基于随机森林算法的推荐系统的设计与实现

沈晶磊[1,2] 虞慧群[1] 范贵生[1] 郭健美[1]
(华东理工大学计算机科学与工程系 上海 200237)[1] (上海市计算机软件评测重点实验室 上海 201112)[2]

摘 要 如今随着推荐系统势头的加强,如何对用户行为进行快速而准确的预测变得愈加重要。通过分析网上社区帖子的点赞和点踩数据,实现了基于随机森林的推荐系统。该系统将实际问题转化为分类模型,并实现了数据处理、特征提取和参数调整。同时,该系统还对用户浏览帖子后是否产生交互行为进行了预测。最后,通过实验仿真并利用F1值对实验结果进行评估。实验结果证明了系统的有效性和效率。

关键词 随机森林,推荐系统,特征提取,机器学习
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.11.024

Design and Implementation of Recommender System Based on Random Forest Algorithm

SHEN Jing-lei[1,2] YU Hui-qun[1] FAN Gui-sheng[1] GUO Jian-mei[1]
(Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)[1]
(Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai 201112, China)[2]

Abstract As the recommender system is gaining momentum nowadays, how to quickly and accurately predict users' behavior is becoming increasingly important. We implemented a recommendation system based on the random forest algorithm through analyzing the praise and stamp data collected from Internet forum. The system transforms the practical problem into a classification model, and it realizes data processing, feature extraction and parameter tuning. Moreover, the system determines whether there is a further interaction behavior after a post has been viewed. Finally, we conducted experiments to evaluate the system in terms of F1-measure. Our experimental results demonstrate the effectiveness and efficiency of our system.

Keywords Random forest, Recommender system, Feature extraction, Machine learning

1 引言

推荐系统是一种与计算机学科和数据挖掘联系紧密的技术,在我们的日常生活中发挥着愈加重要的作用[1]。随着互联网的发展,如今信息已经处于过载的状态。推荐系统被用于向用户推荐用户可能感兴趣的物品和信息,提高了信息提供商信息推送的效率,也提高了商品提供商的广告效率,使得特定信息能传递给特定的人,从而实现商家与用户的双赢。推荐系统领域的研究现已成为各大商务公司和研究机构的发展趋势,具有非常高的应用价值和研究意义。推荐系统已经成为现今电子商务应用固有的一部分[2]。推荐系统还可以在慢性疾病诊断方面提供准确和可信的预测和用药建议[3],在控制疾病方面发挥着重要作用。

本文分析了随机森林算法在推荐系统中的应用,实验数据来自于中国电信学院(China Telecom College)2016年数据分析与挖掘精英挑战赛[4],实现了原始数据从数据分析、特征工程、数据处理、模型建立、结果预测的过程。本文专注于研究随机森林在实际环境中的应用,结合真实数据,对应用过程进行了改进。

2 相关工作

刘红霞研究和分析了当前推荐系统的主要构成:基于内容的推荐、协同过滤、混合方法[5]。其主要对协同过滤推荐算法进行了分析,并指出了其存在的挑战。但是,由于单一的方法往往并不能很好地解决问题,因此更多时候需要数据挖掘、机器学习等多种方法混合使用。

推荐系统中算法的改进主要有:高卫华[6]使用马尔科夫模型来预测用户行为,有效提高了网站的服务质量,最大限度地留住客户。但是该模型的指数型复杂度大大影响了实际应用,而现在面对大数据,寻找一种方便且快捷的算法是非常关键的。Gupta等[2]提出基于层次聚类的算法来减小误差,并将该算法与K-means算法做了对比实验以证明该算法的有

到稿日期:2016-10-18 返修日期:2016-12-01 本文受国家自然科学基金(61602175,61772200),上海市研究生教育创新项目,华东理工大学教育教学规律与方法研究项目资助。

沈晶磊(1992-),男,硕士生,主要研究方向为数据挖掘、软件工程;虞慧群(1967-),男,教授,博士生导师,CCF高级会员,主要研究方向为软件工程、形式化方法,E-mail:yhq@ecust.edu.cn(通信作者);范贵生(1980-),男,副研究员,CCF会员,主要研究方向为软件工程、可信计算;郭健美(1981-),男,副教授,CCF会员,主要研究方向为软件工程、人工智能。

效性。Mustansar 和 Adam[7] 提出了一种级联混合的推荐方法。李欣海[8] 分析了随机森林模型在分类和回归中的应用，并认为随机森林算法通过大量分类树的汇总提高了预测精度，是取代神经网络等传统机器学习方法的新模型，其运算速度快的特点在处理大数据时表现优异。

在实际场景应用方面，Fang 等[9] 利用 Tmail. com 的真实数据分析了用户的在线行为，并提出了一种基于情景的分类方法将用户分为两组：一组是有明显的购买意图的，而另一组没有。当推荐系统使用了这种用户分类特征后，能在预测性能上有显著提高。但这只是用户诸多特征中的一种，若要使得推荐系统能有更好的效果，就需要更深入地挖掘其他用户及商品的特征。刘宁等[10] 基于智能交通数据使用 Elman 神经网络算法设计并实现了实时交通系统中的流量预测，该方法能有效地对交通流量进行预测，但不能满足实时需求。王宇恒[11] 在推荐系统中应用了随机森林算法，分析了该算法在推荐系统中精度高、训练过程快速等优势，还针对特征选择和不平衡分类等问题提出了解决方案，并利用电商购物网站数据进行了实验。该系统对电商购物数据比较适用，而对于网络社区赞踩数据则需要重新设计和实现预测流程。

3　基于随机森林算法的推荐系统设计

3.1　随机森林算法

随机森林（random forest）是一种基于分类树（classification tree）的算法[12]，该算法需要模拟和迭代，被归类为机器学习方法中的一种。分类树算法通过反复二分数据来进行分类或回归，但是会产生过拟合等问题。随机森林是一种集成学习方法，通过产生多个分类树来生成结果，即在特征的选取和数据的选取上进行随机化，生成许多分类树，再汇总分类树的结果。随机森林在复杂度没有显著提高的情况下，提高了预测精度，且对多元线性不敏感，因此对缺失数据和非平衡数据比较稳健。

随机森林是以 K 个决策树$\{h(X,\Theta k),k=1,2,\cdots,K\}$为基本分类单元，进行集成学习后得到的一个组合分类器。随机森林通过随机选择样本和随机选择特征子集来生成大量的树，称之为“随机森林”[13]，如图 1 所示。

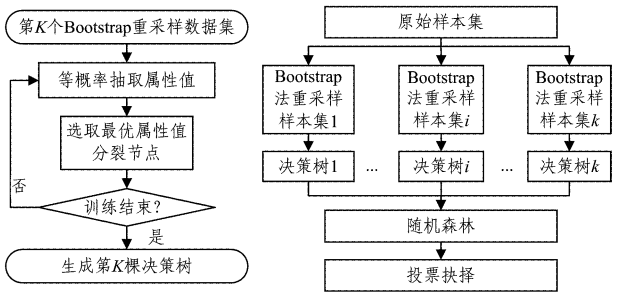

图 1　决策树以及随机森林的生成

在得到随机森林之后，当有一个新的测试样本进入随机森林时，其实就是让每一棵决策树分别进行投票抉择，最终取所有决策树中输出类别最多的那类为分类结果。最终的分类决策可用式(1)表示：

$$H(x)=\arg\max_{Y}\sum_{i}^{k}I(h_i(x)=Y) \tag{1}$$

其中，$H(x)$表示分类组合模型，h_i 是单个决策树分类模型，$I(\cdot)$为示性函数（示性函数是指一个函数使得当集合内有此数时值为 1，当集合内无此数时值为 0），Y 表示目标变量（或称输出变量）。

3.2　问题描述

本文的实验数据提供了“中华万年历”部分用户在 2015 年 11 月前 27 天的帖子浏览记录、交互行为及帖子内容，目标是预测 11 月 28 号到 30 号 3 天用户在 APP 内的点赞、点踩等交互行为。

数据包括一系列网上社区的帖子中被浏览和点击的事件。其中一些用户在浏览或点击详情后会评价事件，评价包括点赞或点踩。问题的目标是预测一次浏览或点击事件是否会产生评价行为，如果产生，具体是点赞还是点踩？点赞意味着用户对该帖子感兴趣并支持其观点，点踩则意味着用户对该帖子感兴趣并反对其观点。这些信息在广告系统中很有价值，信息提供商能给用户推送其感兴趣的文章或是产品。对实验数据进行数据清洗、问题分析、特征提取与建模，最后通过随机森林算法进行结果的预测。

数据共有 3 种。数据 1：用户从 11 月 1 日至 30 日的浏览记录，主要记录项包括用户 id、帖子 id、记录时间、浏览次数和点击次数。数据 2：用户从 11 月 1 日至 27 日的交互行为，主要记录项包括用户 id、帖子 id、记录时间、点赞或是点踩。数据 3：11 月 1 日至 30 日的帖子内容，主要记录项包括帖子 id、帖子发表时间、帖子内容。而目的是根据这些数据预测用户在 11 月 28 日至 30 日是否对帖子有交互行为，如果有，则需指出是点赞还是点踩。

3.3　系统设计

针对该问题，可从数据分析、特征提取、数据处理、模型建立、结果预测几个阶段给出解决方案，具体过程如图 2 所示。

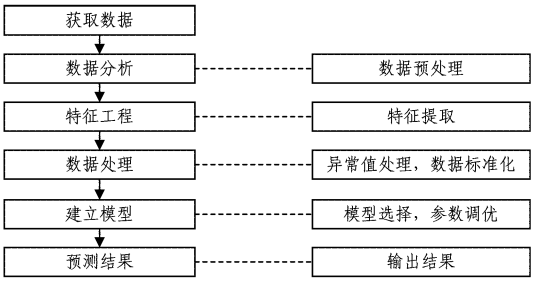

图 2　数据挖掘解决推荐系统问题的主要流程

3.3.1　评价指标

衡量分类模型的常用指标是准确率和召回率。准确率的定义是 $P=TP/(TP+FP)$，是在分类结果中对正类的正确分类的衡量。而召回率 $R=TP/(TP+FN)$，是用来衡量正确分到正类数据占所有正类的指标。但是，单独使用这两个指标进行评价是不够全面的。实际上，两者的调和平均数 $F1=2RP/(R+P)$能较好地反映模型的好坏。

3.3.2　数据分析

(1)数据预处理

由于原始数据中存在大量的脏数据，因此在之后的分析前必须对数据进行预处理。在数据分析中发现，其存在多次

踩赞，需要去重。对于未浏览踩赞等交互记录，需要将这些数据删除。

(2)数据统计分析

数据中共有 9000 多万条浏览记录，47 万条发生交互的踩赞记录，其中用户有 11 万，帖子有 8 万条。

通过分析可以发现，绝大部分用户对浏览过的帖子几乎无交互行为，即他们只会浏览帖子但不会对帖子进行点赞或点踩。因此可以不分析这部分用户的数据，大大减少数据量，并提高精度。这样只保留交互率(赞踩占浏览的比例)大于 0.12%的用户，共 7000 多个。对其分析后发现，用户呈现两种类别，一种是点赞较多，另一种是点踩较多，图 3 给出了用户根据点赞占比进行的分类(赞的数量占交互数量的比例)。

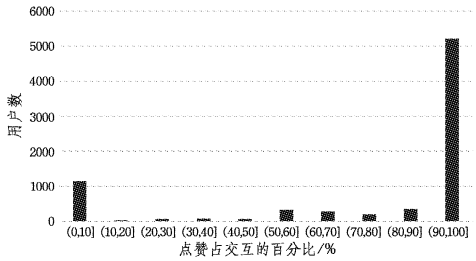

图 3 根据点赞占比分类的用户人数直方图

3.3.3 特征工程

(1)特征提取

在认清数据特征的基础上，根据浏览、交互、时间等因素提取可能有用的特征。所谓特征提取就是将原始特征转换为一组具有明显物理意义或统计意义的特征，从而发现更有意义的潜在的变量，帮助人们对数据产生更深入的了解。

(2)用户特征

选取的用户特征有用户点赞帖子总数、点踩帖子总数、交互帖子总数、浏览帖子总数、点击帖子总数、用户对帖子的点赞率、用户对帖子的点踩率、用户对帖子的交互率、用户对帖子的平均浏览数、用户对帖子的平均点击数、用户的行为特征、用户赞踩数量比、用户浏览点击转化率。

设用户集合为 A，帖子集合为 B，用户对帖子浏览的集合 $C\subseteq A\times B\times N\times T$，其中 N 为次数，T 为时间，用户对帖子点击的集合 $D\subseteq A\times B\times N\times T$，用户对帖子交互的集合 $E\subseteq A\times B\times\{1,2\}\times T$，其中 1 表示点赞，2 表示点踩。

用户 $a_i\in A$ 的点赞帖子总数为 $|F|$，$F=\{x|x\in B,(a_i,x,1)\in E\}$。

用户 $a_i\in A$ 的点踩帖子总数为 $|G|$，$G=\{x|x\in B,(a_i,x,2)\in E\}$。

用户 $a_i\in A$ 的浏览帖子总数为 $|H|$，$H=\{x|x\in B,(a_i,x)\in C\}$。

用户对帖子的点赞率为 $\alpha=|F|/|H|$，用户对帖子的点踩率为 $\beta=|G|/|H|$，用户对帖子的交互率为 $\alpha+\beta$。用户的行为特征的计算公式为：

$$F(x)=\begin{cases}0, & \alpha+\beta<0.012\\ 1, & \alpha+\beta\geqslant 0.012,\alpha<\beta\\ 2, & \alpha+\beta\geqslant 0.012,\alpha>\beta\end{cases}\qquad(2)$$

用户 $a_i\in A$ 的点击帖子总数为 $|K|$，$H=\{x|x\in B,(a_i,x)\in D\}$。

用户浏览点击转化率为 $\gamma=|K|/|H|$。

(3)交互特征

选取的交互特征有用户对该帖子的浏览总数、用户对该帖子的点击总数、用户首次对该帖子的浏览数、用户首次对该帖子的点击数、用户首次看到该帖子的时间特征，其中时间特征为是否在周末以及节假日。

(4)帖子特征

选取的帖子特征有帖子长度等级(1:1～50 个字，2:51～200 个字，3:201～800 个字，4:800 个字以上)、帖子是否包含标题、帖子的主题类型。

3.3.4 模型建立

该问题是一个三分类问题，即最终的分类结果有：无交互、点赞、点踩。可以将此问题分成两个二分类问题，即模型 1 先判断用户对帖子有无交互行为，再将有交互行为的结果输入模型 2 来判断其是点赞还是点踩。问题建模如图 4 所示。

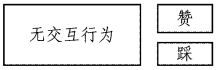

图 4 问题建模

4 实验结果及分析

4.1 代码实现

实验平台采用的是 Canopy，它是 Python 的集成开发环境，其上集成了超过 450 种科学计算的 Python 包，实验使用到了 scikit-learning 包。其中有很多机器学习的工具函数，能有效地减少实验实现的难度。实验采用了 Win10 系统，Core i3 处理器 2.27G 主频，4GB 内存的实验环境。

实验将 11 月 1 日至 27 日的数据作为训练集，将 28 日至 30 日的数据作为测试集，由于推荐系统中存在类不平衡问题，因此需要在实验室采用欠抽样方法来控制正样本和负样本的比例。

实验的主要代码如下。

创建随机森林分类器，其中基分类器的个数为 20：

```
from sklearn.ensemble import RandomForestClassifier
estimator = RandomForestClassifier (n_estimators = 20, min_samples_leaf=10,max_features='auto')
```

对训练集进行训练：

```
estimator.fit(X,y.ravel())
```

对测试集进行预测：

```
y_pred=estimator.predict(X_test)
```

得到训练集上交叉验证的准确率：

```
from sklearn.cross_validation import cross_val_score
score=cross_val_score(estimator,X,y.ravel()).mean()
```

结果集的 F1 值、准确率、召回率的代码如下：

```
from sklearn.metrics import precision_score,recall_score,f1_score
print 'F1 score:',f1_score(y_true,y_pred)
print 'Recall:',recall_score(y_true,y_pred)
print 'Precision:',precision_score(y_true,y_pred )
```

4.2　实验结果

随机森林中决策树的构建是模型建立的核心。决策树个数的多少直接影响着随机森林分类算法的运算速度和分类效果，因此决策树的个数对建模至关重要[13]。如图 5 所示，随着随机森林中决策树棵数的增加，模型交叉验证的准确率是呈上升趋势的；但是当棵数大于 10 时，准确率随决策树棵数增加而上升的趋势已经不明显。如图 6 所示，随着随机森林中决策树棵数的增加，模型的 F1 值是呈上升趋势的；但是当棵数大于 10 时，F1 值随决策树棵数增加而上升的趋势已经不明显。如图 7 所示，随着随机森林中决策树棵数的增加，模型的准确率上升，但召回率变化不显著。

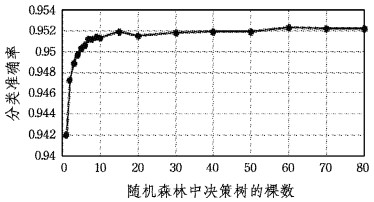

图 5　决策树棵数对交叉验证准确率的影响

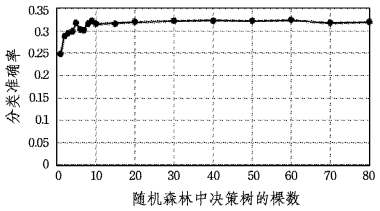

图 6　决策树棵数对 F1 值的影响

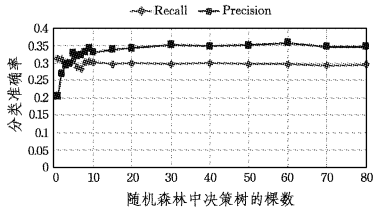

图 7　决策树棵数对准确率召回率的影响

从图 5－图 7 中可以看出，对该数据集而言，当决策树的棵数达到 10 时，随着决策树棵数的增加，随机森林的分类正确率并未持续升高，反而会因为棵数增多而带来时间开销的增加。综合考虑随机森林中包含的决策树的棵树与建模的速度，选取随机森林分类算法中包含 10 棵决策树是比较理想的，这使得推荐系统能快速得出准确的结果，满足推荐系统对时间限制的需要。

为了更好地体现该模型的优点，对于上述样本数据，用 K 近邻和支持向量机进行预测，实验结果如表 1 所列。实验结果表明，使用随机森林算法将有更好的结果，使得基于随机森林算法的推荐系统有更好的推荐精确度。图 8 详细给出了不同近邻数目下的 KNN 算法对分类结果的影响。

表 1　随机森林与其他机器学习方法的比较

method	parameters	cross_val	F1-measure
KNN	n=5	0.850775423	0.152138426
SVM	kernel='rbf'	0.803593457	0.154211257
RF	n=10	0.951236027	0.315708812

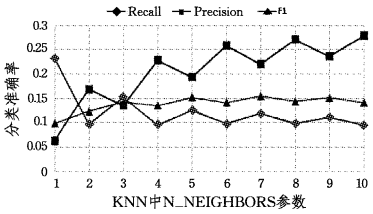

图 8　KNN 算法中 n_neighbors 参数对分类结果的影响

结束语　如今推荐系统在人们的生活中扮演着越来越重要的角色，如何提高推荐的准确度成为现代各大商务公司和研究机构的研究热点。本文介绍了对网上社区帖子交互数据的预测应如何构建随机森林预测模型，如何分析数据及提取特征，并在预测过程中对模型参数进行调优，从而验证方法的有效性。当然，对于随机森林在推荐系统中的应用还有许多可以深入研究的地方，如决策树之间的冗余问题、类不平衡问题的解决等。

参 考 文 献

[1] ASHLEY-DEJO E, NGWIRA S, ZUVA T. A survey of Context-aware Recommender System and services[C]// 2015 International Conference on Computing, Communication and Security (ICCCS). IEEE, 2015: 1-6.

[2] GUPTA U, PATIL N. Recommender system based on Hierarchical Clustering algorithm Chameleon[C]// 2015 IEEE International Advance Computing Conference (IACC). IEEE, 2015: 1006-1010.

[3] HUSSEIN A S, OMAR W M, LI X, et al. Efficient chronic disease diagnosis prediction and recommendation system[C]// 2012 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, 2012: 209-214.

[4] China telecom institute_data mining competition data [EB/OL]. http://yun.baidu.com/share/link?shareid=3742128086&uk=453752322.

[5] LIU H X. A Survey of Collaborative Filtering Technique in Recommendation System[J]. Information Security and Technology, 2016, 7(3): 24-26. (in Chinese)
刘红霞. 基于协同过滤技术的推荐系统综述[J]. 信息安全与技术, 2016, 7(3): 24-26.

[6] GAO W H, XIE K L. New Model and Related Algorithm for the Prediction of Web User's Directions[J]. Computer Applications and Software, 2007, 24(3): 142-144. (in Chinese)
高卫华，谢康林. Web 用户行为预测的一种新模型及算法[J]. 计算机应用与软件, 2007, 24(3): 142-144.

[7] GHAZANFAR M A, PRUGEL-BENNETT A. A scalable, accurate hybrid recommender system[C]// Third International Conference on Knowledge Discovery and Data Mining, 2010 (WKDD'10). IEEE, 2010: 94-98.

[8] LI X H. Using "Random Forest" for Classification and Regression [J]. Chinese Journal of Applied Entomology, 2013, 50(4): 1190-1197. (in Chinese)
李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4): 1190-1197.

（下转第 186 页）

合发表评论信息。而用户“夕阳”更侧重于针对其它类特征进行评论，虽然其表达方式的词性组合比较多样化，但大多使用“{v+[d]+a|n}”的词性组合，其情感倾向程度波动比较大。

通过实验一和实验二可以看出，本文提出的用户评论模式分析方法是有效的，可以较为准确地分析出用户评论的特征类别、词性组合以及情感倾向程度。

结束语 本文提出了一种 APP 软件的用户评论模式分析方法，首先通过分析用户评论信息和 APP 软件信息之间的关系，将用户评论信息分为 3 类；然后分析每类用户的评论信息的词性组合；最后量化每条用户评论信息的情感倾向程度，以分析出 APP 软件的用户评论模式。通过实验证明了本文方法的有效性。但该方法也存在不足，即当用户对 APP 软件的评论特征类别比较分散时只做了简单的频率计算，并没有做深入分析。因此，下一步将对评论特征类别的分析、网络情感词汇的收集以及 APP 软件的用户可信度计算等方面进行深入研究。

参 考 文 献

[1] JIANG H, MA H, REN Z, et al. What makes a good app description? [C]// Proceedings of the 6th Asia-Pacific Symposium on Internetware on Internetware. ACM, 2014: 45-53.

[2] MA S, WANG S, LO D, et al. Active Semi-Supervised Approach for Checking App Behavior Against Its Description[C]// 2015 IEEE 39th Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2015: 179-184.

[3] LI M, WANG X C, ZHANG J, et al. Study on Check-in and Related Behaviors of Location-based Social Network Users [J]. Computer Science, 2013, 40(10): 72-76. (in Chinese)
李敏，王晓聪，张军，等. 基于位置的社交网络用户签到及相关行为研究[J]. 计算机科学，2013，40(10)：72-76.

[4] WANG W, WANG H W, MENG Y. The collaborative filtering recommendation based on sentiment analysis of online reviews [J]. Systems Engineering-Theory & Practice, 2014, 34(12): 3238-3249. (in Chinese)
王伟，王洪伟，孟园. 协同过滤推荐算法研究：考虑在线评论情感倾向[J]. 系统工程理论与实践，2014，34(12)：3238-3249.

[5] ZHANG L, QIAN G Q, FAN W G, et al. Sentiment Analysis Based on Light Reviews [J]. Journal of Software, 2014, 25(12): 2790-2807. (in Chinese)
张林，钱冠群，樊卫国，等. 轻型评论的情感分析研究[J]. 软件学报，2014，25(12)：2790-2807.

[6] LIU Y, LIAO X W, LIU Y. The impact of online review on software and platform's pricing strategies [J]. Journal of Systems Engineering, 2014, 29(4): 560-570. (in Chinese)
刘洋，廖貅武，刘莹. 在线评论对应用软件及平台定价策略的影响[J]. 系统工程学报，2014，29(4)：560-570.

[7] PAGANO D, MAALEJ W. User feedback in the appstore: An empirical study[C]// International Conference on Requirements Engineering. 2013: 125-134.

[8] GUZMAN E, MAALEJ W. How do users like this feature? a fine grained sentiment analysis of app reviews[C]// 2014 IEEE 22nd International Requirements Engineering Conference (RE). IEEE, 2014: 153-162.

[9] ANAM A S M I, YEASIN M. Accessibility in smartphone applications: what do we learn from reviews? [C]// Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, 2013: 35.

[10] KHALID H. On identifying user complaints of iOS apps[C]// International Conference on Software Engineering. 2013: 1474-1476.

[11] HU Z K, ZHENG X L, WU Y F, et al. Product recommendation algorithm based on users' reviews mining [J]. Journal of Zhejiang University (Engineering Science), 2013, 47(8): 1475-1485. (in Chinese)
扈中凯，郑小林，吴亚峰，等. 基于用户评论挖掘的产品推荐算法[J]. 浙江大学学报(工学版)，2013，47(8)：1475-1485.

[12] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions, and reversals [J]. Soviet Physics Doklady, 1966, 10(8): 707-710.

[13] LI X N, ZHANG S W, YANG L, et al. ARES: Autoregressive Emotion-Sensitive Model for Predicting Sales Performance [J]. Journal of Computer Research and Development, 2013, 50(8): 1722-1727. (in Chinese)
李雪妮，张绍武，杨亮，等. ARES：用于预测的情感感知自回归模型[J]. 计算机研究与发展，2013，50(8)：1722-1727.

[14] LIN Q H. Design and Implementation of the Product Reviews Analysis System Based on Affective Computing [D]. Shanghai: Fudan University, 2013. (in Chinese)
林钦和. 基于情感计算的商品评价分析系统设计与实现[D]. 上海：复旦大学，2013.

[15] RAN M, JIANG Y, XIANG Q, et al. Method of Consistency Judgment for App Software's User Comments[C]// International Conference of Young Computer Scientists, Engineers and Educators. Springer Singapore, 2016: 470-483.

(上接第 167 页)

[9] FANG K, ZHANG Q, ZHUANG Z, et al. Making Recommendations Better: The Role of User Online Purchase Intention Identification[C]// 2016 International Conference on Software Networking (ICSN). IEEE, 2016: 1-4.

[10] LIU N, CHEN Y T, YU H Q, et al. Traffic Flow Forecasting Method Based on Elman Neural Network [J]. Journal of East China University of Science and Technology, 2011, 37(2): 204-209. (in Chinese)
刘宁，陈昱颋，虞慧群，等. 基于 Elman 神经网络的交通流量预测方法[J]. 华东理工大学学报，2011，37(2)：204-209.

[11] WANG Y H. Improvement and Application of Random Forest Algorithm in recommender systems [D]. Hangzhou: Zhejiang University, 2016. (in Chinese)
王宇恒. 推荐系统中随机森林算法的优化与应用[D]. 杭州：浙江大学，2016.

[12] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[13] ZHANG Y, GAO Q Q. Water quality evaluation of Chaohu Lake based on random forest method [J]. Chinese Journal of Environmental Engineering, 2016, 10(2): 992-998. (in Chinese)
张颖，高倩倩. 基于随机森林分类算法的巢湖水质评价[J]. 环境工程学报，2016，10(2)：992-998.