述，并分析职业院校教学中现阶段应用先进技术存在的问题，最后针对性的给出了相关建议。

关键词：数字媒体；职业院校；教学改革；思考；美术

近年来，数字媒体技术的发展，给我国各所高等院校的美术教学带来了新的机遇和挑战。以往的大学美术教学活动中，因为教学模式过于单调机械，很多学生难以对对美术课的学习产生热情和兴趣，再加上理论知识的枯燥无趣，导致学生仅仅记住了最基础的理论知识，动手实践操作的技能严重欠缺。幸运的是，数字媒体技术在大学美术教学中的应用，打破了传统教学模式中的这一瓶颈，为满足社会需要，提供更优秀的美术人才，注入了强心剂。穷则思变，变则通，通则久，在数字媒体日新月异进步的时代背景下，高校美术要善于与时俱进，摒弃以往教学模式不合时宜的部分，大刀阔斧地进行美术教育教学改革，通过我国当前各高校美术教学的整体情况分析可知：美术教学存在的一定的问题。

一、现阶段我国职业院校的教学状况分析

（一）教学理念陈旧

1. 不同地区间差异大

现阶段，由于西部与中东部之间的经济水平存在的差异较大，我国的西部地区美术教学水平和中东部地区相差甚远，这其中存在因素有很多，一方面是因为教学资源的分布不均，另一方面原因是老师的教学理念比较陈旧。可以说，现如今还有部分西部院校的美术教师没有真正地领悟到新的社会背景下对美术教学提出的新定位照搬照套，遵循传统的教学理念，认为美术教学就是教会学生画画、教好学生画画，然而在新的时代背景下，这样是远远不够的，何况还是高等教育的美术教学。在时代不断发展的今天，西部地区的教师无法与时俱进，与新时期的教育理念同步是绝对不行的，这会对学生的发展造成很大的消极影响，无法适应新时代社会对于美术人才的需求。

2. 不同院校之间差异大

不同职业院校由于经济实力和对美术教育重视程度的不同，美术课的课时设置与师资水平均不一样，这也会导致培养出的美术系毕业生的能力与素养差异很大。在高校开设美术教学的时候，应当充分意识到当前美术教学的突出问题与缺陷，并结合本校数字媒体的技术优势，着力提升美术专业的教学实力，弥补缺陷。我国部分职业院校在美术教学方面不重视以人为本的理念，他们

最后给出测试用户（用户编码 =1000）的两种推荐结果按类别进行展示，如下表所示：

基于ALS的菜品推荐

序号	菜品	类别
1	蒜蓉荷兰豆	素菜
2	当归红枣蛋	蛋
3	干煸苦瓜	素菜
4	虾仁西兰花	海鲜
5	柠檬海蜇头	海鲜
6	炸猪排升级版	猪肉
7	锅塌豆腐	其他
8	萝卜烧肉	猪肉
9	自制番茄酱	佐餐
10	家传红烧肉	猪肉

基于物品的菜品推荐

序号	菜品	类别
1	荔枝虾球	海鲜
2	干煸四季豆	素菜
3	润肺清补凉汤	汤品
4	咖喱猪肉饭	猪肉
5	鲜笋焖饭	猪肉
6	泉州炸醋肉	猪肉
7	纯纯的豆浆	饮品
8	五香熏鱼	鱼
9	彩椒烤鸡串	烧烤
10	豆腐皮烤菜卷	烧烤

这样，就将这 10 个菜品推荐给测试用户（用户编码 =1000），更满足用户的口味和风格。

（二）向所用户推荐新菜品

当经过评估选定模型后，可以应用模型对所有用户生成推荐数据集。以下以基于物品与基于 Spark ALS 的推荐模型为例，导出推荐结果集。由于基于物品与基于用户的两种推荐模型，在计算推荐结果集时的方法非常类似，在此不展开说明。基于 Spark ALS 的推荐模型，它本身就已经包括了几个常用的推荐方法接口，可以非常方便导出推荐结果。

1. 为指定用户进行 topN 推荐，即推荐 N 个预测评分最高的物品。

2. 为“用户 – 物品”对进行预测评分，即预测某用户对某物品的评分。

3. 为所有用户推荐 TOP N 个物品。

4. 为所有物品推荐 TOP N 个用户。

5. 最终的推荐结果集，保存到 Hadoop 的 HDFS 文件系统中。

6. 也可以根据业务需要，直接保存在外部数据库，比如 MySQL 或 HBase 上。

六、结语

本文详细阐述了基于 Spark 的餐饮服务行业中大数据推荐系统实现过程，从案例背景、实现目标、系统整体架构及流程设计等方面展开，分步骤实现了系统建模。同时，针对系统实现的各个过程，包括前期方案设计、数据探索、数据预处理，到后期数据建模、模型寻优、模型评价及推荐等，都提供了分析思路与参考代码。