

基于天猫交易数据的分析与挖掘*

王希梅

清华大学 软件学院

北京 100084

wxm17@mails.tsinghua.edu.cn

于千山

清华大学 软件学院

北京 100084

yqs17@mails.tsinghua.edu.cn

陳善宇

清华大学 软件学院

北京 100084

abc321094@gmail.com

摘要

我们利用 2014 年双 11 (11 月 11 日) 前六个月天猫的用户行为日志进行数据挖掘任务。在本次任务中, 我们提出四个问题, 并分别设计算法进行解决, 同时给出了实验结果并进行讨论。这几个问题的解决, 对于商家降低促销成本, 提高投资回报率 (ROI) 非常重要。

关键词

数据挖掘, 聚类分析, 行为预测

1. 问题描述

我们提出以下四个问题:

1. 商品频繁模式挖掘: 即分析哪些商品会被一起购买。
2. 同一商家用户的分类: 商家会拥有很多用户, 对于商家来说, 如果能够把握用户的行为, 清楚用户的类别, 就可以更好地提供服务、贩卖商品, 并提高投资回报率。在该问题中, 我们考虑将同一商家的用户分为忠实客户、潜在忠实客户和非忠实客户三类。
3. 同一商家常被购买物品种类 or 某些特定物品 前者意味着还要对物品进行特征提取和分类: 通过对用户行为日志的分析, 抽取出用户喜爱的商品种类。对于商家来说, 这有助于把握用户喜好, 以便于进行针对性促销。
4. 用户重复购买的预测: 为了吸引大量的新买家, 商家有时会在特定的日子进行大促销。然而, 很多买家都是一次性买家, 这些促销活动可能对销售的影响不大。我们期望预测未来哪些特定商家的新买家将成为忠实客户。这些新买家将来再次购买同一批商品的可能性。

*具体数据集信息参见 <https://tianchi.aliyun.com/datalab/dataset.htm?id=5>

2. 方法设计

这一节中, 我们考虑对第 1 节提出的问题进行方法设计。

2.1 商品频繁模式挖掘

我们使用 FPGrowth 算法进行商品关联规则挖掘,

该算法...**Todo: 算法 review**

数据...**Todo: 描述使用到的数据**

2.2 同一商家用户的分类

我们要将同一商家的用户分为忠实客户、潜在忠实客户和非忠实客户三类, 首先需要对用户进行特征提取。我们考虑以下特征: **Todo: 给一个 table**

2.3 同一商家常被购买物品

如果不考虑分类就只需要对同一商家所有用户发生四种操作 (要加权, 例如 “购买” 权重最高) 的物品进行计数, 选出最高的几个即可。

2.4 用户重复购买的预测

我们提取了 `user_gender` 等... 个特征...**Todo: 给一个 table**

然后选用几种模型, 使用 `train_format1.csv` 中 80% 的数据训练网络, 另外 20% 的数据进行预测。

Todo: 特征工程

Todo: 算法框架

Todo: 模型、融合

3. 结果与分析

我们所有的实验均在 **Todo: 实验环境** 的主机上进行运行。

- CPU: 2.3GHz Xeon E5 CPU 72 Cores
- RAM: 250GB
- OS: ...
- Kernel: ...

- Python: ...

Todo: 我们得到如下结果

4. 总结

本文我们用... 数据, ... 方法, 分析了... 问题, 得到了... 结果

5. 参考文献

- [1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.
- [2] J. Braams. Babel, a multilingual style-option system for use with latex’s standard document styles. *TUGboat*, 12(2):291–301, June 1991.
- [3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.
- [4] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.
- [5] L. Lamport. *LaTeX User’s Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.
- [6] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.