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



### 4.1.2 推荐算法

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" \t "_blank)(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。

### 4.1.3 数据结构以及数据库设计

### 4.1.4 系统架构

MyBatis

Spring

SpringMVC

### 4.1.5 测试用例

# 5.项目计划进度以及预期结果