文本复制检测报告单(全文标明引文)

№:ADBD2018R_2018053015312720180530154845440174234834　　检测时间:2018-05-30 15:48:45

检测文献:　53141119_盛傢伟_计算机科学与技术（网络与信息安全）_基于文本处理技术的App分类系统的设计与实现

作者:　盛傢伟

检测范围:　中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

图书资源

优先出版文献库

大学生论文联合比对库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

互联网文档资源

CNKI大成编客-原创作品库

个人比对库

时间范围:　1900-01-01至2018-05-30

可能已提前检测，检测时间：2018/5/19 9:29:28，检测结果：4.1%

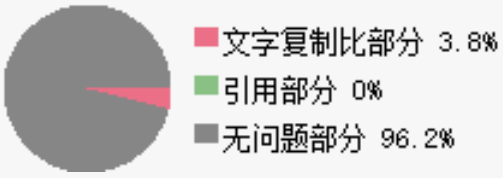
检测结果

总文字复制比：3.8%　　跨语言检测结果：0%

去除引用文献复制比：3.8%　　去除本人已发表文献复制比：3.8%

单篇最大文字复制比：1.2%（用深度学习解决大规模文本分类问题 - starzhou的专栏 - CSDN博客）

重复字数：	[1296]	总段落数：	[8]
总字数：	[33685]	疑似段落数：	[4]
单篇最大重复字数：	[393]	前部重合字数：	[886]
疑似段落最大重合字数：	[716]	后部重合字数：	[410]
疑似段落最小重合字数：	[55]		



指　标：☐疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表格：0　　公式：0　　疑似文字的图片：0　　脚注与尾注：0

0%（0）	中英文摘要等（总2673字）
18%（440）	第1章绪论（总2441字）
24.8%（716）	第2章相关工作（总2889字）
0.5%（55）	第3章分类体系的构建与评价_第1部分（总10444字）
0%（0）	第3章分类体系的构建与评价_第2部分（总8130字）
0%（0）	第4章分类算法的训练与评价（总5246字）
0%（0）	第5章分类系统原型的设计与实现（总815字）
8.1%（85）	第6章总结与展望（总1047字）

（注释：无问题部分　文字复制比部分　引用部分）

1. 中英文摘要等　　总字数：2673

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

吉林大学学士学位论文（设计）承诺书

本人郑重承诺：所呈交的学士学位毕业论文（设计），是本人在指导教师的指导下，独立进行实验、设计、调研等工作基础上取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写的作品成果。对本人实验或设计中做出重要贡献的个人或集体，均已在文中以明确的方式注明。本人完全意识到本承诺书的法律结果由本人承担。

学士学位论文（设计）作者签名：

2018年5月20日

摘要

基于文本处理技术的App分类系统的设计与实现

近年来随着移动设备的广泛使用，已经有大量的为移动用户开发的移动应用程序（App）。移动应用程序在日常生活中扮演着重要角色，研究移动应用程序的使用可以帮助了解用户的偏好，从而激发许多智能个性化服务，如应用推荐，用户分组和广告投放。基于应用商店的App分类体系和App分类标注数据，通过使用文本处理技术可以实现App的自动分类。然而存在以下问题：由于App名称和包名本身携带的信息很少，难以仅使用名称和包名对App进行分类；没有公认的App分类体系，现有的分类体系可能存在类别不合理问题；App的类别由上传者手工标注，可能存在错误标注的样本。这些问题给App分类带来了困难。

本文使用Google Play中App的描述文本来扩展App的文本数据，并使用Google Play中的App分类体系作为原始分类体系，主要内容有：

（1）构建和评价App分类体系：通过对中文文本描述进行文本聚类得到符合文本相似关系的聚类结果，结合聚类结果和原始分类体系，手工构建新的分类体系，并通过实验验证新的分类体系的合理性。根据文本相似度设计算法将原始标注的App标注到新的类别中。

（2）训练和验证App分类算法:在带有类别标注的App文本描述数据集上，提取文本特征，构建多种文本分类器，并通过对比实验分别验证各个分类器的性能。其中，使用向量空间模型的SVM分类器在数据集上达到了最佳的分类性能。

在上述实验的基础上，设计和实现基于文本处理技术的App分类系统原型，用以解决App分类问题。

关键字： App分类，文本聚类，文本分类

Abstract

Design and Implementation of App Classification System Based on Text Processing Technology

In recent years, there have been a large number of mobile Applications (Apps) developed for mobile users. Researching the use of mobile Applications can help understand user preferences and provide many smart and personalized services, such as Application recommendations, user grouping, and ad placement. Using text processing technology can achieve automatic classification of Apps. However, there are problems in App classification: It is difficult to classify Apps because of the lack of App text information; there are not a widely accepted App taxonomy; App categories are labeled by the uploader manually, and there may be incorrectly labeled Apps.

We use the description text of the App in Google Play to extend the data of Apps, and use the categories of App in Google Play as the original App taxonomy. The main contents in this article are:

(1) Build and evaluate the App taxonomy: Using text clustering of Chinese text descriptions, we obtain clusters according to the similarity of texts. We combine the clusters with the original App taxonomy, and construct a new taxonomy manually, then evaluate the taxonomy with experiments. According to the text similarity, we design an algorithm to label the Apps automatically.

(2) Training and evaluate App classifier: We design three methods to extract text features on the labeled text description data set and build a lot of text classifiers to verify the performance of classifiers by comparing experiment results. Among the classifiers, the SVM classifier using the vector space model achieved the best classification performance on the data set.

Based on the above experiments, a prototype of an App classification system based on text processing technology is designed and implemented to solve the problem of App classification.

Keywords: App classification, text clustering, text classification

目录

2. 第1章绪论

总字数：2441

相似文献列表　文字复制比：18%(440)　疑似剽窃观点：(0)

1	文本分类综述 靳小波; -《自动化博览》- 2006-12-30	14.5%（354） 是否引证：否
2	基于AdaBoost实现文本数据分类 聂焕君 -《大学生论文联合比对库》- 2017-05-31	13.0%（317） 是否引证：否
3	文本分类 - lionzl的专栏 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2013	13.0%（317） 是否引证：否
4	文本分类概述 - bluenight专栏 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	12.7%（310） 是否引证：否
5	交叉覆盖算法下文本分类的研究 李家兵(导师：张燕平) -《安徽大学硕士论文》- 2007-04-01	12.0%（294） 是否引证：否
6	基于图模型的Web文档分类方法研究 张炼(导师：孟海东) -《内蒙古科技大学博士论文》- 2010-06-30	11.7%（285） 是否引证：否
7	KNN算法的改进及其在文本分类中的应用 卜凡军(导师：钱雪忠) -《江南大学硕士论文》- 2009-08-01	11.7%（285） 是否引证：否
8	基于科技文献的中文文本分类算法研究 王俊英(导师：郭景峰) -《燕山大学硕士论文》- 2007-03-01	11.6%（283） 是否引证：否
9	网络舆情分析关键技术研究 刘泽光(导师：刘辉林) -《东北大学博士论文》- 2013-06-01	10.3%（252） 是否引证：否
10	汉语主客观文本分类及预处理研究 - 豆丁网 -《互联网文档资源（http://www.docin.com）》- 2017	9.8%（240） 是否引证：否
11	汉语主客观文本分类及预处理研究 张霄凯(导师：姚天昉) -《上海交通大学硕士论文》- 2009-01-01	9.8%（240） 是否引证：否
12	基于最优分割策略的高性能文本分类方法 万狄飞(导师：王国胤;樊兴华) -《重庆邮电大学硕士论文》- 2008-05-01	9.7%（237） 是否引证：否
13	基于朴素贝叶斯和BP神经网络的中文文本分类问题研究 王雅玢(导师：夏幼明) -《云南师范大学硕士论文》- 2008-05-28	8.7%（213） 是否引证：否
14	KNN算法的改进及其在文本分类中的应用 - 豆丁网 -《互联网文档资源（http://www.docin.com）》- 2013	8.1%（198） 是否引证：否
15	基于新浪微博数据的用户情感分析系统的设计与实现 刘志嘉 -《大学生论文联合比对库》- 2016-05-17	7.4%（180） 是否引证：否
16	一种潜在语义索引差异模型及其应用 米晓芳(导师：宋宜斌;王立宏) -《烟台大学硕士论文》- 2007-03-31	6.1%（149） 是否引证：否
17	中文短文本分类的相关技术研究 崔争艳(导师：胡小华) -《河南大学博士论文》- 2011-05-01	6.0%（146） 是否引证：否
18	短信自动分类技术研究与应用 李继刚(导师：李锋) -《东华大学博士论文》- 2012-11-01	6.0%（146） 是否引证：否
19	文本内容智能化划分系统设计与实现 李凌 -《大学生论文联合比对库》- 2015-06-05	5.8%（141） 是否引证：否
20	基于神经网络的文本自动分类系统研究 王志玲(导师：王效岳) -《山东理工大学硕士论文》- 2007-04-10	5.4%（131） 是否引证：否
21	基于神经网络的文本挖掘在专利自动分类中的研究与应用 马芳(导师：王效岳) -《山东理工大学硕士论文》- 2009-04-16	5.4%（131） 是否引证：否
22	文本分类中基于TFIDF的特征选择算法研究 杨晨 -《大学生论文联合比对库》- 2017-05-25	4.1%（100） 是否引证：否
23	基于机器学习的文本处理技术研究与应用 薛松(导师：景晓军) -《北京邮电大学博士论文》- 2015-03-09	3.9%（94） 是否引证：否
24	学位论文预审分配管理系统研究 吕斐斐(导师：钱国明) -《哈尔滨工业大学博士论文》- 2010-06-01	1.3%（31） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第1章绪论

1.1 研究背景与意义

九十年代以来，互联网迅速发展，逐渐走进人们的日常生活中来。互联网中包含了海量的信息，其中信息的形式包括文本、声音和图像等。文本信息是网络中信息的主要的传递形式，与其它形式相比，它具有更强的表达能力，且更易于信息的上传和下载。基于机器学习方法的文本分类系统，可以在给定的分类模型下，根据内容自动对文本进行类别划分，从而有助于人们进行文本数据的挖掘，因此，成为信息处理领域中重要的研究方向。

另一方面，随着近年来移动设备的广泛使用，已经有大量的为移动用户开发的移动应用程序。据相关数据统计，2017年的Google Play下载量达到了640亿次，相比2016年增长了16.7%，而App Store的下载量则是280亿次，相比2016年增加了6.7%。移动应用程序在日常生活中扮演着重要角色，研究移动应用的使用可以帮助了解用户的偏好，从而激发许多智能个性化服务，如应用推荐，用户分组和广告投放。

使用文本处理技术可以实现App的分类系统。然而，由于App名字信息很短，得到的信息十分稀疏，所以需要通过扩展App的文本数据，来提高分类系统的性能。本课题使用扩展的App描述文本数据作为实验数据集，采用文本处理技术，建立适当的App分类体系，提取文本特征并训练分类器，以达到较好的分类效果。对移动App进行分类，将有助于进一步分析用户行为，对推荐系统的设计也有着重要的意义。

1.2 研究现状

首先介绍文本分类问题的相关研究。文本分类问题，通过在预先分类好的文本集上进行训练，建立一个判别规则或者分类器，从而对未知类别的新样本进行自动归类。1971 年，Rocchio提出了在用户查询中不断通过用户的反馈来修正类权重向量，来构成线性分类器的方法。1979 年，Van Rijsbergen对信息检索领域的研究做了系统的总结，提出了一系列关于信息检索的概念和评估标准。文中还讨论了信息检索的概率模型，后来的文本分类研究大多数是建立在概率模型的基础上。1995年，Vipnik 基于统计理论提出了支持向量机(Support Vector Machine)方法，基本思想是寻找最优的高维分类超平面。支持向量机方法有着扎实的理论基础，与传统的算法相比，它有着良好的分类性能，并且在不同的数据集上显示了算法的鲁棒性。至今，支持向量机的理论和应用仍是分类研究的热点之一。

虽然App分类问题是近些年来产生的一个新的应用问题，但是仍然可以通过转换为短文本分类问题来解决。短文本分类问题有如下一些研究：Phan等人[1]提出了使用隐含主题来提高短文本的分类性能；Sahami等人[2]提出一种新的短文本相似度度量方法，这种方法可以被一种核函数证明；Yih等人[3]引入额外的学习过程来提高度量方法等。事实上，上述一些方法可以用来解决App分类问题，Ma等人[4]提出使用文本信息，通过余弦相似度来划分App类别。在这些论文中，Zhu等人[5]发表的论文，给出了较为综合的使用扩展信息的App分类方法，既使用了基于Web的文本信息来提取显式文本特征和隐式文本特征，同时也使用了App的日志信息来提取与类别相关的其它特征，建立最大熵分类器来实现App自动分类的功能。

本课题将使用App的描述文本数据作为扩展的数据集，通过文本聚类等方法建立分类指标，应用文本处理技术对App的描述文本进行处理，训练分类器以达到较好的分类效果。

1.3 研究主要难点

App分类研究中的主要难点如下：

1. 没有明确公认的App分类体系。当前有很多App应用商店提供App的下载服务，但是这些App应用商店并没有统一的App分类体系，不同的应用商店根据自己平台的特点设置App分类体系，因而可能会导致一些分类的不合理现象，如部分类别的概念过于模糊、过于广泛或者过于狭窄等。

2. App原始类别标注的不准确。由于App应用商店中的App类别大多为上传者手工标注，且很多App应用商店中的分类体系并不明确，所以不可避免的会出现类别标注的错误，无法直接使用App数据集所带有的原始类别标注。

3. App信息过少，难以直接用来解决分类问题。App自身携带的信息主要为App的名字、App的包名和App的版本号，这些信息只有很少的特征用于训练分类模型。对于App的名字而言，其包含的文本非常短，且包含的词汇相当稀少，很难建立App与相应类别的对应关系。

4. App描述文本的稀疏性和高维性。在应用商店中一般都会有对App的文字描述，这些描述文本是能获得到的与App最相关的文本数据。但是对于文本数据而言，文本特征较为稀疏，且文本本身具有较高的维度，这将会影响到训练分类器的性能。

1.4 本文的主要内容与章节安排

本文使用Google Play中App的中文描述文本来扩展App的文本信息，使用Google Play的分类体系作为原始分类体系。具体工作如下：

1. 结合原始分类体系和文本描述的聚类结果，重新构建App分类体系，使之在符合原始分类的基础上，尽可能符合文本相似度的分类。

2. 设计和实现App类别的标注方法，纠正App类别标注的错误。

3. 在标注后的App数据上训练分类器，实现App自动分类的功能。

4. 设计和实现基于文本处理技术的App分类系统的原型。

本文共分为6章，章节安排如下：第2章，相关工作，介绍和综述现有的文本处理技术；第3章，分类体系的构建与评价，分析原始App分类体系存在的问题，为解决这些问题，结合聚类结果，构建新的App分类体系，并评价新的App分类体系；在此基础上，标注没有划分类别的App；第4章，分类算法的训练和评价，在标注好的App数据集上训练分类算法，比较和

分析不同分类算法的效果；第5章，设计与实现App分类系统原型；第6章，对全文进行总结，并对未来工作进行展望。

图1-1 本文的章节结构

指　标
疑似剽窃文字表述

1. 文本分类问题的相关研究。文本分类问题，通过在预先分类好的文本集上进行训练，建立一个判别规则或者分类器，从而对未知类别的新样本进行自动归类。1971 年，Rocchio提出了在用户查询中不断通过用户的反馈来修正类权重向量，来构成线性分类器的方法。1979 年，Van Rijsbergen对信息检索领域的研究做了系统的总结，提出了一系列关于信息检索的概念
2. 文中还讨论了信息检索的概率模型，后来的文本分类研究大多数是建立在概率模型的基础上。
3. 与传统的算法相比，它有着良好的分类性能，并且在不同的数据集上显示了算法的鲁棒性。至今，支持向量机的理论和应用仍是分类研究的热点之一。

3. 第2章相关工作

总字数：2889

相似文献列表　文字复制比：24.8%(716)　疑似剽窃观点：(0)

1	用深度学习解决大规模文本分类问题 - starzhou的专栏 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	13.6%（393） 是否引证：否
2	用深度学习（CNN RNN Attention）解决大规模文本分类问题 - 综述和实践（转载）- u013818406的博客 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	13.6%（393） 是否引证：否
3	基于内容和用户标识的混合型垃圾弹幕识别与过滤研究 张树华(导师：王晓耘) -《杭州电子科技大学博士论文》- 2017-02-01	5.7%（164） 是否引证：否
4	sklearn基本用法----SVM - clover -《网络（http://blog.csdn.net）》- 2017	4.0%（116） 是否引证：否
5	机器学习——支持向量机（SVM）- 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	4.0%（116） 是否引证：否
6	孟秋辰_用深度学习方法分析基因表达调控关系 孟秋辰 -《大学生论文联合比对库》- 2017-05-30	3.9%（114） 是否引证：否
7	234-3013204088-董晨龙 董晨龙 -《大学生论文联合比对库》- 2017-05-26	3.6%（105） 是否引证：否
8	行人检测方法研究 伍子毅 -《大学生论文联合比对库》- 2017-06-06	2.9%（84） 是否引证：否
9	计算机辅助定密系统关键技术研究与实现 何流(导师：杜晔) -《北京交通大学博士论文》- 2017-06-01	2.4%（70） 是否引证：否
10	基于稀疏编码空间金字塔匹配的高中统计图形识别 吴婷(导师：魏艳涛) -《华中师范大学博士论文》- 2017-05-01	2.2%（64） 是否引证：否
11	2013b04004_薛志排_测绘工程 薛志排 -《大学生论文联合比对库》- 2017-05-10	1.4%（40） 是否引证：否
12	2013b04004_薛志排_测绘工程 薛志排 -《大学生论文联合比对库》- 2017-05-24	1.4%（40） 是否引证：否
13	基于烟雾多特征的吸烟行为识别算法研究 苏翔宇(导师：胡春海) -《燕山大学博士论文》- 2014-05-01	1.4%（40） 是否引证：否
14	基于支持向量机的分类器设计 王曼 -《大学生论文联合比对库》- 2015-05-07	1.4%（40） 是否引证：否
15	基于支持向量机的分类器设计 王曼 -《大学生论文联合比对库》- 2015-05-11	1.4%（40） 是否引证：否
16	基于支持向量机的分类器设计 王曼 -《大学生论文联合比对库》- 2015-05-06	1.3%（39） 是否引证：否
17	基于支持向量机的分类器设计 王曼 -《大学生论文联合比对库》- 2015-05-06	1.3%（39） 是否引证：否
18	基于多层感知机的分类算法研究 江风涛 -《大学生论文联合比对库》- 2016-05-09	1.3%（39） 是否引证：否
19	广州市城市扩张过程及热环境演变研究	1.3%（39）

	李雁(导师：陈颖彪) -《广州大学博士论文》- 2011-05-01	是否引证：否
20	压缩感知遥感图像融合及分类方法研究	1.3%（39）
	阮涛(导师：那彦) -《西安电子科技大学博士论文》- 2012-03-01	是否引证：否
21	银川平原城市空间扩展分析及情景预测研究	1.3%（39）
	张春梅(导师：张建明) -《兰州大学博士论文》- 2013-05-01	是否引证：否
22	高维分类问题的Logistic回归惩罚经验似然方法	1.3%（39）
	徐雯雯(导师：罗季) -《浙江财经大学博士论文》- 2014-12-01	是否引证：否
23	62100208-赵代全-地理信息系统-湿地水体遥感提取方法对比分析	1.3%（39）
	赵代全 -《大学生论文联合比对库》- 2014-05-29	是否引证：否
24	测绘专工程业英语词汇及名词解释	1.3%（39）
	-《互联网文档资源（http://wenku.baidu.c）》- 2016	是否引证：否
25	测绘专业英语词汇及名词解释	1.3%（39）
	-《互联网文档资源（http://wenku.baidu.c）》- 2017	是否引证：否
26	测绘专业英语词汇及名词解释	1.3%（39）
	-《互联网文档资源（http://wenku.baidu.c）》- 2017	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第2章相关工作

2.1 文本处理技术综述

采用文本处理技术，可以解决移动应用程序的分类问题。根据所处理文本数据是否含有类别信息，文本处理技术大致上可分为无监督的文本聚类问题和有监督的文本分类问题。

有监督的文本聚类问题主要依据文本相似度，对文本进行分簇，从而使得同一簇中文档相似度较高，不同簇中文档相似度较低。而无监督的文本分类问题则是在已有数据集上训练文本分类模型，把文本数据映射到给定类别中的某一个，从而应用于文本的类别预测。

无论是文本聚类还是文本分类，在进行处理之前，一般需要进行文本特征化过程。文本的特征化一般包括文本预处理、特征提取和文本表示三个部分，目的是将文本数据转换为便于计算机处理的形式，并封装足够的信息量，用于文本的聚类或者分类。

2.2 文本特征化方法

1. 文本预处理

中文的文本预处理过程包括文本分词和去除停用词两个部分。研究发现，文本的单词粒度要好于文本的字粒度，其原因在于文本中的单词比字包含更强的语义信息，显然，字粒度损失了过多的“n-gram”信息。然而，由于中文文本没有类似于英文文本中的空格分隔符来区分单词，所以在进行中文文本处理之前，需要通过文本分词技术来分隔单词。而去除停用词则是去除介词、连词等对文本处理无意义的词，通常需要维护停用词表，忽略停用词表中的停用词。

2. 文本表示

文本表示目的是将文本预处理之后的文本转换为便于计算机理解的形式。传统的方法是词袋模型（BOW，Bag Of Words）或向量空间模型（VSM，Vector Space Model）。词袋模型将文本中出现的每一个词作为一个维度，每个维度对应单词的词频作为该维度的值，其不足在于忽略文本的上下文关系，并且无法表征语义信息。另一方面，因为词库中单词的数量较多，因此词袋模型往往具有高纬度、高稀疏性，这将不利于文本的后续处理。

除了词袋模型，还有基于语义的文本表示方法，比如LDA主题模型，假设文档具有多个主题，将文档的主题作为特征的一部分构建文档特征。还有基于词嵌入（word embedding）的文本表示方法。Mikolov在2013年提出的Word2vec方法是一种词嵌入文本分布式表示方法[6][7][8]，基于CBOW 和 Skip-Gram两种模型训练神经网络模型，将单词表示为词向量，使单词封装更多的语义信息。

3. 特征提取

向量空间模型的特征提取，包括特征选择和特征权重计算两部分。

特征选择主要包括过滤式（filter）评价策略和封装式（wrapper）评价策略。过滤式特征选择方法，基本思路是根据评价指标对原始单词特征（term）进行评分排序，从中保留得分最高的一些特征项，过滤其余的特征项。根据是否需要文本的类别信息，可以分为无监督的特征选择评价指标和有监督的特征选择评价指标。

无监督的特征选择评价指标如下[9]：

（1）文档频率（DF，Document Frequency）

（2）单词权（TS, Term Strength）

（3）单词熵（EN, Entropy-based Feature Ranking）

（4）单词贡献度（TC, Term Contribution）

（5）单词变化度（TV, Term Variance）

另一方面，有监督的特征选择评价指标较之无监督的特征选择评价指标，更能选择出具有类别区分性的特征。常见的有监督的特征选择评价指标如下[10]：

（1）卡方检验（Chi-square）

（2）信息增益（IG, Information Gain）

（3）基尼系数（Gini Index）

特征权重主要是基于经典的TF-IDF方法，主要假设是一个词的重要度与在文档内的词频（term frequency, TF）成正比，与在所有文档中出现的次数成反比。TF-IDF权重的计算公式如下：TFIDFi,j=TFi,j×IDFi,j (2-1)TFi,j=ni,jknk,j (2-2)IDFi=log|D||{j:ti∈dj}| (2-3)

其中，TFIDFi,j表示单词特征i在文档j中的权重；TFi,j为单词特征i在文档j中的词频；IDFi为单词特征i的逆文档频率。

2.3 无监督文本聚类方法

文本聚类的目标是将文本相似度较高的文本划入同一类别，文本相似度较低的文本划入不同类别。聚类算法按照聚类的基本策略分为凝聚式聚类和点分配式聚类。

凝聚式聚类算法一开始将每个点都看成一个簇，在算法执行过程中，合并相似度最高的簇，直到达到预定的簇数目或紧密度标准。凝聚式聚类的典型算法是层次聚类。

另一类算法为点分配式聚类，这类算法按照某个顺序依次考虑每个点，并将它分配到合适的簇中。这类算法一般需要有初始化簇的过程。点分配式聚类的典型算法之一是K-means算法。

在聚类中需要定义距离测度，距离测度包括Jaccard距离、欧式距离、余弦距离、编辑距离和海明距离等。用于衡量文本距离的常用测度为余弦距离。文本向量的夹角余弦一般被称为余弦相似度，文本的余弦距离定义为1减去其余弦相似度。因此，文本的余弦相似度越高，文本之间的余弦距离越接近。

2.4 有监督文本分类方法

分类问题是机器学习非常重要的一个组成部分，它的目标是根据已知样本的某些特征，判断一个新的样本属于哪种已知的样本类。分类问题也被称为监督式学习(supervised learning)，根据已知训练区提供的样本，通过计算选择特征参数，建立判别函数以对样本进行的分类。大部分机器学习方法都在文本分类领域有所应用，下面简要介绍朴素贝叶斯分类算法和支持向量机两种分类器。

在机器学习中，朴素贝叶斯分类器（Naïve Bayes，NB）是一系列以假设特征之间强（朴素）独立下运用贝叶斯定理为基础的简单概率分类器。朴素贝叶斯分类器是以词频为特征判断文本所属类别或其他（如垃圾邮件、合法性、体育或政治等等）的问题。通过适当的预处理，它可以与这个领域更先进的方法（包括支持向量机）相竞争。

支持向量机（support vector machine，SVM）是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。其基本原理为，将实例表示为空间中的点，通过寻找两个类别实例之间的间隔，将两个类别分隔到两侧。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。除了进行线性分类之外，SVM还可以使用核方法来处理非线性的分类问题，即将其输入映射到高维特征空间中来寻找类别之间的间隔的方法。SVM分类器在文本分类、图像分类等方面有着较为广泛的应用。

2.5 本章小结

本章综述了文本处理相关的技术，并且介绍了文本特征化方法、无监督的文本聚类方法和有监督的文本分类方法。在这些技术的基础上，本文的后续部分将使用文本处理技术对App描述文本进行聚类，构建App分类体系，然后使用该分类体系标注文本描述数据的类别，训练文本数据分类器，实现App自动分类的功能。

指　标
疑似剽窃文字表述

1. 分类。
   2.2 文本特征化方法
   1. 文本预处理
   中文的文本预处理过程包括文本分词和去除停用
2. 文本的后续处理。
   除了词袋模型，还有基于语义的文本表示方法，比如LDA主题模型，
3. 然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。除了进行线性分类之外，SVM还可以使用核方法来处理非线性的分类问题，即将其输入映射到高维特征空间

4. 第3章分类体系的构建与评价_第1部分

总字数：10444

相似文献列表　文字复制比：0.5%(55)　疑似剽窃观点：(0)

1	数据挖掘技术在客户保持中的应用研究	0.5%（55）
	田小霞,刘晓霞,范全润 - 《计算机时代》- 2003-02-25	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第3章分类体系的构建与评价

3.1 引言

本文使用Google Play中的App文本描述来扩展App数据，并使用Google Play中的App分类体系作为原始分类体系，在此基础上调整App分类体系。

本章首先对原始分类体系进行介绍，分析原始分类体系中的问题，然后对App文本描述进行特征选择，并进行文本聚类。结合聚类结果和原始分类体系，手工标注App类别，构建新的App分类体系，并对新的分类体系进行评价。在新的分类体系之下，对未划分类别的App数据进行类别的标注，按文本相似度的方式划分App数据，并在最后对App数据类别的标注进行了评价。

3.2 原始分类体系问题分析

在构建分类体系之前，本节首先简述原始分类体系的详细信息以及原始分类体系中的问题，从而说明重新构建App分类体系的必要性。

3.2.1原始分类体系简介

本文中使用的原始分类体系为Google Play中对App的分类体系，数据来源为Google Play中的中文App文本描述。Google Play对App的分类体系共包含49个类别，其中，游戏类App包含17类，非游戏类App包含32类。Google Play中的App的类别标注方法为App上传者手工标注，上传者根据Google Play提供的类别人为选择适当的类别进行标注。Google Play的App分类体系及数据集大小如表3-1。

表3-1 Google Play的App分类体系及数据集大小

非游戏类App类别 App数量游戏类App类别 App数量

ART_AND_DESIGN AUTO_AND_VEHICLES BEAUTY BOOKS_AND_REFERENCE BUSINESS COMICS COMMUNICATION DATING EDUCATION ENTERTAINMENT EVENTS FINANCE FOOD_AND_DRINK HEALTH_AND_FITNESS HOUSE_AND_HOME LIBRARIES_AND_DEMO LIFESTYLE MAPS_AND_NAVIGATION MEDICAL MUSIC_AND_AUDIO NEWS_AND_MAGAZINES PARENTING PERSONALIZATION PHOTOGRAPHY PRODUCTIVITY SHOPPING SOCIAL SPORTS TOOLS TRAVEL_AND_LOCAL VIDEO_PLAYERS WEATHER
166198174300300209300157300300127300116300144300300300300300300172300300300300300300300300300
GAME_ACTION GAME_ADVENTURE GAME_ARCADE GAME_BOARD GAME_CARD GAME_CASINO GAME_CASUAL GAME_EDUCATIONAL GAME_MUSIC GAME_PUZZLE GAME_RACING GAME_ROLE_PLAYING GAME_SIMULATION GAME_SPORTS GAME_STRATEGY GAME_TRIVIA GAME_WORD
300300300300300204300300166300300300300300300291227

实验中，App描述文本的数据总量为13151条，其中游戏类App数量为4788条，非游戏类App数量为8363条。

3.2.2原始分类体系的数据统计

对原始数据集进行初步处理和统计，进而分析原始分类体系的问题。首先对原始文本数据进行中文分词，去除停用词，建立向量空间模型。定义距离测度为余弦距离，对不同类别中心之间的距离和不同类别的半径进行统计分析。

1. 不同类别中心之间的余弦距离统计

定义类别的中心为类别中文本向量的均值中心，求算各类别中心之间的余弦距离，得到各类别中心的余弦距离矩阵。类别中心代表了该类别数据在向量空间中的位置，类别中心之间的余弦距离能够在一定程度上反映各个类别之间的相似程度。如表3-2为部分类别的余弦距离矩阵。

表3-2 原始分类体系中部分类别的余弦距离矩阵

ART_AND_DESIGN AUTO_AND_VEHICLES BEAUTY BOOKS_AND_REFERENCE BUSINESS COMICS COMMUNICATION

ART_AND_DESIGN 0.000 0.887 0.609 0.847 0.840 0.670 0.871

AUTO_AND_VEHICLES 0.887 0.000 0.906 0.819 0.670 0.888 0.754

BEAUTY 0.609 0.906 0.000 0.887 0.843 0.778 0.880

BOOKS_AND_REFERENCE 0.847 0.819 0.887 0.000 0.664 0.718 0.714

BUSINESS 0.840 0.670 0.843 0.664 0.000 0.813 0.508

COMICS 0.670 0.888 0.778 0.718 0.813 0.000 0.797

COMMUNICATION 0.871 0.754 0.880 0.714 0.508 0.797 0.000

表格中元素为类别中心之间的余弦距离，可以看到，BUSINESS和COMMUCATION之间的余弦距离最近，为0.508；COMICS和AUTO_AND_VEHICLES之间的余弦距离最远，为0.888。类似的，对于全部类别，余弦距离最近的TOP 5类别和余弦距离最远的TOP 5类别如下。

表3-3 余弦距离最近的TOP 5类别

('GAME_ACTION', 'GAME_ARCADE')0.235

('GAME_ACTION', 'GAME_ROLE_PLAYING')0.323

('PRODUCTIVITY', 'TOOLS')0.328

('GAME_EDUCATIONAL', 'PARENTING')0.335

('GAME_ACTION', 'GAME_STRATEGY')0.335

表3-4 余弦距离最远的TOP 5类别

('ART_AND_DESIGN', 'FOOD_AND_DRINK') 0.953

('GAME_CASINO', 'PARENTING') 0.953

('GAME_MUSIC', 'WEATHER') 0.953

('GAME_EDUCATIONAL', 'WEATHER') 0.961

('GAME_CASINO', 'WEATHER') 0.966

由表3-3和表3-4对比可以发现，原始分类体系下的各类别数据的余弦距离有着较大的差异。类别中心的余弦距离较远，说明这些类别在文本相似度上有着较大的区分，类别成分的交叉可能较少；而类别中心的余弦距离较近，说明这些类别在文本相似度上难以有明确的区别，可能存在着成分的交叉。

另外可以发现，游戏类App和非游戏类App差别不大。如表3-3中，GAME_EDUCATION和PARENTING的中心余弦距离为0.335，存在一定的交叉，这意味着游戏类App和非游戏类App在文本相似度上没有明显的区分。举例来说，GAME_EDUCATION中包含“儿童学数学”App，PARENTING中包含“速算”App，两者的描述文本十分相似。

2. 各类别的半径统计

定义类别的半径为类别中各数据到类别中心的余弦距离的均值。类别的半径一定程度上反映了类别的松散程度，类别的半径大，说明类别中各数据相似度较低，类别的成分较为松散；类别的半径小，说明类别中数据相似度较高，类别的成分较为集中。

表3-5 半径最大的TOP 5类别

LIFESTYLE 0.874

ENTERTAINMENT 0.874

GAME_CASUAL 0.870

BUSINESS 0.853

TOOLS 0.844

表3-6 半径最小的TOP 5类别

GAME_EDUCATIONAL 0.689

DATING 0.683

GAME_CASINO 0.678

FOOD_AND_DRINK 0.663

WEATHER 0.613

从表3-5和表3-6可以看到，LIFESTYLE，ENTERTAINMENT，GAME_CASUAL，BUSINESS，TOOLS这些类别较为松散，类别数据较为宽泛，类别中可能包含多种成分。举例来说，LIFESTYLE中包含“美团”App，“日语学习”App和“易办违章”App，应该分属于不同的类别。而GAME_EDUCATIONAL，DATING，GAME_CASINO，FOOD_AND_DRINK，WEATHER这些类别较为集中，类别数据较窄，类别中的成分比较单一。

3.2.3原始分类体系的问题分析

综合上述两种统计指标，可以发现数据的原始分类体系存在如下问题：

（1）原始类别中有些类别之间概念存在交叉，类别之间的相似度较高，难以区分

（2）原始类别的概念范围不准确，导致有些类别的较为松散，可能内部包含多个真实类别；有些类别的划分过小，导致类别数量过多

（3）原始类别为App上传者手工标注，不同的标注者对原始分类的理解不同，可能导致一些App数据被划入错误的类别中

举例来说：对于（1），表3-3中TOOLS和PRODUCTIVITY有着交叉，TOOLS包含各种工具，PRODUCTIVITY包含生产力工具；对于（2），表3-5中LIFESTYLE意为生活方式，概念较宽，而各种GAME划分较细，产生了大量的类别

因为上述问题的存在，如果直接采用原始类别，将不利于后续分类问题的处理。所以在本实验中，需要借助聚类方法来对App的描述文本数据进行处理，得到符合文本相似度的App的类别，参考于聚类结果对原始分类体系进行合并、拆分和重组，使新的分类体系更加合理。

3.3 文本特征的构建方法设计

在进行中文文本聚类之前，需要对文本进行特征选择，用来提高文本聚类的性能。为了选择出更加具有普遍性的特征，本文没有使用原始类别来进行有监督的特征选择，而是使用无监督的特征选择方法。本节首先介绍常用的无监督特征评价指标，然后设计特征选择方法，进行三次特征选择，最后采用评价指标来确定最终的特征数量。

3.3.1常用的无监督特征评价指标

用于表示文本的向量的维度较高，为了便于聚类，需要对构成文本向量的特征进行特征选择。相关工作中提到5种常见的无监督特征选择评价指标，具体定义如下。其中D表示文档集，M表示特征的总维数，N表示文档在文档集中的数量。

（1）文档频率（DF，Document Frequency）

文档频率定义为，单词特征在文档中出现的文档数量。这是一种最简单的特征选择指标，且非常适合在大规模文档的进行特征选择。虽然这种指标非常简单，但是被实验中证明是有效的。

（2）单词权（TS, Term Strength）

单词权定义为，在相似的文档对中，单词特征在第一篇文档中出现的条件下，在第二篇文档中出现的条件概率。公式定义如下：TSt=pt∈djt∈di, di, dj∈Dnsimdi, dj>β (3-1)

其中β为判定文档对相似的阈值。因为公式中需要计算每对文档对的相似度，所以复杂度是O(N2)。

（3）单词熵（EN, Entropy-based Feature Ranking）

单词熵定义为，在移除单词特征后，文档集的熵的减少程度。公式定义如下：Et=-i=1Nj=1N(Si,j×logSi,j+(1-Si,j)×log(1-Si,j)) (3-2)

其中Si,j为文档di和文档dj的相似度，定义公式如下：Si,j=e-α×disti,j, α=-ln(0.5)dist (3-3)

其中，disti,j为文档di和文档dj的在单词特征t移除后的距离，dist是单词特征t被移除后所有文档的平均距离。最大的问题是其复杂度为O(MN2)，所以对于大规模数据集一般需要采样。

（4）单词贡献度（TC, Term Contribution）

由于文本聚类的结果依赖于文档之间的相似度，所以每个单词特征的贡献度可以看作是它对文档集整体的相似度的贡献程度。因此，单词贡献度定义如下：TCt=i,j∩i≠jf(t,di)×f(t,dj) (3-4)

其中，f(t,d)表示文档d中单词特征t的TF-IDF权重。当所有单词特征的权重都相等时，即可简化为ft,d=1，此时TC(t)可以表示为：TCt=DF(t)(DFt-1) (3-5)

因为TC(t)单调递增且DF(t)≥0，所以可知DF(t)其实是TC(t)的一种特例。使用倒排索引技术，TC的时间复杂度为O(MN2)，其中N为每个单词特征在文档中的平均出现次数。

（5）单词变化度（TV, Term Variance）

单词变化度是单词特征的权重在文档集中的方差。它衡量了单词特征在文档集中的变化程度，变化程度高的单词特征可能对文档区分性较强，变化程度低的单词特征可能对文档区分性较弱。公式定义如下：TVt=i=1nf(t,di)-f(t)2 (3-6)

其中，f(t,di)为单词特征t在di中的TF-IDF权重，f(t)为单词特征t在文档集中的平均权重。

3.3.2特征选择方法设计

依据3.3.1中介绍的特征评价指标，实验中对特征进行三次特征选择。通过特征选择，在尽可能保留文本相似度信息的情况下，降低文本向量的维度。具体特征选择方法如下：

（1）基于词性对特征进行过滤：名词和动词是构成文本的主要成分，去除形容词、副词、助词等成分

（2）基于文档频率对特征进行过滤：DF值是单词在文档中出现的频率。DF值过高的单词，说明大多数文档都具有该词，不具备区分性；DF值过低的单词，说明该词只出现在极个别的文档中，可能为少量文档所独有，也不具备区分性

（3）基于文本聚类的特征选择指标进行过滤：按TV和TC指标过滤，将特征分别按TV和TC指标降序排序，过滤掉后k（%）的交集部分。也就是过滤掉在TV和TC看来都比较差的部分，尽可能多的保留特征。

3.3.3 特征维度的最终确定

在第三次特征选择的过程中，需要决定参数k的值。为此，需要采取一定方法在求得最佳的参数值。假设原始数据的原始分类体系具有一定的合理性，在进行特征选择后，原始分类体系在衡量分类体系的指标上应该至少不比特征选择之前差。

常用于衡量分类体系的指标有DBI和轮廓系数（Silhouette Coefficient），具体定义如下。

（1）DBI指标：DBI= 1Kimaxj,j≠i1nix∈Cidx,ci+1njx∈cjdx,cj/dci,cj(3-7)

其中，K为类别数量，x为类别中样本，c为类别的均值中心，d为距离测度，此处使用余弦距离作为其距离测度。

DBI评价了类别内部数据的紧密程度和类别之间数据的离散程度。DBI指标越低，分类体系的效果越好，反之越差。值得一提的是，DBI在定义中重点反映了分类体系在最坏类别中的分类效果，即最为松散的类别半径和最为接近的类别之间的距离，因而并不能反映分类体系整体的效果。

（2）轮廓系数:SIL= 1Kk=1K1nki∈Iksi, si=bi-aimaxai,bi=1-aibi, if bi>a(i)(3-8)

其中，K为类别数量，b(i)为样本i到所有不包含样本i的类别的所有点的平均距离，a(i)为样本i到样本i所处类别内的所有点的平均距离。

与DBI指标一样，轮廓系数同时评价了类别内部数据的紧密程度和类别之间数据的离散程度。轮廓系数越高，分类体系的效果越好，反之越差。与DBI不同的是，轮廓系数考察了所有的样本点的情况。

应用上述两种评价指标，在原始分类体系的基础上，逐步增大特征过滤比例k值的取值，DBI指标和轮廓系数随过滤比例k的变化曲线如图3-1和图3-2所示。

图3-1 DBI随文本特征过滤比例k变化曲线

图3-2 轮廓系数随文本特征过滤比例k变化曲线

由图3-1和图3-2得如下结论：

（1）DBI在聚类的特征选择过程中不具备衡量作用

如图3-2所示，随着特征数量的减少，DBI指标呈现逐渐增大的趋势。DBI指标衡量了所有类别中最差的类别随着特征减少

的指标变化趋势，即，随着特征数量的减少，原来最相近的类别变得更为相近。无法决定应过滤的特征数量。

（2）SIL轮廓系数能够在特征选择中起到衡量作用

从图3-3来看，轮廓系数呈现从稳定无明显变化到逐渐减小的趋势，拐点出现在0.81处，在0.98处出现负值。轮廓系数考察了每一个点在分类体系下的距离指标，随着特征数量的减少，分类体系的类别之间的松散程度和类别内部的紧密程度的差值增大，意味着类别内的数据距离变得更小，类别间的数据距离变得更大，改善了原有分类体系的效果。在大于0.97处出现负值，意味着类别间的数据距离小于类别内的数据距离，数据丢失过多信息，使原始分类体系出现了错误。因此可以选择0.81作为k的参考值，即第三次过滤特征的比例。

表3-7 三次过滤特征的特征数量统计表

过滤特征次数选择特征方法特征数量

第0次过滤特征中文分词，去停用词 72841

第1次过滤特征词性 n, v 52825

第2次过滤特征 DF [7, 1000] 11137

第3次过滤特征 TC + TV 2506

表3-8 最终的特征按TF-IDF降序排序，TOP10的特征统计表

单词特征 TF-IDF

孩子天气地图新闻宝宝记录文件壁纸故事水平

0.012670.010310.010240.009520.009180.009070.008920.008780.008630.00851

表3-7展示了三次过滤特征的详细情况。表3-8展示了经过了特征选择，文本向量中平均TF-IDF最高的单词特征及其权重，可以看到，通过特征选择方法，得到的TF-IDF最高的单词特征带有较强的类别区分信息。

3.4 文本聚类方法设计

本节将使用层次聚类方法和K-means聚类方法对App文本描述进行聚类。首先对层次聚类方法和K-means聚类方法进行简要介绍，然后对App描述文本进行聚类，最后对聚类结果进行评价。

3.4.1文本聚类方法

本文中使用层次聚类方法和K-means聚类方法对App描述文本进行聚类，并对聚类结果进行分析。

1. 层次聚类方法

首先介绍层次聚类方法，层次聚类方法可以分为自底向上和自顶向下两类。自底向上的聚类方法，首先把每一个样本视为一个簇，随时间推移每次寻找最接近的簇进行合并；自顶向下的则相反，首先把所有样本视为一个簇，随时间推移每次拆分距离最远的簇。下面重点介绍自底向上的层次聚类方法。

自底向上的层次聚类方法的基本步骤：

（1）把每个样本归为一簇，计算每两个簇之间的距离；

（2）寻找各个类之间最近的两个簇，把二者归为一簇；

（3）重新计算新生成的这个簇与每个旧簇之间的距离；

（4）重复2和3，直到满足结束条件退出。

类别之间距离的计算方法主要有如下几种：最近距离（Single），最远距离（complete），平均距离（average）和离差平方和（ward）。最近距离定义为两个簇中最近的两个样本点之间的距离；最远距离定义为两个簇中最远的两个样本点之间的距离；平均距离定义为两个簇中任意两个样本点之间的距离的平均值；离差平方和为结果簇内所有样本点到质心的距离平方和。在层次聚类中，不同的距离定义可能导致完全不同的聚类过程和聚类结果。

层次聚类的最大优点是聚类结果便于理解，且不需要提前设定类别数目。层次聚类在聚类过程中不断合并相似的子簇，最终可以生成一个簇间合并的树形结构，反映了簇间的相似关系。因此也就可以实现所谓的“随取随用”，根据需要选择合适的类别数目和聚类结果。

2. K-means聚类方法

K-means是典型的基于点分配策略的聚类算法，由于K-means方法的时间复杂度较低，因此在文本聚类问题中得到广泛的应用。其算法基本步骤如下：

（1）选择K个点作为初始质心；

（2）将每个点指派到最近的质心，从而形成K个不同的簇；

（3）重新计算每个簇的质心；

（4）重复（2）、（3），直至簇不发生变化或达到最大迭代次数。

在K-means聚类中，一般使用误差平方和（Sum of the Squared Error, SSE）作为聚类的度量函数，多次运行并选择最佳的一次运行结果作为最终的聚类结果。SSE定义如下，其中k表示聚类数目，Ci表示第i个聚类中心，dist表示欧几里得距离：SSE=i=1Kx∈Cidist(Ci,x)2 (3-9)

虽然K-means算法非常简单且使用广泛，但是K-means依然有一些缺陷，如：K值需要预先指定；K-means对初始选取的聚类中心十分敏感；K-means算法并不适合于所有数据类型；对离群点的数据进行聚类，K-means也会出现问题。

3. 聚类方法设计

本实验中聚类的用途为：作为重新标定数据类别的参考，比较原始类别的分类体系和聚类结果的区别，用以手工标定新的类别体系。因此为了便于比较，人为选定和原始分类体系相同的类别数（即K = 49），得到聚类结果。

因为层次聚类有着便于理解、易于分析簇间相似关系的优点，所以实验中使用层次聚类进行聚类，简单依据相似关系切分成49类，并使用K-means方法得到聚类结果进行对比。

实验设置：

（1）层次聚类方法：以average和ward两种距离度量进行聚类，使用余弦距离作为距离测度，并按簇间合并的层次结构切分为49类。

（2）K-means聚类方法：以欧几里得距离作为距离测度，设置超参数K为49，使用K-means++方法初始化簇心，执行10次按SSE取最佳聚类结果。

3.4.2 聚类评价指标

在进行结果分析之前，定义评价聚类结果的指标。在使用DBI指标和轮廓系数之外，为便于分析结果，定义如下指标：

（1）平均外部距离(out_dist): 簇中心间距离的均值out_dist=1Ki1K-1j,j≠idist(ci,cj) (3-10)

（2）平均外部最近距离(out_min): 簇中心间最近距离的均值out_min=1Kiminj,j≠i{dist(ci,cj)} (3-11)

（3）平均内部平均距离(in_dist): 簇内距离的均值in_dist=1Ki1nixiCidist(xi,ci) (3-12)

（4）平均内部最远距离(in_max): 簇内最大距离的均值in_max=1Kimaxxi∈Ci{xi,ci} (3-13)

3.4.3 聚类结果评价

实验中使用了k-means，以ward为度量的层次聚类和以average为度量的层次聚类，得到三种聚类结果，与原始分类体系进行比对，不同聚类方法对应指标统计如表3-9所示。

表3-9 不同聚类方法对应指标统计

聚类方法 K DBI 轮廓系数平均外部距离(outer_dist) 平均外部最近距离(outer_min) 平均内部平均距离(inner_dist) 平均内部最远距离（inner_max） 不同类别样本数量的方差

原始类别 49 3.749 0.011 0.782 0.448 0.753 0.968 3280.687

K-means聚类 49 2.182 0.045 0.904 0.672 0.594 0.787 286439.953

层次聚类（ward） 49 2.116 0.018 0.923 0.665 0.511 0.810 307815.789

层次聚类（average） 49 1.921 0.016 0.917 0.772 0.593 0.813 191297.014

其中，DBI反映了聚类结果中最差的簇的聚类情况；轮廓系数反映了聚类结果的平均情况；四种指标反映了类间的松散程度和类内的聚集程度；不同类别样本数量的方差衡量了聚类结果的平衡程度，较低的值反映出聚类结果中各类别的样本数量更为平衡。

可以得到如下结论：

（1）三种聚类结果从各个指标上看都优于原始类别；

（2）从轮廓系数上看，k-means优于两种层次聚类；两种层次聚类差别不大；

（3）从DBI上看，层次聚类（average）优于k-means和层次聚类（ward）；

（4）从方差来看，层次聚类（average）比k-means和层次聚类（ward）更平衡；

其中，层次聚类方法（average）的簇间层次结构如图3-3所示，横线为切分49类的位置，可以看到样本数据的聚类过程和簇间的相似度关系。

图3-3 层次聚类方法（average）的簇间层次结构

图3-4展示了原始分类体系、K-means聚类、层次聚类（ward）和层次聚类（average）的可视化结构，从图中可以看到：原始分类体系的结构最为混乱；K-means聚类结果中，存在一个明显的“大簇”混杂在其余各簇之中；层次聚类（ward）和层次聚类（average）均得到了较为清晰的聚类结果。

图3-4 原始分类体系、k-means聚类、ward层次聚类和average层次聚类的t-SNE可视化图

综合上述分析，以average为度量的层次聚类方法，与K-means方法相比，能够得到类别间的相似关系，同时避免了K-means聚类结果中样本数量不平衡的情况；与以ward为度量的层次聚类方法相比，聚类评价的指标更好。为便于手工划分新

的分类体系，实验中选择以average为度量的层次聚类方法进行下一步处理。

3.5 分类体系的构建方法设计

原始类别为上传者手工标注的，在一定程度上符合于人的认知；而聚类划分出来的类别，对于描述文本而言，更符合统计上的文本相似度。手工划分新的类别的原则为，在原始类别划分的基础上，参考聚类结果，调整原始类别使其尽可能符合文本相似度。

手工划分过程如下：

（1）调整聚类结果。忽略样本数量过小的簇（样本数量<30）；对于样本数量过大的簇（样本数量>1000），如果相对集中则不作处理，相反，则将该类别按层次聚类的过程，适当拆分出符合层次聚类结构的子簇。

（2）建立原始类别和聚类结果之间的对应关系，保留对应关系中较大的部分。具体来说，如果簇中原始类别的样本的数量占簇中样本总数的比例超过30%，则视为原始类别与聚类结果具有较大的对应关系。

（3）将原始类别按聚类结果进行拆分、合并和重组，得到新的类别划分。

其中保留原始类别不应该合并、拆分的类别。举例来说，原始类别GAME_SPORTS和SPORTS在聚类后被划分至同一簇中，说明这两类从文本相似的角度上讲是相似的，但是从App分类的背景知识上来说这两类是不同的两个类别，不应该因为文本聚类结果上的相似而合并。类似的保留其它的一些原始类别划分。

表3-10展示了原始类别与新类别之间的对应关系表，表3-11展示了新的分类体系下各个类别信息。

表3-10 原始类别与新类别的对应关系

旧类别标签新类别标签 (新,旧)数量旧标签大小新标签大小

MAPS_AND_NAVIGATION0129195310

TRAVEL_AND_LOCAL0181181310

AUTO_AND_VEHICLES1110110176

MAPS_AND_NAVIGATION166195176

GAME_RACING2221221285

GAME_SIMULATION264158285

HOUSE_AND_HOME3565656

表3-11 新的分类体系下，各类别信息

类别标号别名样本数量 TFIDF TOP5 来源

0 地图旅行 310['地图', '旅行', '路线', '酒店', '旅游']{'TRAVEL_AND_LOCAL', 'MAPS_AND_NAVIGATION'}

1 汽车车辆 176['汽车', '车辆', '车', '二手车', '驾驶']{'AUTO_AND_VEHICLES', 'MAPS_AND_NAVIGATION'}

2 赛车游戏 285['赛车', '驾驶', '汽车', '卡车', '停车']{'GAME_RACING', 'GAME_SIMULATION'}

3 公寓房子 56['公寓', '装饰', '房子', '想法', '卧室']{'HOUSE_AND_HOME'}

4 交易理财 323['交易', '银行', '投资', '股票', '客户']{'BUSINESS', 'FINANCE', 'SHOPPING'}

5 购物美食 332['购物', '食谱', '商品', '优惠', '美食']{'FOOD_AND_DRINK', 'SHOPPING', 'LIFESTYLE'}

6 新闻阅读 221 ['新闻', '阅读', '文章', '杂志', '报纸']{'NEWS_AND_MAGAZINES', 'BOOKS_AND_REFERENCE'}

7 电影视频 122['电影', '电视', '播放器', '观看', '直播']{'VIDEO_PLAYERS', 'ENTERTAINMENT'}

8 音乐音频 198 ['歌曲', '播放器', '播放列表', '专辑', '均衡器'] {'MUSIC_AND_AUDIO'}

9 音乐游戏 103['钢琴', '歌曲', '节奏', '鼓', '音符']{'GAME_MUSIC'}

续表： 表3-11 新的分类体系下，各类别信息

10 个性设置 311['壁纸', '图标', '键盘', '桌面', '锁']{'ART_AND_DESIGN', 'TOOLS', 'PERSONALIZATION', 'BEAUTY'}

11 天气预测 233['天气', '天气预报', '预测', '温度', '雷达']{'WEATHER'}

12 聊天通话 387['聊天', '约会', '通话', '短信', '发送']{'BUSINESS', 'COMMUNICATION', 'SOCIAL', 'DATING'}

13 各种工具 398['文件', '浏览器', '笔记', '备份', '演示'] {'LIBRARIES_AND_DEMO', 'PRODUCTIVITY', 'COMMUNICATION', 'VIDEO_PLAYERS', 'TOOLS'}

14 拍摄拍照 393['效果', '滤镜', '拍摄', '拍', '图像']{'ART_AND_DESIGN', 'ENTERTAINMENT', 'PHOTOGRAPHY', 'BEAUTY', 'VIDEO_PLAYERS'}

15 漫画卡通 80['漫画', '头像', '阅读', '绘画', '卡通']{'COMICS'}

16 医疗药物 124['医疗', '药物', '医生', '临床', '医学']{'MEDICAL'}

17 锻炼运动 228['锻炼', '运动', '训练', '健身', '睡眠']{'SPORTS', 'MEDICAL', 'HEALTH_AND_FITNESS'}

18 单词词典 175['单词', '英语', '翻译', '词典', '发音']{'EDUCATION', 'BOOKS_AND_REFERENCE'}

19 文字游戏 229['字', '单词', '词', '字母', '文字游戏']{'GAME_TRIVIA', 'GAME_WORD'}

20 休闲游戏 132 ['发型', '头发', '公主', '女孩', '宠物'] {'BEAUTY', 'GAME_EDUCATIONAL', 'GAME_CASUAL'}

21 育儿孩子 415['宝宝', '孩子', '儿童', '巴士', '动物']{'EDUCATION', 'GAME_TRIVIA', 'GAME_EDUCATIONAL', 'PARENTING'}

22 卡牌游戏 375['老虎机', '扑克', '纸牌', '赌场', '牌']{'GAME_CASINO', 'GAME_CARD'}

23 体育游戏 162['足球', '篮球', '球', '球员', '球队']{'GAME_SPORTS'}

24 体育项目 113['足球', '统计', '球员', '球队', '俱乐部']{'SPORTS'}

25 棋类游戏 139['国际象棋', '棋盘', '棋子', '象棋', '骰子']{'GAME_BOARD'}

26 飞行游戏 36['飞机', '飞行', '直升机', '模拟器', '飞行员']{'GAME_SIMULATION'}

续表： 表3-11 新的分类体系下，各类别信息

27 综合游戏 1380 ['敌人', '英雄', '武器', '冒险', '升级'] {'GAME_STRATEGY', 'GAME_CASUAL', 'GAME_ADVENTURE', 'GAME_TRIVIA', 'GAME_ROLE_PLAYING', 'GAME_ACTION', 'GAME_BOARD', 'GAME_SIMULATION', 'GAME_ARCADE', 'GAME_PUZZLE'}

需要注意的是，手工划分分类体系存在一些文本相似但依然未被合并的类别，如表3-12所示，其中类别序号1中的类别与类别序号2中的类别在文本上有相似关系。

表3-12 新的分类体系下，文本相似但依然未被合并的类别

类别序号1 类别别名1 类别序号2 类别别名2

1 汽车车辆 2 赛车游戏

7 电影视频 8 音乐音频

8 音乐音频 9 音乐游戏

18 单词词典 19 文字游戏

20 休闲游戏 21 育儿孩子

23 体育游戏 24 体育项目

3.6 分类体系的评价

本节将从整体和局部两个角度对手工构建出的分类体系进行评价。

3.6.1 整体角度评价

从整体角度看，使用数据子集的新的类别和子集的原始类别进行对比，从而得到新的分类体系的效果评价。

表3-13 新旧分类体系下，各评价指标对比

类别 K DBI 轮廓系数平均外部距离(outer_dist) 平均外部最近距离(outer_min) 平均内部平均距离(inner_dist) 平均内部最远距离（inner_max）

原始类别 49 3.127 0.037 0.867 0.495 0.683 0.941

新的类别 28 2.204 0.049 0.884 0.677 0.675 0.946

从表3-13中可以看到，和原始类别相比，新的类别比原始类别更符合统计上的文本相似度。

在手工标注类别的过程中，去除了原始分类和聚类结果之间的对应关系中较为零散的原始类别样本，只保留了较大的原始类别样本，因而手工标注类别的样本是所有样本的一个子集。与图3-4中的原始分类图相比，图3-5显示了所有标注了新类别的样本在原始分类体系下的分类情况，去除了所有未标注新类别的样本。

图3-5 标注数据的原始分类体系的t-SNE可视化结构

图3-6 标注数据的新的分类体系的t-SNE可视化结构

从图3-5和图3-6对比可知，除文本相似但仍保留为不同类别的类别（如表3-12所示）之外，新的分类体系较原始分类体系相似关系更为清晰，相似度高的簇被合并至同一类，相似度低的类别被拆分成不同类别。

3.6.2 局部角度评价

从局部角度看，原始分类体系下的类别在新的类别中被拆分、合并和重组。选择一部分类别子集，对类别子集进行评价。如表3-14所示，在对应关系中MEDICAL被拆分为18、19；SPORTS被拆分为19、27；HEALTH_AND_FITNESS与SPORTS和MEDICAL合并。

表3-14 在新的类别中，抽取子集如下

18 医疗药物 124['医疗', '药物', '医生', '临床', '医学']{'MEDICAL'}

19 锻炼运动 228['锻炼', '运动', '训练', '健身', '睡眠']{'SPORTS', 'MEDICAL', 'HEALTH_AND_FITNESS'}

27 体育项目 113['足球', '统计', '球员', '球队', '俱乐部']{'SPORTS'}

对MEDICAL、SPORTS、HEALTH_AND_FITNESS重新聚类，将聚类结果与划分出的新类别进行比对。如表3-15和表3-16所示，可以看到，聚类结果中除了类别0和类别4这两个较小的簇，其它的簇与手工划分的类别在TF-IDF上有着很高的相似度。

表3-15 手工划分的类别和TFIDF分布如下

类别类别大小 TFIDF TOP10

18 124 ['药物', '医疗', '医生', '临床', '医院', '医学', '治疗', '病人', '医师', '挂号']

19 228 ['锻炼', '训练', '运动', '健身', '睡眠', '记录', '体重', '教练', '测量', '活动']

27 113 ['足球', '统计', '球员', '球队', '团队', '俱乐部', '网球', '新闻', '高尔夫', '得分']

表3-16 聚类结果的类别和TFIDF分布如下

类别类别大小 TFIDF TOP10

0 3 ['消息', '页面', '组织', '相册', '巴黎', '地理', '观点', '报纸', '指南针', '问答']

1 110 ['药物', '医疗', '医生', '临床', '医院', '医学', '治疗', '病人', '挂号', '医师']

2 225 ['锻炼', '训练', '运动', '健身', '睡眠', '记录', '体重', '测量', '身体', '活动']

3 126 ['足球', '球员', '统计', '球队', '团队', '俱乐部', '新闻', '网球', '得分', '教练']

4 1 ['降低', '记忆力', '人类', '下降', '兴趣', '性', '应', '导致', '注意力', '软件']

3.7 样本类别的标注方法设计

前文使用了无监督的特征评价指标进行了特征选择，并在特征选择的基础上进行聚类，对聚类结果进行手工标注，得到了符合文本相似度和App分类的背景知识的App分类体系。然而，在手工标注的过程中，保留了一部分按相似度区分不明显，但仍被分开的类别，比如类别1（汽车车辆）和类别2（赛车游戏）。因此，为了更加明显的区分出这些类别，首先进行有监督的特征选择，用手工标注后的分类体系作为类别标签，选择出具有类别区分度的特征。在特征选择的基础上，将未标注类别的样本按相似度划入已标注的类别，并将最后未标注的类别作为单独的“其它类别”处理。

3.7.1有监督的特征选择

卡方检验方法是常用的有监督的特征选择方法。在按词性和按DF指标过滤特征之后，按卡方检验的指标来过滤特征。仿照前文所述依据轮廓系数，选择最佳过滤特征比例。三次特征过滤数量如表3-17所示。从表3-18可以看到，通过卡方检验方法，较之无监督的特征选择方法，在评价指标上有了一定程度的提高，同时也说明了，带有类别信息的特征选择方法可以选择出更具有类别区分性的特征。

表3-17 三次过滤特征的特征数量统计表

过滤特征次数特征选择方法特征数量

第0次过滤特征中文分词，去停用词 52538

第1次过滤特征词性 n, v 39727

第2次过滤特征 DF [7, 1000] 7947

第3次过滤特征 Chi-square 2782

表3-18 有监督特征过滤方法和无监督的特征过滤方法对比

特征选择方法 K DBI 轮廓系数平均外部距离(outer_dist) 平均外部最近距离(outer_min) 平均内部平均距离(inner_dist) 平均内部最远距离（inner_max） 特征维度

无监督特征选择方法 28 2.204 0.049 0.884 0.677 0.675 0.946 2506

有监督特征选择方法 28 2.227 0.052 0.880 0.655 0.658 0.940 2782

3.7.2未标注样本的分布情况

称已标注的类别为候选类别。在划分剩余样本之前，统计样本到候选类别的距离分布。定义距离为：样本到类别内所有样本的余弦距离的平均值。

剩余样本到已标记类别的距离分布图3-7，其中，横轴为样本到最近的候选类别的距离。

图3-7 样本到已标记类别的最近距离分布情况

候选类别之间距离的方差的平均值分布如图3-8，其中，横轴为样本到最近的候选类别的距离。方差越大，说明候选类别之间的区分越明显。方差越小，说明候选类别之间的区分越不明显，难以按样本到候选类别的距离来正确划分样本。

图3-8 样本到候选类别距离的方差分布

综上所述，随着样本到候选类别距离的增大，样本到候选类别之间的区分越来越不明显，特征也逐渐变得稀疏。为保证标注样本的准确性，划分剩余数据应该设定一定的距离阈值，在低于阈值时，样本按距离划入候选类别；在高于阈值时，划入待定类别。

3.7.3样本类别的标注方法

在进行卡方检验的特征选择之后，按距离和原始类别信息来划分未标注的数据，使其符合文本相似度的同时，尽可能利用原始类别信息。

流程如下：

（1）对于每个没有划分类别的数据a，计算到每个候选类别的距离

（2）如果距离最近的类别大于阈值，则划入待定类别

（3）对于小于阈值的所有候选类别：如果候选类别中存在包含a的原始类别的类别，则划入距离最近的类别中；否则划入距离最近的候选类别

流程需要确定相似度阈值。以轮廓系数作为评价指标，选择使得样本数据整体的轮廓系数为最高的相似度阈值。此时样本数据的整体分类效果是较好的。

3.7.4样本类别的标注结果评价

使用上述样本类别的标注方法进行样本类别的标注。样本数据的总量为13151，在进行按相似度标注类别之前，已标注类别的数据量为8048，未标注类别的数据量为5103；在按相似度标注类别之后，总共标注的样本数量为10858，仍未标注类别的数据量为2293。

对于大于阈值的样本，由于到现有类别的距离都比较远，难以标注到现有的类别之中，因而可以将未划分类别的样本单

独作为一个“其它类别”。

仿照手工标注的类别描述，“其它类别”的描述如表3-19所示。从TF-IDF最高的TOP10单词特征来看，“其它类别”的TOP10 TF-IDF单词特征不在已标注的分类体系的TOP10 TF-IDF单词特征中，说明“其它类别”与已标注的类别具有较大的区分度，可以作为一个单独的类别。

表3-19 ‘其它类别’的类别描述

类别标号别名大小 TF-IDF TOP 10 来源

29 其它类别 2293 ['生活', '服务', '程序', '连接', '产品', '系统', '活动', '控制', '屏幕', '数据'] 各个原始类别

将“其它类别”作为单独的一个类别，与已标注的分类体系进行合并，作为最终的App分类体系。为评价带有“其它类别”的分类体系，计算相应的分类体系评价指标。如表3-20所示，包含了“其它类别”的样本分类体系仍然好于原始的分类体系。

表3-20 按相似度划分未标注数据后与原始分类体系的评价指标对比

分类体系 K DBI 轮廓系数平均外部距离(outer_dist) 平均外部最近距离(outer_min) 平均内部平均距离(inner_dist) 平均内部最远距离（inner_max）

原始分类体系 49 3.749 0.011 0.782 0.448 0.753 0.968

包含其它类别的分类体系 29 2.803 0.029 0.819 0.569 0.710 0.946

3.8 本章小结

本章提出了一种基于文本聚类的App分类体系的构建方法。在分析了原始分类体系的问题之后，结合文本聚类结果和原始分类体系，构建了新的App分类体系。从整体和局部两个角度对新的App分类进行评价，与原始分类体系相比，新的分类体系在评价指标上更合理，即同一类别内的样本数据更为集中，不同类别的样本数据之间距离更为分散。

在得到新的分类体系之后，提出了一种按相似度划分未标注样本类别的方法。使用该方法对未标注的样本进行类别标注，并将无法按相似度标注的样本划入“其它类别”。“其它类别”中样本与新的分类体系中的类别距离都比较远，因而可以将“其它类别”作为单独的一个类别，与新的分类体系中的类别进行合并。通过使用分类体系的评价指标进行验证，最终的样本分类好于原始的样本分类。

6. 第4章分类算法的训练与评价

总字数：5246

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第4章分类算法的训练与评价

4.1 引言

通过使用App描述文本，可以扩充App的相关信息，进而提高App分类系统的性能。本章将使用已构建的App分类体系，需要采用分类算法，训练App描述文本分类器，以达到自动化App分类的功能。

本章中首先介绍了三种文本特征的构建方法，包括传统VSM模型，文档Word2vec向量均值特征和VSM模型与Word2vec相结合的文本特征，然后使用构建的文本特征，输入文本特征至朴素贝叶斯分类器和SVM分类器中，训练分类器并对比分类器的性能。

4.2 文本特征的构建

对于基于统计的文本分类算法，文本特征的构建将会直接影响到文本分类器的性能。本节将介绍三种文本特征的构建方法，为下一步的训练分类器打下基础。

4.2.1 VSM向量空间模型

使用传统的VSM向量空间模型来构建文本特征。对于向量空间模型，有如下有监督的特征选择评价指标，其中C为类别标号，m为类别的数量，N为文档集中文档的数量。

（1）卡方检验（Chi-square）

卡方检验衡量了单词特征与类别之间的相关性。公式定义如下：ℵ2t,c=N×(pt,c×pt,c-p(t,c)×p(t,c))2p(t)×p(t)×p(c)×p(c) (4-1)ℵ2t=avgi=1m{ℵ2t,c} (4-2)

（2）信息增益（IG, Information Gain）

信息增益衡量了单词特征在出现与缺失情况下，对带有类别信息的文档集的熵的变化。公式定义如下：IGt=-i=1mpCilogpCi+pti=1mpCitlogpCit+p(t)i=1mp(Ci|t)log(p(Ci|t)) (4-3)

（3）基尼系数（Gini Index）

基尼系数衡量了单词特征与类别的相关程度。公式定义如下：GIt=i=1MptCi2p(Ci|t)2 (4-4)

其中，ptCi是类别Ci下出现单词特征t的概率，p(Ci|t)是出现单词特征t情况下属于类别Ci的概率。

使用上述方法，可以选择对类别最具区分性的单词特征，构建向量空间模型VSM作为文本特征。其中每一个文本对应一个空间向量，向量的每一维度对应一个单词特征。显然，这种文本特征的构建方法很容易造成较高的维度和较高的稀疏性，增加计算资源的消耗和分类器训练的难度。

4.2.2 Word2vec向量均值

Word2vec通过训练双层的神经网络，将文档中每一个单词映射到相同长度的向量中。每个单词用多个维度的向量表示，保留了单词与单词之间的相似关系和语义信息。本文使用每篇文档的特征为文档中所有的单词向量的和的均值。

4.2.3 结合VSM和Word2vec向量

论文[11]中提到，可以使用Word2vec训练得到的向量来降低VSM的稀疏性。Word2vec训练得到的单词向量具有相似关系，根据单词与单词之间的相似关系，可以对文本向量中为0单词特征进行权重填充。

具体方法为：

（1）对文本向量中的0项单词特征，寻找与之最相似的非0项单词特征；

（2）令该0项单词特征的权重等于最相似的非0项单词特征的权重乘以两者的相似度；

（3）重复（1）、（2）至文本向量中的所有0项单词特征全部被填充。

简言之，填充0项的思想为预测没有出现的单词特征，在假设出现时的TF-IDF权重。为避免出现维度过高的稠密向量，实际中往往可以选择相似度最高的TOP n项进行填充。

4.3 文本分类方法介绍

对于文本分类问题，使用较多的是朴素贝叶斯方法和支持向量机模型，本节将对朴素贝叶斯方法和支持向量机模型的基本原理进行简要介绍。

4.3.1 朴素贝叶斯法的基本原理

朴素贝叶斯法是一种基于概率模型的生成式学习方法，它使用训练数据集学习P(X| Y)和P(Y)的估计，从而得到联合概率分布，即：PX, Y=PYPXY) (4-5)

其基本假设是条件独立性，即影响类别的不同因素相互独立，即：PX=xY=ck=j=1nP(Xj=xj|Y=ck) (4-6)

显然，利用贝叶斯定理，在已学习到联合概率模型情况下，有：PYX=P(X,Y)P(X)=P(Y)P(X|Y)YP(Y)P(X|Y) (4-7)

因此，在进行分类时，后验概率最大的类y就是输入类别x的预测类别，即：y=argmaxckP(Y=ck)j=1nP(Xj=xj|Y=ck) (4-8)

4.3.2 支持向量机的基本原理

支持向量机是一种处理二分类问题的线性分类模型，通过使用核技巧(kernel)，使它成为实质上的非线性分类器。支持向量机的学习策略是求解分类间隔的最大化，其原始最优化问题为：minw,b12|w|2 (4-9)s.t. yiw•xi+b-1≥0, i=1,2,⋯,N (4-10)

求得最优化问题的解为w*,b*,可得线性可分支持向量机，分离超平面为w*•x+b=0 (4-11)

分类决策函数是fx=sign(w*•x+b) (4-12)

可以使用软间隔和核技巧来扩展线性可分支持向量机的使用范围。

支持向量机本身是一种处理二分类问题的分类器，但是对于多分类问题，可以使用一对多法(one-versus-rest, OVR SVMs)，一对一法(one-versus-one, OVO SVMs)和层次支持向量机来进行解决。

4.4 实验验证与性能分析

在本节中将应用特征选择方法，训练朴素贝叶斯分类器和支持向量机模型，通过实验对比，评价不同特征选择方法和分类器的性能。

4.4.1 文本特征构建

实验中使用App描述文本进行文本的特征构建。首先对App描述文本进行中文分词，去除停用词，在此基础上，使用如下文本特征构建方法。

1. VSM向量空间模型

依据词性和文档频率，对文本进行特征选择，并在此基础上，使用卡方检验方法，保留使得样本整体分类体系的轮廓系数最高的特征过滤比例。具体特征过滤信息如表4-1：

表4-1 特征选择信息

过滤特征次数特征选择方法特征数量

第0次过滤特征中文分词，去停用词 52538

第1次过滤特征词性 n, v 39727

第2次过滤特征 DF [7, 1000] 7947

第3次过滤特征 Chi-square 2782

2. Word2vec向量均值

实验中使用App描述文本和维基百科中文语料集两种语料来训练单词向量。

（1）App描述文本语料

使用现有App描述文本作为语料集，样本数量为97432。训练Word2vec采用skip-gram算法，窗口大小为5，忽略词频较低的单词，训练向量的维度为200维。

（2）维基百科中文语料

较大的数据集可以提高训练向量的精确度。维基百科中文语料集包含所有中文的维基百科词条文本数据，能够较全面的涵盖中文词汇。训练word2vec采用CBOW算法，窗口大小为5，忽略词频较低的单词，训练向量的维度为400维。

3. 结合VSM和Word2vec向量

实验中选择相似度TOP 30的0项单词特征进行填充，表4-2展示了相似度最高的TOP 5 零项单词特征，可以看到，对应的预测权重具有一定的合理性。

表4-2 相似度最高的TOP 5 零项单词特征信息

零项特征预测TF-IDF 最相似非零项特征最相似非零项特征TF-IDF 零项与非零项相似度

放大 4.053908 缩放 5.123037 0.79131

图片 2.813369 图像 3.792245 0.741874

素描 5.01107 草图 6.957969 0.720192

文字 3.070872 文本 4.381771 0.700829

网页 3.369233 浏览器 4.88221 0.690104

4.4.2 分类器设置

1. 朴素贝叶斯分类器实验设置

实验中使用多项式分布的朴素贝叶斯方法，使用VSM和结合Word2vec的VSM两种文本特征，输入至朴素贝叶斯分类器中。对数据集进行随机排序，并进行层次抽样，使数据集尽可能平衡。为避免实验偶然性，采用5折交叉检验的方法，对分类器性能进行检验。

2. SVM分类器实验设置

实验中使用一对一法，核函数选用线性核函数，对使用VSM、Word2vec单词向量和结合Word2vec的VSM的三种文本特征的数据集进行训练。与朴素贝叶斯的实验过程相同，对数据集进行随机排序，并进行层次抽样，使数据集尽可能平衡。为避免实验偶然性，采用5折交叉检验的方法，对分类器性能进行检验。

4.4.3 分类器实验结果对比

在进行实验性能对比与评价之前，首先定义评价指标。对分类器的评价指标一般选用精确率(Precision)、召回率(Recall)和F值(F1-Measure)。这里针对多分类问题，选用宏平均的评价指标定义。具体公式定义如下：Macro_P=1Ki=1KPi (4-13)Macro_R=1Ki=1KRi (4-14)Macro_F=1Ki=1KFi (4-15)

其中K为类别总数，P为类别精确率，R为类别召回率，F为类别F1值。

在分类器的训练过程中，使用构建好的App分类体系，包括“其它类别”。应用不同的文本特征和不同分类器进行训练，实验结果对比如表4-3所示。

表4-3 分类器实验结果对比

序号分类器特征构成特征维度平均非零维度 Macro_P Macro_R Macro_F

1 SVM VSM 2782 26.167 0.786 0.774 0.778

2 SVM VSM（L2 Norm） 2782 26.167 0.846 0.825 0.832

3 SVM word2vec（1） 200 N/A 0.772 0.679 0.688

4 SVM word2vec（2） 400 N/A 0.723 0.720 0.719

5 SVM VSM+word2vec（1） 2782 56.163 0.836 0.819 0.824

6 NB VSM 2782 26.167 0.785 0.789 0.792

7 NB VSM（L2 Norm） 2782 26.167 0.818 0.777 0.773

8 NB VSM+word2vec（2） 2782 56.163 0.809 0.766 0.763

表4-3中，Macro_P，Macro_R和Macro_F为五折交叉检验结果的平均值；Word2vec（1）是由App描述文本数据集训练得到的向量；Word2vec（2）是由维基中文语料集训练得到的向量。L2 Norm为向量归一化；VSM+word2vec（1）是使用维基中文语料集的word2vec，与VSM进行结合，并进行向量归一化；VSM+word2vec（2）是使用维基中文语料集的word2vec，与VSM进行结合，且不做向量归一化。

不同分类器在宏平均下的准确率、召回率和F值的对比如图4-1所示。

图4-1 不同分类器在宏平均下的准确率、召回率和F值对比

从实验结果对比中可以看到：

（1）L2 Norm对比：对于SVM分类器，归一化要好于未归一化；对于NB分类器，未归一化要好于归一化。（1和2、6和7）

（2）分类器对比：使用SVM分类器和NB分类器的最好情况进行对比，SVM优于NB。（2和6）

（3）Word2vec与VSM对比：以Word2vec均值为特征的分类器性能差于以VSM为特征的分类器性能，但是特征维度上远低于VSM，计算速度上快于VSM。（2和4）

（4）VSM+Word2vec和VSM对比：使用Word2vec相似度来降低VSM稀疏程度的方法，并没有在分类器性能上表现出明显提升；原因可能在于在填充0维度之前，文本向量已经有明显的相似关系，在填充0之后可能引入了噪声，造成了分类器性能的降低。（2和5、6和8）

（5）word2vec对比：维基中文语料好于App描述文本训练出来的vector，原因可能是较大的数据集训练出来的vector准确度高于小数据集。（3和4）

综上所述，使用归一化VSM模型的SVM分类器在分类性能上达到了最好的效果。

4.5 本章小结

本章设计了多种基于文本分类技术的App分类器，并对不同的分类器进行了性能对比。首先，使用三种特征构建方法对App描述文本进行特征提取，包括传统VSM特征，文档Word2vec向量均值特征和VSM模型与Word2vec相结合的文本特征。然后，将文本特征输入至朴素贝叶斯分类器和支持向量机分类器中，训练得到文本分类器。最后，通过五折交叉检验对分类器进行了性能评价。通过实验验证，使用归一化VSM模型的SVM分类器达到了最好的分类性能，能够对App的描述文本进行有效的自动化分类。

7. 第5章分类系统原型的设计与实现

总字数：815

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第5章分类系统原型的设计与实现

5.1 引言

第3章使用App描述文本构建了App分类体系，并对App样本数据进行了类别标注。第4章使用多种文本特征，建立了不同的App文本描述分类器。本章在第3章和第4章基础上，设计并实现App分类系统的原型。

5.2 分类系统原型设计与实现

综合前文实验方法和结果，可以利用现有App文本描述数据集训练分类器，设计和实现App分类系统原型。为便于比较预测，分类系统原型使用不同的分类器和不同的输入特征，分别对输入文本进行预测。

App分类系统原型的运行流程如下：

1. 加载词典和分类器

在读入数据之前，首先加载中文词典和分类器和一些预设置，便于后续中文分词、特征化和分类器的使用。

2. 读入App描述文本

读入App描述文本，限制App描述文本长度。对于过短的App描述文本，可能难以做出有效的分类。

3. 描述文本特征化

对读入的文本进行中文分词，去除停用词，并进行相应的特征化。

4. 特征输入分类器

将特征化的文本输入到训练后的分类器中。

5. 输出分类结果

分类器对文本特征进行分类，输出预测的类别。

系统原型实现中，运行环境为：windows 10操作系统，python 3.6.3编程语言，使用pickle做分类器模型的持久化。

5.3 分类系统原型运行效果展示

如图5-1所示，读入任意App描述文本，在系统后台自动进行相应的特征提取过程，并输入至不同的分类器。可以看到，不同的分类器对文本的预测类别结果不尽相同。对比之下，使用向量空间模型特征的SVM模型的预测结果更符合于对App的认知。

图5-1 App分类系统原型分类示例

5.4 本章小结

本章利用App描述文本扩展App的相关数据，使用构建的App分类体系和多种App分类器，设计并实现了简易的App分类系统模型。可以看到，该分类系统对App描述文本具有一定的分类能力。

8. 第6章总结与展望

总字数：1047

相似文献列表　文字复制比：8.1%(85)　疑似剽窃观点：(0)

1	150801202708+黄家玉+会计1527 -《大学生论文联合比对库》- 2017-04-01	7.8%（82） 是否引证：否
2	150801202708+黄家玉+会计1527 -《大学生论文联合比对库》- 2017-04-18	7.4%（78） 是否引证：否
3	12132612 孙佳月 孙佳月 -《大学生论文联合比对库》- 2017-04-17	6.7%（70） 是否引证：否
4	20130808515_康丽_静海区县域经济发展问题研究 康丽 -《大学生论文联合比对库》- 2017-05-23	4.1%（43） 是否引证：否
5	程家鑫_S201204S201041_财政金融学院_传统产业转型升级下的商业银行信贷风险的控制 程家鑫 -《大学生论文联合比对库》- 2016-05-02	3.8%（40） 是否引证：否

6	马文韬_201213010209_旅游管理学院_山海关古城旅游形象塑造思考	3.8%（40）
	马文韬 - 《大学生论文联合比对库》- 2016-05-04	是否引证：否
7	程家鑫_201213010252_旅游管理学院_晋商文化旅游感知度研究————基于太原市民的调查	3.8%（40）
	程家鑫 - 《大学生论文联合比对库》- 2016-05-04	是否引证：否
8	程家鑫_S201204S201041_财政金融学院_传统产业转型升级下的商业银行信贷风险的控制	3.8%（40）
	程家鑫 - 《大学生论文联合比对库》- 2016-05-08	是否引证：否
9	陕西富盛科贸有限公司存货管理中存在的问题与对策	3.8%（40）
	张云 - 《大学生论文联合比对库》- 2016-05-26	是否引证：否
10	议程设置与受众兴趣的联系	3.8%（40）
	黎子宁 - 《大学生论文联合比对库》- 2017-05-13	是否引证：否
11	金融学院-13050503-曹玉燕	3.8%（40）
	金融学院 - 《大学生论文联合比对库》- 2017-05-19	是否引证：否
12	基于Android系统的线上房产中介平台的设计与实现	3.8%（40）
	叶志超 - 《大学生论文联合比对库》- 2017-05-20	是否引证：否
13	水田土壤有效锌含量空间分布特征及影响因素分析	3.8%（40）
	郝秋杉 - 《大学生论文联合比对库》- 2017-05-21	是否引证：否
14	11614210_郭小培_对于我国商业银行呆账现状的分析及对策研究	3.8%（40）
	郭小培 - 《大学生论文联合比对库》- 2017-05-30	是否引证：否
15	150801202719+吴春花+会计1527	3.8%（40）
	- 《大学生论文联合比对库》- 2017-04-01	是否引证：否
16	130401494421 +王洁丽+ 金融1344	3.8%（40）
	- 《大学生论文联合比对库》- 2017-04-12	是否引证：否
17	150801202719+吴春花+会计1527	3.8%（40）
	- 《大学生论文联合比对库》- 2017-04-18	是否引证：否
18	论宋词中的杏花意象	3.8%（40）
	冯晓彤 - 《大学生论文联合比对库》- 2017-05-02	是否引证：否
19	电子商务的发展对我国经济增长的作用分析	3.8%（40）
	程文英 - 《大学生论文联合比对库》- 2017-05-06	是否引证：否
20	网络文学的现实影响与发展方向	3.8%（40）
	赵玉麒 - 《大学生论文联合比对库》- 2017-05-18	是否引证：否
21	节能照明系统控制部分的研究	3.8%（40）
	易乾 - 《大学生论文联合比对库》- 2017-05-19	是否引证：否
22	12131314陈洛瑶	3.7%（39）
	陈洛瑶 - 《大学生论文联合比对库》- 2017-04-14	是否引证：否
23	人格特质、职业类别、财富观与个人理财之研究	3.7%（39）
	严振阳 - 《大学生论文联合比对库》- 2017-05-04	是否引证：否
24	人格特质、职业类别、财富观与个人理财之研究	3.7%（39）
	严振阳 - 《大学生论文联合比对库》- 2017-05-10	是否引证：否
25	12131314陈洛瑶论文	3.7%（39）
	12131314陈洛瑶论文 - 《大学生论文联合比对库》- 2017-05-15	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第6章总结与展望

6.1 文本工作总结

研究移动应用App分类，有助于进一步分析用户行为，对推荐系统的设计也有着重要的意义。然而在App分类研究中，存在着如下问题：没有明确公认的App分类体系；App原始类别标注的不准确；App信息过少，难以直接用来解决分类问题；App描述文本的稀疏性和高维性，难以对文本进行准确的分类。

本文使用App描述文本来扩充相关的文本数据，便于App分类体系的构建和App分类器的训练。为了构建合理的App分类体系，使用文本聚类方法，对描述文本进行聚类，结合聚类结果和原始的App分类体系，重新构建新的App分类体系。与原始分类体系相比，新构建的App分类体系更符合文本相似度的类别划分，即同一类别内文本相似度较高，而不同类别间文本相似度较低。同时，设计了一种按相似度标注App类别的方法，对未标注类别的App文本描述进行类别标注。

在构建了App分类体系的基础上，提取多种文本特征，设计并实现App描述文本分类器。通过实验验证，基于VSM模型的SVM文本分类器具有较高的分类性能，可以用于实现App分类系统原型。

6.2 未来工作的展望

本文使用App描述文本数据，构建了较为合理的App分类体系，和具有较好性能的App分类器。然而，本文仅仅使用了

App描述文本数据，并没有使用App名字和App包名等数据，以及应用商店中的App评论数据。下一步工作将结合App名字、App包名等App自身的文本数据，及App应用商店中的评论等文本数据，进一步扩充App相关数据，提高分类器的性能。此外，将使用基于深度学习的文本分类方法，进一步提高App分类的准确性。

参考文献

[1] Phan X H, Nguyen C T, Le D T, et al. A Hidden Topic-Based Framework toward Building Applications with Short Web Documents[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(7):961-976.

[2] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in Proc. WWW, Edinburgh, U.K., 2006, pp. 377–386.

[3] Metzler D, Dumais S, Meek C. Similarity Measures for Short Segments of Text[C]// AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. 2007:1489.

[4] H. Ma, H. Cao, Q. Yang, E. Chen, and J. Tian, “A habit mining approach for discovering similar mobile users,” in Proc. WWW, Lyon, France, 2012, pp. 231–240.

[5] Zhu H, Chen E, Xiong H, et al. Mobile App Classification with Enriched Contextual Information[J]. IEEE Transactions on Mobile Computing, 2014, 13(7):1550-1563.

[6] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. 2014, 4:II-1188.

[7] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[8] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information[J]. 2016.

[9] Liu T, Liu S, Chen Z, et al. An evaluation on feature selection for text clustering[C]// Twentieth International Conference on International Conference on Machine Learning. AAAI Press, 2003:488-495.

[10] Uysal A K. An improved global feature selection scheme for text classification[M]. Pergamon Press, Inc. 2016.

[11] Yao D, Bi J, Huang J, et al. A word distributed representation based framework for large-scale short text classification[C]// International Joint Conference on Neural Networks. IEEE, 2015:1-7.

[12] Y. Yang, J. Pedersen, A comparative study on feature election in text categorization. International conference on Machine Learning(ICML), 1997.

[13] Berardi G, Esuli A, Fagni T, et al. Multi-store metadata-based supervised mobile app classification.[J]. 2015:585-588.

[14] Lindorfer M, Neugschwandtner M, Platzer C. MARVIN: Efficient and Comprehensive Mobile App Classification through Static and Dynamic Analysis[C]// IEEE, Computer Software and Applications Conference. IEEE Computer Society, 2015:422-433.

[15] Loet. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index[J]. Journal of the Association for Information Science & Technology, 2014, 59(1):77-85.

[16] Christiani T, Pagh R. Set similarity search beyond MinHash[C]// ACM Sigact Symposium on Theory of Computing. ACM, 2017:1094-1107.

[17] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038-2048.

[18] Dong L, Wesseloo J, Potvin Y, et al. Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression[J]. Rock Mechanics & Rock Engineering, 2016, 49(1):183-211.

[19] Chen C H. Improved TFIDF in big news retrieval: An empirical study[J]. Pattern Recognition Letters, 2016, 93.

[20] Qin P, Xu W, Guo J. A novel negative sampling based on TFIDF for learning word representation[M]. Elsevier Science Publishers B. V. 2016.

[21] Joachims T. Transductive Inference for Text Classification using Support Vector Machines[C]// Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1999:200-209.

[22] Mccallum A. A comparison of event models for naive bayes text classification[C]// Proc. AAAI-98 Workshop on Learning for Text Categorization. 1998:41--48.

致谢

首先要感谢我的论文指导老师，吉林大学计算机科学与技术学院的刘小华老师，对我的毕业设计提供帮助。

特别感谢中国科学院信息工程研究所的客座导师、国家计算机网络与信息安全管理中心的王丽宏老师，王老师为我的毕业设计提供了良好的工作条件，提供了很多指导，同时创造了很多机会和空间让我学习和成长，付出了很多心血与精力。感谢国家计算机网络与信息安全管理中心的刘婧老师，在毕业设计中提供了帮助与支持，并在撰写论文的过程中提出了许多有益的改善性意见。感谢中国科学院信息工程研究所的王斌老师，在我的毕业设计中提供帮助。感谢实验室的钟盛海师兄、顾杰师兄、郭杰师姐、黄洪仁师兄和王士承师兄等师兄师姐对我的帮助与支持。

最后，衷心感谢我的家人、朋友，以及同学们，在他们的鼓励和支持下我才得以顺利完成此论文。

指　标
疑似剽窃文字表述

1. 致谢
   首先要感谢我的论文指导老师，吉林大学计算机科学与技术学院的刘小华老师，对我的
2. 最后，衷心感谢我的家人、朋友，以及同学们，在他们的鼓励和支持下我才得以顺利完成此论文。

说明：1.总文字复制比：被检测论文总重合字数在总字数中所占的比例

2.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例

3.去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例

4.单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比

5.指标是由系统根据《学术论文不端行为的界定标准》自动生成的

6.红色文字表示文字复制部分;绿色文字表示引用部分

7.本报告单仅对您所选择比对资源范围内检测结果负责