浙 江 大 学

硕士学位论文开题报告

**（专业学位）**

论文题目： 团购商品个性化推荐系统的设计与实现

姓 名： 李朝阳

学 号： Z134325272

专 业： 软件工程

院 别： 软件学院

导 师： 李际军,朱小军

二零×× 年 ×

# 1.项目来源，课题依据，背景

团购 -团体购物，将消费者联系起来与商家进行谈判，以求最优价格的购物方式，团购商品在中国已日益盛行，涉及衣食住行等各个方面， 由于每个团购平台的栏位有限以及消费者差异性等，一个或者几个团购专栏不可能满足每个消费者， 再有广告位的投入产出比等考虑价值因素，为消费者进行个性化推荐已日渐成为每个团购平台的考虑的重要因素。

# 2.中国团购个性化推荐现状分析以及问题

1. 随着互联网的不断发展，信息爆发式的增长，团购商品信息已经出现过载的情况；用户希望快速定位到自己需要的信息上去， 如果网站不能满足那么消费者就回转向其他网站，造成用户流失，用户的转化率降低， 商品的成交率下降。

2. 不能充分利用资源；如平台有大量的团购商品要推荐，但商品栏位有限；而一种商品不可能适应所有的用户；这时候个性化推送就显得尤为重要了。

3. 个性化商品推荐的性能问题；根据推荐算法计算推荐商品， 然后推送给用户，整个过程的处理时间可能达到消费者的忍耐极限，从而造成损失；

4. 冷启动问题， 当一个新的用户过来的时候，没有任何关于他的个人信息，无法个性推荐；同理当一个新的商品过来的时候，没有任何评价（显式数据）或者点击行为（隐形数据）与用户关联。

# 3.系统目标以及可行性分析

## 3.1系统预期目标

1. 推荐的质量

个性化推荐，推荐个体所想所需，你需要的是根据用户的历史数据分析用户可能需要的， 如果推荐的内容与用户所想要南辕北辙，那么你的推荐就是失败的，保证推荐的质量是每个推荐系统所必须的。

2. 可扩展性

一种推荐算法不可能适合所有的推荐业务需求，例如适合推荐美食的，根据用户的美食消费记录， 以及用户所在的位置，消费的价格等综合加权分值并排序，取Top-m项即可；但若要推荐啤酒与尿布，那就需要另一种算法，来计算其关联规则了，因此推荐系统的扩展性是必须的，应该随着业务的增长而增长。

3. 高效性

一般用户的历史数据是放置在网络日志文件中的， 其存放方式是分布式的， 而且量也比较大， 所以对于一些实时性要求比较高的推荐系统就很麻烦了；例如：新浪微博的“推荐好友”功能，当你关注一个话题，或者好友的时候，推荐系统要实时的计算你可能感性的人；同时这也引出了推荐系统的另一个目标；

4. 推荐理由

系统向用户推荐了一个产品或者好友，需要个推荐规则也就是推荐理由，推荐规则如何表现，数学表示比较生硬，不容易理解，当然一些简单的是很容易理解的，如商品的销量排行， 用户的评价等， 但是啤酒与尿布背后的推荐理由如何解释？

## 3.2 可行性分析

### 3.2.1技术可行性分析

本系统采用比较成熟的框架实现，使用文档类型数据库存储网络日志文件， 使用MapReduce进行推荐计算处理， 推荐算法采用组合推荐（混合常用的推荐算法）采用关系型数据库存储分析结果，并采用MyBaits+Spring+SpringMVC搭建Web application，前段显示采用HTML5+CSS+JAVASCRIPT。以上是系统的大体描述，各项技术都比较成熟完全可应用于生产。

### 3.2.2 操作可行性分析

1. 算法的设计，或者说是算法框架的设计， 根据本次项目的具体需求设计；

2. 数据结构设计以及数据库的设计

3. 系统架构搭建

4. 测试用例的设计（爬虫， 数据的处理， 数据的计算）

5. 系统实现

6. 验证

# 4.概要设计

### 4.1.1 基础架构图

**整体图**



**推荐部分：**



### 4.1.2 推荐系统

#### 团购特点

团购商品往往分为两类:地域性商品和非地域性商品。地域性商品指只为本

地用户提供的商品，如KTV券、按摩券等需要在本地商店消费的服务类商品。非地域性商品指销售不受地域范围影响的商品，如服装、玩具等面向全国消费者销售的商品。一般团购网站这两种商品同时存在。地域性商品推荐，推荐系统的训练集数据为同一地域的商品和用户;非地域性商品推荐，推荐系统的训练集数据为全部的非地域性商品和购买它们的用户。两者仅仅是数据规模不同，推荐方法相同。团购电子商务网站的特点主要有以下几点:

1)商品上架时间短

团购网站的销售方式是限时折扣，所以商品在线销售的时间很短，一般是一天或个位数天数。而且团购商品有数量的限制，售出一定数量后也会结束销售，所以热卖商品预计上架一天也可能因为在一天内售完数量而提前下架。相比于传统电子商务网站商品的长期供应，团购商品的上架时间很短。

2)商品变化快

由于每种商品上架时间很短，所以商品种类变化很快。团购网站一般是每天

会更新一批商品上架，有些团购网站如淘宝的聚划算每天在固定时间开团，更换

所有商品。

3)用户反馈慢

由于团购网站的商品变换频繁，在用户收到商品并给商品评分前该商品很可

能已经下架，所以不能及时获得用户对商品的评分。难以通过用户反馈的信息来

获得用户对于商品的喜好程度。

4)上架商品数量少

团购网站同天销售的商品数较少，月\_数量比较固定，一般都在几百范围内。

不像传统电子商务网站动辄上力商品，甚至达数亿商品。

5)用户注册信息少

用户注册信息的完整性很低，有效信息很少，绝大多数用户没有填充完整他

们的个人信息，不易参照注册信息进行用户划分。因此参照注册信息、，用统计学

提供推荐的方法准确率很低。

6)数据的稀疏性

用户有过行为的商品数量占总商品数量的比例较小，即使在团购网站较少的

商品数量情况下稀疏性问题也不容忽略。如本实验数据中取某一天数据，``M上

#### 基于关联规则的推荐算法

根据[韩家炜](http://baike.baidu.com/view/1623478.htm" \t "_blank)等观点，关联规则定义为：假设[http://c.hiphotos.baidu.com/baike/s%3D220/sign=da6901a2b8a1cd1101b675228913c8b0/a50f4bfbfbedab6421faa852f536afc378311e93.jpg](http://baike.baidu.com/picture/1076817/1076817/0/a50f4bfbfbedab6421faa852f536afc378311e93?fr=lemma&ct=single)是项的集合。给定一个交易数据库D，其中每个事务(Transaction)t是I的非空子集，即，每一个交易都与一个唯一的[标识符](http://baike.baidu.com/view/390932.htm" \t "_blank)TID(Transaction ID)对应。关联规则在D中的[支持度](http://baike.baidu.com/view/4335695.htm" \t "_blank)(support)是D中事务同时包含X、Y的百分比，即[概率](http://baike.baidu.com/view/45320.htm" \t "_blank)；[置信度](http://baike.baidu.com/view/434404.htm)(confidence)是D中事务已经包含X的情况下，包含Y的百分比，即[条件概率](http://baike.baidu.com/view/965891.htm" \t "_blank)。如果满足最小支持度[阈值](http://baike.baidu.com/view/409216.htm)和最小[置信度](http://baike.baidu.com/view/434404.htm)阈值，则认为关联规则是有趣的。这些阈值是根据挖掘需要人为设定。

关联规则挖掘过程主要包含两个阶段：第一阶段必须先从资料集合中找出所有的高频项目组(Frequent Itemsets)，第二阶段再由这些高频项目组中产生关联规则(Association Rules)。

关联规则挖掘的第一阶段必须从[原始资料](http://baike.baidu.com/view/2462344.htm" \t "_blank)集合中，找出所有高频项目组(Large Itemsets)。高频的意思是指某一项目组出现的频率相对于所有记录而言，必须达到某一水平。一项目组出现的频率称为[支持度](http://baike.baidu.com/view/4335695.htm)(Support)，以一个包含A与B两个项目的2-itemset为例，我们可以经由[公式](http://baike.baidu.com/view/645857.htm" \t "_blank)(1)求得包含{A,B}项目组的支持度，若支持度大于等于所设定的最小支持度(Minimum Support)门槛值时，则{A,B}称为高频项目组。一个满足最小支持度的k-itemset，则称为高频k-项目组(Frequent k-itemset)，一般表示为Large k或Frequent k。算法并从Large k的项目组中再产生Large k+1，直到无法再找到更长的高频项目组为止。

关联规则挖掘的第二阶段是要产生关联规则(Association Rules)。从高频项目组产生关联规则，是利用前一步骤的高频k-项目组来产生规则，在最小信赖度(Minimum Confidence)的条件门槛下，若一规则所求得的信赖度满足最小信赖度，称此规则为关联规则。例如：经由高频k-项目组{A,B}所产生的规则AB，其信赖度可经由[公式](http://baike.baidu.com/view/645857.htm" \t "_blank)(2)求得，若信赖度大于等于最小信赖度，则称AB为关联规则。

#### 基于内容的推荐算法

基于内容的信息推荐方法的理论依据主要来自于[信息检索](http://baike.baidu.com/view/45496.htm)和[信息过滤](http://baike.baidu.com/view/506542.htm)，所谓的基于内容的推荐方法就是根据用户过去的浏览记录来向用户推荐用户没有接触过的推荐项。主要是从两个方法来描述基于内容的推荐方法：启发式的方法和基于模型的方法。启发式的方法就是用户凭借经验来定义相关的计算公式，然后再根据公式的计算结果和实际的结果进行验证，然后再不断修改公式以达到最终目的。而对于模型的方法就是根据以往的数据作为数据集，然后根据这个数据集来学习出一个模型。一般的推荐系统中运用到的启发式的方法就是使用[tf-idf](http://baike.baidu.com/view/1228847.htm)的方法来计算，跟还有tf-idf的方法计算出这个文档中出现权重比较高的关键字作为描述用户特征，并使用这些关键字作为描述用户特征的向量；然后再根据被推荐项中的权重高的关键字来作为推荐项的属性特征，然后再将这个两个向量最相近的（与用户特征的向量计算得分最高）的项推荐给用户。在计算用户特征向量和被推荐项的特征向量的相似性时，一般使用的是cosine方法，计算两个向量之间夹角的cosine值。

#### 基于协同过滤的推荐

基于协同过滤的推荐算法理论上可以推荐世界上的任何一种东西。图片、音乐、样样可以。 协同过滤算法主要是通过对未评分项进行评分 预测来实现的。不同的协同过滤之间也有很大的不同。

基于用户的协同过滤算法: 基于一个这样的假设“跟你喜好相似的人喜欢的东西你也很有可能喜欢。”所以基于用户的协同过滤主要的任务就是找出用户的最近邻居，从而根据最近邻 居的喜好做出未知项的评分预测。这种算法主要分为3个步骤：

一，用户评分。可以分为显性评分和隐形评分两种。显性评分就是直接给项目评分（例如给百度里的用户评分），隐形评分就是通过评价或是购买的行为给项目评分 （例如在[有啊](http://baike.baidu.com/view/1909861.htm)购买了什么东西）。

二，寻找最近邻居。这一步就是寻找与你距离最近的用户，测算距离一般采用以下三种算法：　1.皮尔森相关系数。　2.余弦相似性。　3调整余弦相似性。　调整余弦 相似性似乎效果会好一些。

三，推荐。产生了最近邻居集合后，就根据这个集合对未知项进行评分预测。把评分最高的N个项推荐给用户。 这种算法存在性能上的瓶颈，当用户数越来越多的时候，寻找最近邻居的复杂度也会大幅度的增长。

因而这种算法无法满足及时推荐的要求。基于项的协同过滤解决了这个问题。 基于项的协同过滤算法 根基于用户的算法相似，只不过第二步改为计算项之间的相似度。由于项之间的相似度比较稳定可以在线下进行，所以解决了基于用户的协同过滤算法存在的性能瓶颈。

#### 组合推荐

由于各种推荐方法都有优缺点，所以在实际中，组合推荐（Hybrid Recommendation）经常被采用。研究和应用最多的是内容推荐和协同过滤推荐的组合。最简单的做法就是分别用基于内容的方法和协同过滤推荐方法去产生一个推荐预测结果，然后用某方法组合其结果。尽管从理论上有很多种推荐组合方法，但在某一具体问题中并不见得都有效，组合推荐一个最重要原则就是通过组合后要能避免或弥补各自推荐技术的弱点。

在组合方式上，有研究人员提出了七种组合思路：  
　　λ1）加权（Weight）：加权多种推荐技术结果。  
　　λ2）变换（Switch）：根据问题背景和实际情况或要求决定变换采用不同的推荐技术。  
　　λ3）混合（Mixed）：同时采用多种推荐技术给出多种推荐结果为用户提供参考。  
　　λ4）特征组合（Feature combination）：组合来自不同推荐数据源的特征被另一种推荐算法所采用。  
　　λ5）层叠（Cascade）：先用一种推荐技术产生一种粗糙的推荐结果，第二种推荐技术在此推荐结果的基础上进一步作出更精确的推荐。  
　　λ6）特征扩充（Feature augmentation）：一种技术产生附加的特征信息嵌入到另一种推荐技术的特征输入中。  
　　λ7）元级别（Meta-level）：用一种推荐方法产生的模型作为另一种推荐方法的输入。

#### 推荐算法设计

由上述团购网站特才点可知，团购与传统电子商务有许多不同，因此应用于团

购网站的个性化推荐系统也必须根据这些特点量身定做。由于团购网站的推荐有

较强的实时性要求，因此个性化推荐算法在保证准确率的前提一下必须有较强的实

时一性。

团购网站有商品变化快的特点，若应用关联规则推荐，有的团购网站每日更

新所有商品，则刚上架时必然无法达到最小支持度和置信度，使得在每日团购开

始时无法形成个性化推荐，且由于最小支持度的存在，某些销量较少的商品不能

进入推荐列表，推荐的覆盖度不足。

基于内容的推荐需要从商品介绍的文本内容里挖掘出商品的相似性，由于团

购网站每日上架商品中类别跨度大，属于同一类别的商品较少，将会导致同类商

品必定出现在推荐列表中，效果与非个性化推荐的同类商品推荐方法相似，不能

很好地起到个性化推荐的作用。

隐语义模型方法在很多实际应用中反映出较好的准确性，但相对地，其实时

性很差，新用户或新项目的加入，需要通过重新扫描所有数据才能构建模型。用

户在网站中的新行为不能立即改变推荐列表，因此不适合用在实时性要求较高的

领域。

基于项目的协同过滤实时性较好。由于团购网站上架商品较少，用户数量远

远大于商品数，且用户增加较快，若应用基于用户的协同过滤，则由于用户数量

庞大寻找最近邻用户将耗费大量时间，而通过找到专家用户的方法又自动忽略了

大批用户对推荐准确率会造成影响。而项目的相似性稳定性较高，许多研究证明

基于项目的协同过滤在用户数大大多一于商品数的环境下应用的效果较好，在实际

应用中基于项目的协同过滤应用得最)一泛，因此本文将基于项目的协同过滤方法

应用到团购电子商务领域中。

不同于传统电子商务网站巨架商品稳定，团购网站的上架商品经常处于变动

中，用户上次购买的商品很可能已经不出现在本日的上架商品中了。因此，传统

的基于项目的协同过滤推荐方法中计算所有项目相似度的力一法在团购领域EI'.难

以应用。本文设计的推荐算法考虑用户本}飞表现出来的偏好以及历史行为表现出

的偏好，综合得出用户对商品的偏好。

### 4.1.3 数据结构以及数据库设计

#### 用户行为日志

由于用户的行为的数据量很大， 而且数据的数据结构不是一成不变的，一次最好采用分布式的，非关系型数据设计；

#### 商品信息， 推荐数据

商品的信息数据结构一般是不变的， 而推荐结果集同样是固定的，因此可以考虑部分放入关系型数据库中；

### 4.1.4 系统架构

团购商品推荐系统通过用户接口接受推荐请求，在推荐引擎的帮助下选择不

同的推荐模块产生推荐，保证最符合用户偏好的商品处在个性化推荐列表之中。

推荐系统还可以从用户不断增加的行为日志中学习用户的偏好，修正本日推荐模

型和用户历史偏好模型，进一步改善推荐质量，从而提高个性化推荐的准确率。

目前多数电子商务网站推荐系统的架构，推荐引擎和推荐算法是一体化的，

这意味着面对任何情况推荐系统都是用同一种算法实现推荐，导致推荐功能单

一，难以满足推荐应用中不同的推荐需求。本文由于在推荐策略中考虑了多种推

荐需求，因此，采用一个多模型的推荐引擎架构，以满足不同用户类型的推荐需

求。

### 4.1.5 测试用例

#### 推荐预测准确率

这是一个重要的离线评测指标，根据评分预测和TopN推荐这两种不同场合

而采用不同的方法。现在大多数研究集中于评分预测，这是因为推荐系统早期研

究组Grouplens的研究主要是基于电影评分数据库进行的。对此，Amazon前科

学家Greg Linden持不同看法，09年他在Communications of the ACM网站发

表了一篇文章指出电影推荐的目的在于找到用户最感兴趣的电影而不是猜测用

户看了电影后会给电影怎样的评分，因此TopN更符合实际应用的需求。

(1)评分预测

适用于预测用户对项目评分的场合l\-，般通过平均绝对误差(MAE)方法

计算。

(2)ＴopN推荐

网站在提供推荐服务时，一般给用户一个长度为N的推荐表，这种推荐叫TopN推荐。

除上述常用评估指标外，还有一些评估方法来评估推荐算法。

多样性评估方法认为给用户的推荐列表需要满足用户广泛的兴趣，即兴趣的

多样性。多样性描述了推荐列表中物品两两之间的不相似性，与相似性是对应的。

新颖性评估是指给用户推荐的物品里有多少是他们以前没听说过的。在网站

中实现新颖性的最简单方法是，把那些用户之前在网站中对其有过行为的物品从

推荐列表中过滤掉。0' scar Celma(2010)在博l.论文中研究了新颖度的评测，

最简单的方法是利用推荐结果的评价流行度，因为越不热门的物品越可能让川)”

觉得新颖。

惊喜度是最近儿年推荐系统领域最热门的话题。Guy Shani(2011)的沦文认

为，如果推荐结果和用户的历史兴趣不相似，但却让用户觉得满意，那么就可以

说推荐结果的惊喜度很高。与新颖性的区别在于，新颖性仅仅取决于用户是否听

说过这个推荐结果。

信任度评估用户对推荐系统的信任程度，度量方法只能通过问卷调查方式。

提高信任度的方法主要有两种:首先是增加推荐系统的透明度，Henriette

Cramer (2008)等人的研究认为，增加推荐系统透明度的主要方法是提供推荐解

释。其次是考虑用户的社交网络信息，利用好友给用户做推荐，并且用好友进行

推荐解释。

实时性也是一个重要指标，在某些需要快速产生推荐的网站里尤其重要。实

时性包括两个方面:一方面推荐系统需要实时地更新推荐列表来满足用户新的行

为变化，另一方面是推荐系统需要能够将新加入系统的物品推荐给用户。

健壮性指标衡量一个推荐系统抗击作弊的能力，评测方法主要利用模拟攻

击。Neil Hurley(2011)专门为2011年推荐系统大会编制了一个关于推荐系统健

壮性的教程。

#### 测试数据收集设计

计并实现一个爬虫系统， 获取有关团购美食类的各种美食数据；并将数据进行一定的推荐计算， 存放如MySQL中，作为目标数据；

用户行为记录可通过经典用户行为数据库进行分析；

# 5.项目开发计划以及预期成果

**计划表：**

|  |  |  |
| --- | --- | --- |
| **编号** | **起止日期** | **计划完成内容** |
| 1 | 2014年10 - 2014年12 | 查阅文献资料，编写课题开题报告 |
| 2 | 2014年12- 2015年2 | 推荐算法选定和设计 |
| 3 | 2015年 2 -2015年3 | 系统的文档和功能迭代 |
| 4 | 2015年3 – 2015年4 | 系统调试和维护 |
| 5 | 2015年5月 | 撰写论文正文 |

**成果：**

1. 推荐系统设计文档
2. 推荐系统的系统实现