数据挖掘大作业题目： 在 Expedia 数据集上的旅馆推荐

姓　　名： 李苏畅 魏思杰 李杨 罗伟

摘　要

随着旅馆的普及，人们住宿的选择越来越多。近年来，旅馆的类型逐渐增多，给了用户更多的选择。为了研究用户对旅馆选择的偏好，我们参与了kaggle上的一个利用Expedia数据集的研究项目。

本文主要研究的是如何通过大量用户浏览以及预定旅馆的数据，找到不同用户对旅馆的偏好模式，根据用户的情况进行旅馆的推荐。

本文以python为平台，使用PCA对提供的数据集进行降维，提取出关键的数据特征。然后，对处理后的数据利用GBDT进行回归分析，经拟合后对用户进行旅馆的推荐，进行平台测试后得出准确率为63%。

关键词： 数据挖掘；算法；GBDT；PCA

目　录

第1章 引 言

1.1 选题背景

旅馆产业是一个飞速发展的产业，人们对旅馆的需求呈越来越快的增长趋势。而随着旅馆数量的逐年上升，越来越多的人对旅馆的选择有了不同的标准。这也对旅馆行业对自身的提高提出了进一步的要求。不同类型的旅馆往往是为不同类型的游客准备的。

用户对旅馆的要求包含许多特征：用户是否希望旅馆有游泳池，是否希望每天有服务员打扫房间，是否喜欢购买旅馆提供的物品等。不同的游客，其旅馆选择习惯不可能完全相同，也就导致了对游客推荐的旅馆不可能完全相同。

而随着行业竞争越来越激烈，越来越多的旅馆有了自己的特色，对游客来说，选择合适的旅馆就变成了一个更加复杂的问题，所以，本文通过对大量用户的旅馆预定记录的分析，进行针对用户的旅馆推荐，期望帮助游客住到最合适自己的旅馆。

1.2 数据分析简介

数据分析，就是使用各种各样的统计分析方法来对大量数据进行分析的方法。通过数据分析，可以从大量数据中提取有用的信息，通过对这些信息的分析，可以总结形成可信的结论。通过数据分析所得出的结论，可以指导生活和生产实践，创造更大的价值。

数据分析，从目的方面划分，可分为探索性数据分析与验证性数据分析两种类型。探索性数据分析的目的是从数据中找出新的结论，以便人们对数据的价值进行更大的发掘；而验证性数据分析则是已经假设了一个结论，数据分析的目的是对这个结论进行进一步验证。两种数据分析目的虽然不同，但是分析的手段类似，区别在于探索性数据分析需要从多种角度对数据进行分析，而验证性数据分析则只需针对性地对目的进行分析即可。

数据分析正在人类的生产和生活中起着越来越大的作用。尤其是计算机的发明和广泛应用，让数据分析可以分析的数据量越来越大，分析的速度越来越快，可以提取的信息越来越多。越来越多的个人和企业因为数据分析而获利。

当今社会已经进入“大数据时代”，社会中的数据总量已经达到了惊人的地步，这些数据中也包含着各种各样的信息。在这种现实下，如果合理地使用数据分析方法对各种各样的数据进行分析，就可以挖掘出更多的信息，让它们为人所用。

在本文中，对原始数据进行处理的过程也属于数据分析，同样需要使用各种统计分析方法。通过对游客浏览和预定旅馆的数据进行分析，可以形成游客的预定偏好，概括出游客的喜好模式，便于对游客进行适合的旅馆推荐。

第2章 数据导入和初步处理

2.1 数据集分析

2.1.1 Expedia 数据集简介

Expedia 提供了大量的用户行为数据。包括了用户在搜索什么，他们如何看待搜索结果（点击还是预定），搜索结果是不是一个旅行计划的一部分。数据是随机选择的，并且不具有对整体数据的代表性。

Expdeia 关注的是用户倾向于预定哪种旅馆类型。Expedia 已经通过内部算法构建了旅馆的分类，将价格、用户评价、相距城市中心等因素作为考虑的变量将旅馆进行了聚类。这些旅馆类别将会准确地代表用户将要预测的旅馆类型，这样就避免了有的新旅馆没有历史数据的问题。 比赛的目标就是预测每一个用户将会预定哪一个旅馆类。

训练数据集和测试数据集在时间上是分开的：训练数据集来自 2013 年和 2014 年的数据，而测试数据集来自 2015 年的数据。训练数据包括了日志中的所有用户，包括了点击事件和预定事件，而测试数据集只包括预定事件。

2.1.2 数据内容

数据集包含了丰富的条目。其中仅训练数据就达到了 4.07G 的大小。从图 2-1 中我们可以看到，数据集中含有巨大的信息，所以我们做的第一步就是对数据之间的关联性进行分析。

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tinyint
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Checkin date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Numer of similar events in the context of the same user session	bigint
hotel_cluster	ID of a hotel cluster	int

图 2-1　使用线性插值法进行插值的结果

2.2 数据预处理

由于数据集是随机抽取并保护了用户信息的数据集，所以其中存在大量的数据缺失，由图 2-2 可得，数据项之间的关联程度很低，所以我们为了更准确地进行数据分析，首先对数据进行了清洗，将数据集分成了多个子数据集。个别缺失数据较多的数据集就被放弃了。训练数据集中共涉及 120 万用户，旅馆类别 100 类。旅馆类别明显较少，故我们对每一类的旅馆分别进行了绘图后的数据分析。

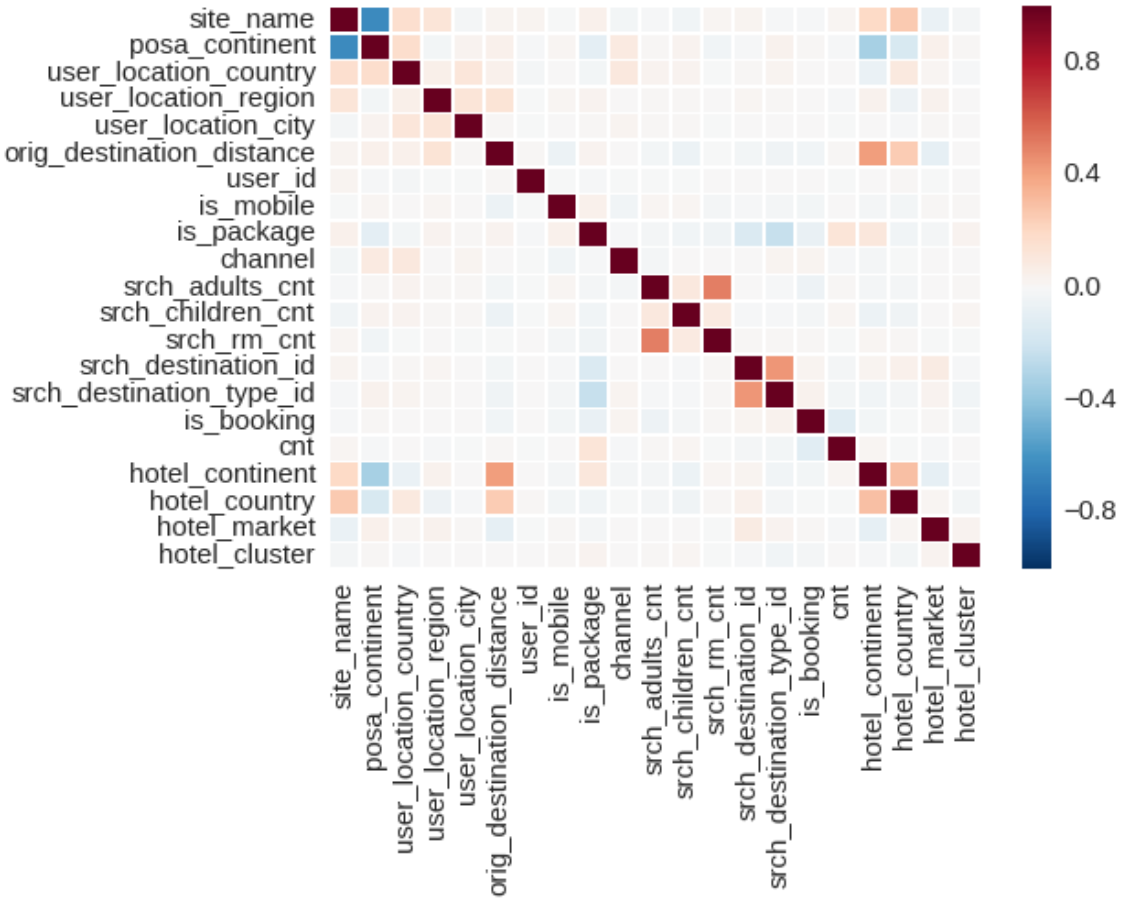

图 2-2　数据之间的相关性

由于数据之间缺少相关性，所以我们采用 PCA 来对输入数据进行降维。

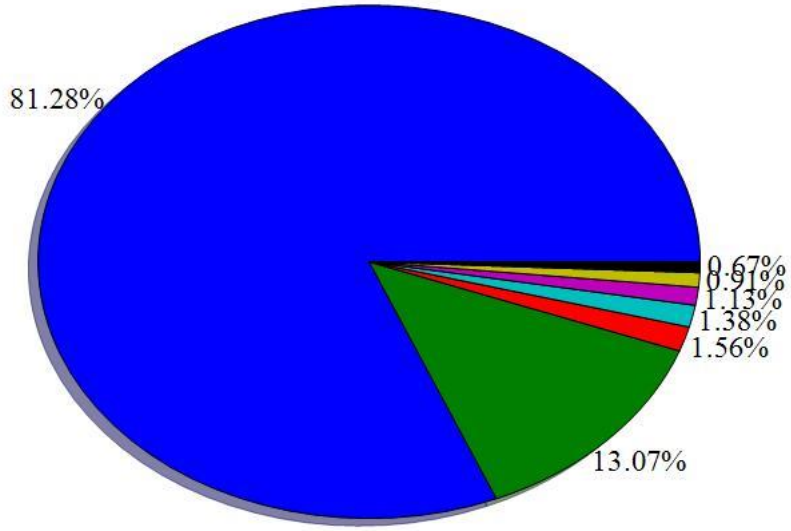

图 2-3　不同出发点的游客占旅馆总人数的比例

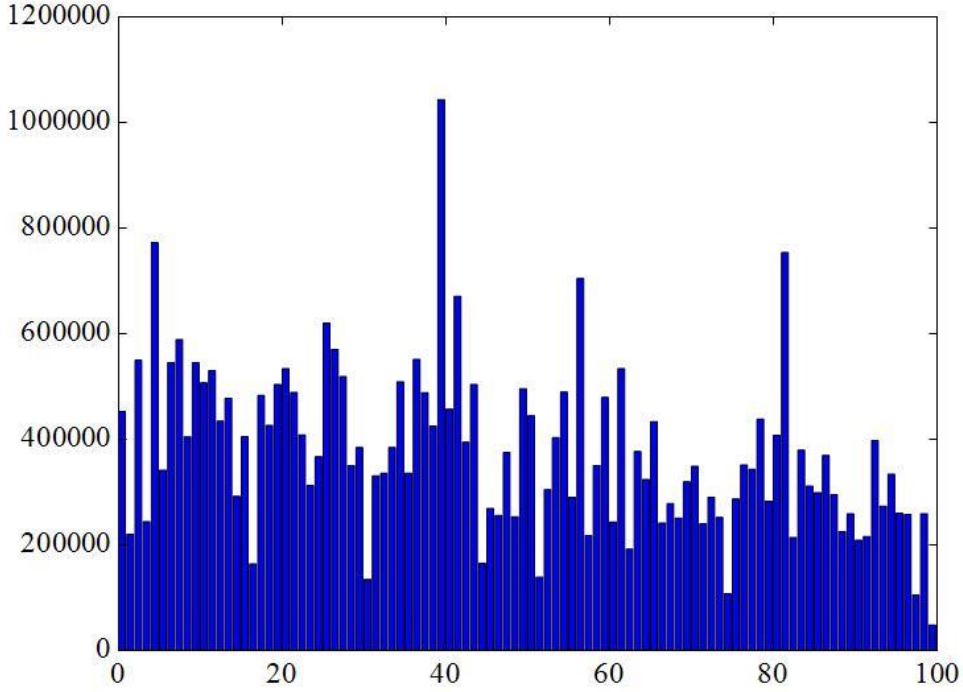

图 2-4　不同旅馆的游客数比较

同样图 3-1 和图 3-2 是对旅馆进行相关数据分析的截图。经过这些完整的分析，我们最终确定了使用 GBDT 对数据进行拟合的方法。

第3章 算法简介

3.1 GBDT 算法简介

GBDT(Gradient Boosting Decision Tree)又叫 MART(Multiple Additive Regression Tree)，是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案。它在被提出之初就和 SVM 一起被认为是泛化能力(generalization)较强的算法。近些年更因为被用于搜索排序的机器学习模型而引起大家关注。

GBDT 几乎可用于所有回归问题(线性/非线性),相对 logistic regression 仅能用于线性回归，GBDT 的适用面非常广。GBDT 亦可用于二分类问题（设定阈值，大于阈值为正例，反之为负例)。

GBDT 主要由三个概念组成：Regression Decision Tree（即 DT),Gradient Boosting（即 GB)，Shrinkage（算法的一个重要演进分枝，目前大部分源码都按该版本实现)。

3.1.1 回归树(Regression Decision Tree)

依据决策树分为两大类，回归树和分类树。前者用于预测实数值，如明天的温度、用户的年龄、网页的相关程度；后者用于分类标签值，如晴天/阴天/雾/雨、用户性别、网页是否是垃圾页面等。前者的结果加减是有意义的，如 10 岁+5 岁-3 岁=12 岁，后者则无意义，如男+男+女没有意义。GBDT 的核心在于累加所有树的结果作为最终结果，就像前面对年龄的累加，而分类树的结果显然是没办法累加的，所以 GBDT 中的树都是回归树，不是分类树。

以对人的性别判别/年龄预测为例,每个 instance 都是一个我们已知性别/年龄的人，而 feature 则包括这个人上网的时长、上网的时段、网购所花的金额等。

C4.5 分类树在每次分枝时，穷举每一个 feature 的每一个阈值，找到使得按照 feature≤阈值和 feature>阈值分成的、两个分枝的熵最大的 feature 和阈值（熵最大的概念可理解成尽可能每个分枝的男女比例都远离 1:1)，按照该标准分枝得到两个新节点,用同样方法继续分枝直到所有人都被分入性别唯一的叶子节点，或达到预设的终止条件。若最终叶子节点中的性别不唯一，

则以多数人的性别作为该叶子节点的性别。

回归树总体流程也是类似，不过在每个节点（不一定是叶子节点）都会得一个预测值，以年龄为例，该预测值等于属于这个节点的所有人年龄的平均值。分枝时穷举每一个 feature 的每个阈值找最好的分割点，但衡量最好的标准不再是最大熵，而是最小化均方差，即$\frac{\sum(\text{每个人的年龄}-\text{预测年龄})^2}{N}$，或者说是$\frac{\text{每个人的预测误差平方和}}{N}$。这很好理解，被预测出错的人数越多，错的越离谱，均方差就越大，因此可以通过最小化均方差来找到最靠谱的分枝依据。分枝直到每个叶子节点上人的年龄都唯一或者达到预设的终止条件（如叶子个数上限），若最终叶子节点上人的年龄不唯一，则以该节点上所有人的平均年龄作为该叶子节点的预测年龄。

3.1.2 梯度迭代(Gradient Boosting)

Boosting，迭代，即通过迭代多棵树来共同决策。GBDT 是把所有树的结论累加起来做最终结论的，所以可以想到每棵树的结论并不是年龄本身，而是年龄的一个累加量。GBDT 的核心就在于，每一棵树学习的是之前所有树结论和的残差，这个残差就是一个加入预测值后可以得到真实值的累加量。举例说明：假如 A 的真实年龄是 18 岁，但第一棵树的预测年龄是 12 岁，即残差为 6 岁。那么在第二棵树里我们把 A 的年龄设为 6 岁去学习。如果第二棵树真的能把 A 分到 6 岁的叶子节点，那累加两棵树的结论就是 A 的真实年龄；如果第二棵树的结论是 5 岁，则 A 仍然存在 1 岁的残差，第三棵树里 A 的年龄就变成 1 岁，然后继续学习。这就是梯度迭代在 GBDT 中的意义。

3.1.3 缩减(Shrinkage)

Shrinkage 的思想认为，每次走一小步逐渐逼近结果的效果，要比每次迈一大步很快逼近结果的方式更容易避免过拟合。即它不完全信任每一个棵残差树，它认为每棵树只学到了真理的一小部分，累加的时候只累加一小部分，然后通过多学几棵树弥补不足。

经验证明，和 Adaboost 一样，Shrinkage 也能减少过拟合的发生。

3.2 PCA 工作原理

对同一个体进行多项观察时，必定涉及多个随机变量 X_1，X_2，…，X_p，它

们都是的相关性，一时难以综合。这时就需要借助主成分分析(principal component analysis)来概括诸多信息的主要方面。我们希望有一个或几个较好的综合指标来概括信息，而且希望综合指标互相独立地各代表某一方面的性质。

任何一个度量指标的好坏除了可靠、真实之外，还必须能充分反映个体间的变异。如果有一项指标，不同个体的取值都大同小异，那么该指标不能用来区分不同的个体。由这一点来看，一项指标在个体间的变异越大越好。因此我们把“变异大”作为“好”的标准来寻求综合指标。

3.2.1 主成分的一般定义

设有随机变量 X_1，X_2，…，X_p，其样本均数记为$\overline{X_1}$，$\overline{X_2}$，…，$\overline{X_p}$，样本标准差记为 S_1，S_2，…，S_p。首先作标准化变换

$$x_i = \frac{X_i - \overline{X_i}}{S_i}$$

我们有如下的定义：

(1)若 $C_1=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p$，$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$，且使 $Var(C_1)$最大，则称 C_1为第一主成分；

(2)若 $C_2=a_{21}X_1+a_{22}X_2+\cdots+a_{2p}X_p$，$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1$，($a_{21}$，$a_{22}$，…，$a_{2p}$)垂直于($a_{11}$，$a_{12}$，…，$a_{1p}$)，且使 $Var(C_2)$最大，则称 C_2为第二主成分；

(3)类似地，可有第三、四、五…主成分，至多有 p 个。

3.2.2 主成分的性质

主成分 C_1，C_2，…，C_p具有如下几个性质：

(1)主成分间互不相关，即对任意 i 和 j，C_i和 C_j的相关系数

$$Corr(C_i,\ C_j)=0 \qquad i \quad j$$

(2)组合系数(a_{i1}，a_{i2}，…，a_{ip})构成的向量为单位向量，$a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$

(3)各主成分的方差是依次递减的，即

$Var(C_1) \geqslant Var(C_2) \geqslant \cdots \geqslant Var(C_p)$

(4)总方差不增不减， 即

$Var(C_1)+Var(C_2)+\cdots+Var(C_p)=Var(x_1)+Var(x_2)+\cdots+Var(x_p)$

这一性质说明,主成分是原变量的线性组合,是对原变量信息的一种改组,主成分不增加总信息量，也不减少总信息量。

(5)主成分和原变量的相关系数 $Corr(C_i, x_j)=a_{ij}=a_{ij}$

(6)令 X_1，X_2，…，X_p 的相关矩阵为 R，$(a_{i1}, a_{i2}, \cdots, a_{ip})$则是相关矩阵 R 的第 i 个特征向量(eigenvector)。而且，特征值λ_i就是第 i 主成分的方差，即

$Var(C_i)=\lambda_i$

其中λ_i为相关矩阵 R 的第 i 个特征值(eigenvalue) $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$。

3.2.3 主成分的数目的选取

前已指出，设有 p 个随机变量，便有 p 个主成分。由于总方差不增不减，C_1，C_2 等前几个综合变量的方差较大，而 C_p，C_{p-1} 等后几个综合变量的方差较小，严格说来，只有前几个综合变量才称得上主（要）成份，后几个综合变量实为“次”（要）成份。实践中总是保留前几个，忽略后几个。

保留多少个主成分取决于保留部分的累积方差在方差总和中所占百分比（即累计贡献率），它标志着前几个主成分概括信息之多寡。实践中，粗略规定一个百分比便可决定保留几个主成分；如果多留一个主成分，累积方差增加无几，便不再多留。

3.2.4 主成分回归

主成分分析本身往往并不是目的，而是达到目的的一种手段。因此，它多用在大型研究项目的某个中间环节。例如，把它用在多重回归中，便产生了主成分回归。另外,它还可以用于聚类、判别分析等。本节主要介绍主成分回归。

在多重回归曾指出，当自变量间高度相关时，某些回归参数的估计值极不稳定，甚至出现有悖常理、难以解释的情形。这时，可先采用主成分分析产生若干主成分，它们必定会将相关性较强的变量综合在同一个主成分中，而不同的主成分又是互相独立的。只要多保留几个主成分，原变量的信息不致过多损失。然后，以这些主成分为自变量进行多重回归就不会再出现共线性的困扰。如果原有 p 个自变量 X_1，X_2，…，X_p，那么，采用全部 p 个主成分所作回归完全等价于直接对原变量的回归;采用一部分主成分所作回归虽不完全等价于对原变量的回归，但往往能摆脱某些虚假信息，而出现较合理的结果。

以上思路也适用于判别分析，当自变量高度相关时，直接作判别分析同样有多重共线性问题，可先计算自变量的主成分，然后通过主成分估计判别函数。

第4章 总结与展望

根据网站给出的返回结果，我们算法的最终得分是 0.50186。虽然预测的效果与最优结果相比还有一定的距离（排名大概 150 名），但是相比于使用的纯统计学方法已经有了长足的提高。未来有机会我们会考虑尝试进一步优化特征提取的结果，并争取通过对参数的改进来提高 GBDT 的回归效果，希望能够获得更好的实验结果。

参考文献

[1]张鹏. 基于主成分分析的综合评价研究[D].南京理工大学,2004.

[2]王涛. 基于 PCA 人脸图像压缩与重建算法的研究与实现[D].昆明理工大学,2014.

[3]陈佩. 主成分分析法研究及其在特征提取中的应用[D].陕西师范大学,2014.

[4]汪爱娟. 基于 PCA 的多变量系统故障检测研究[D].郑州大学,2014.

[5]邹润. 基于模型组合算法的用户个性化推荐研究[D].南京大学,2014.

[6]王辉. 基于核主成分分析特征提取及支持向量机的人脸识别应用研究[D].合肥工业大学,2006.

[7]马云龙. 基于主成分分析的 RBF 神经网络预测算法及其应用[D].吉林大学,2015.

[8]杨胜凯. 基于核主成分分析的特征变换研究[D].浙江大学,2014.

[9]陶思羽. 基于主成分分析和粗糙集的聚类分析在经济指标数据中的应用[D].吉林大学,2012.

[10]陈诺言. 基于个性化推荐引擎组合的推荐系统的设计与实现[D].华南理工大学,2012.

[11]尹轲. 基于群体智能的 PCA 人脸识别算法的优化研究[D].青岛理工大学,2015.

[12]孙万龙. 基于 GBDT 的社区问题标签推荐技术研究[D].哈尔滨工业大学,2015.

[13]李清. 基于 MovieLens 数据集的协同过滤推荐系统研究[D].西安电子科技大学,2014.

[14]宋宇轩. 基于搜索日志和点击日志的同义词挖掘的研究和实现[D].北京交通大学,2011.

[15]曹份槟. 基于 PCA 和 SVM 的货车故障检测[D].北京交通大学,2011.

[16]张书娟. 基于电子商务用户行为的同义词识别[D].哈尔滨工业大学,2011.

[17]赵忠盖. 基于 PCA 统计过程监控的若干问题研究[D].江南大学,2007.

[18]廖贵明. 个性化推荐引擎系统研究[D].电子科技大学,2013.

[19]刘剑波. 基于协同过滤和行为分析的微博推荐系统[D].南京理工大

学,2014.

[20]郝立燕. 协同过滤技术中若干问题的研究[D].华侨大学,2013.