Viiision

APP评论数据分析系统

目标与服务模型

ComVision 团队

目录

1. 项目目标

1.1. 项目背景

1.1.1. 公司背景

虹软公司是全球领先的专业计算摄影与计算机视觉技术公司。创建于 1994 年，总部设在美国加利福尼亚硅谷，同时在欧洲、东京、首尔、台北、上海、杭州、南京都设有区域性的商业与研发基地。

2004 年，虹软捕捉到手机相机市场蕴藏巨大潜力，开始专注于手机平台的影像处理和拍摄技术，成为世界上最早进入移动领域的传统影像软件公司。之后，随着主流手机平台的发布，虹软基于一向专注的多媒体、图像等领域开展了深度研发，技术成果随着巨量移动设备进入到消费者手中。并随后在竞争日益激烈的智能手机市场环境中，凭借在计算摄影与计算机视觉领域的研发领导地位，集中开发了一系列独特的技术和产品，稳固地帮助各大厂商、互联网公司建立了差异化的产品和良好的口碑，成为了他们的最佳拍档。

多年来，一直专注于计算摄影与计算机视觉技术领域的虹软公司，结合市场需求并引领技术趋势，不断自主研发和创新，拥有此领域强势的核心技术能力，已为全球数十亿台的硬件产品提供了解决方案，给全球消费者带来了更好的用户体验和真正的价值。

1.1.2. 技术背景

Viiision—APP 评论数据分析系统是一个集爬虫技术、分词技术、关键词提取、机器学习与模式匹配、数据可视化技术等新一代信息技术为一体的多样性系统。致力于通过对评论的处理和分析，将统计分析后的重要信息通过多类型多维度的图表方式直观地呈现给用户，使得用户能够方便快捷地获取所需信息。

1、爬虫技术

随着网络的高速发展和评论的自由化，越来越多的用户会在不同平台上对 APP 进行评论，面对评论信息相对分散而冗杂的情况，通过爬虫技术获取评论信息已成为数据挖掘工程师的首要选择。本项目 Viiision—APP 评论数据分析系统根据用户所需，通过爬虫技术实时地从各大应用商店中将用户对 APP 的评价抓取过来，进而对评论数据进行解析，结构化后保存到数据库中，作为后续数据统计分析的重要来源。

2、分词技术

用户往往会在各大应用商店和论坛发表自己使用 APP 后的感受和想法，而这些感受来自用户的直接感受，对于企业来说相当重要。而如何更好地从评论中获取重要信息取决于对其的分词处理程度。Viiision—APP 评论数据分析系统通过调用包含两万多条词条和词性的词典，借用结巴分词器对评论进行分词处理。

3、关键词提取技术

本系统选取 TF-IDF 算法作为关键词选取的主要算法，该算法在搜索引擎等实际应用中广泛使用，主要用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

4、机器学习算法与模式匹配

机器学习分类方式具有训练简单，分类精度高的优点；缺点主要是对训练集数据的数量和质量有严格的要求。当训练集数量不全面或代表性不强时，将严重影响分类效果。而基于模式匹配的方式采用对评论进行情感特征词的匹配的方式进行情感分析。因此对于评论极性不明显的方式采用该方法比较有效。本系统对产品特征进行情感分析时，采用两者相结合的方式以提高情感分析的准确性。

5、数据可视化技术

在这个大数据时代，具有丰富的交互功能及更多的可视化效果的图表越来越受到人们的欢迎。近些年，数据可视化技术也在迅猛发展，数据可视化是利用计算机图形学的图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。在本系统中，主要是将分析后的数据呈现给用户，如何充分地利用数据可视化技术就显得尤为重要。本项目组致力于通过多样式多维度的图表呈现方式使用户能够更加快捷、更加清晰地了解到相关信息。

1.1.3. 业务背景

随着智能机的普及，APP 的开发和使用也层出不穷；同时加之网络的迅猛发展，拓宽了用户表达自我意见的渠道。对于每一款手机应用，都会在不同的平台做发布，用户会在各个平台评论，这些评论因为来自用户的直接感受，显得十分重要。这些评论信息不仅能让用户了解现在的 APP 市场，而且更重要的是能够为产品决策者和开发者提供宝贵的反馈信息，精确地获取到对企业有用的信息以便改进产品的质量。然而，因其分布范围较为广泛，且人们不可能逐条阅读海量的用户评论，因此，一款基于评论的分析系统显得尤为重要。

近几年，也陆续出现了多款基于评论的分析系统，如：WeTest 腾讯质量开放平台、ASO114、Cobub 等，他们大多更加偏重于数据的丰富度，而在数据的统计分析、呈现方式、结果反馈方面并没有进行深入的研究与优化。而本系统旨在给公司决策者和开发者提供反馈信息使得他们能够精确获取信息并作出相应的调整，因此对数据的深入分析、直观呈现及结果反馈是本系统的重点研究方向。

为了更好地迎合企业的需求，为企业提供个性化服务，我们团队致力于 APP 评论数据分析系统的解决和完善。

1.2. 项目目标

- **系统的适配性**

系统在电脑端和移动端都能正常访问，通过加强用户交互以带给用户更好的用户体验，满足用户随时随地查看信息的需求。

- **系统的高效性**

实现用户在对相关数据进行导入时，占用资源较少，并同时保证查询页面依然可以正常访问，多人可以同时流畅地在线使用。

- **评论筛选的自定义性**

用户可根据时间颗粒度、评论好坏程度、应用商店、评论关键字等有选择性地查看相关的评论详情。

- **数据呈现的直观性**

用户可以根据自己的需求方便快捷地从平台上获取相关的信息，且信息的呈现方式直观，易于被广大用户接纳。

- **问题反馈的针对性**

企业内各部门职能不一，各部门期望获得有针对性的问题反馈，以便更好地提出后续的解决方案。

- **用户使用的个性化**

用户可对关注的 APP 进行增加和修改，可根据需求对不同的 APP 进行增加分词的操作并在下次查询时优先显示。

- **管理员管理的简易性**

为管理员设置本地导入和网上导入功能，使其可以方便快捷地对数据进行导入和管理。

2. 项目服务模型

2.1. 可行性服务模型

2.1.1. 技术可行性

数据分析型业务的痛点在于如何从海量的数据中找到用户所需的信息并以一种直观的方式展现给用户。因此对于数据源的要求是分类明确并且获取平台多样，对于呈现方式的要求是能让用户第一眼便看到自己想要的数据。由此可见对于信息的获取及筛选和筛选结果呈现方式的选择尤为重要。

对于信息的获取和筛选，我们通过爬虫从各大应用平台获得相应的评论，结构化后保存到数据库中，并通过分词和关键词提取技术进行产品特征的提取，再结合情感分析，最后将分析结果以邮件的形式推送给不同的部门，使得其可根据获得的信息对产品进行改进和完善。

对于信息的展示，我们选用了 Echarts，通过动态可交互的柱状图，曲线图，地区分布图、词云图等来个性化地展示统计分析结果。

2.1.2. 社会环境可行性

在当今的大数据时代，其战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

智能手机的发展使得所有的工具模块都以 Application 的形式存在，更甚包括蒂姆·库克在发布 Apple TV 时，也表示 Apple TV 的未来是 APP。可见在数码设备上，APP 的地位是越来越稳定的。在这样的条件下，对于 APP 及时的反馈就显得尤为重要。

Viiision 这个平台将用户对 APP 全面的反馈信息进行筛选和总结，以多类型的图表形式呈现，为 APP 的良性发展提供针对性的建议。

2.1.3. 政策可行性

2015 年 7 月 4 日，国务院印发《国务院关于积极推进“互联网+”行动的指导意见》，政府大力支持“互联网+”产业。Viiision 通过数据可视化为移动应用的良性发展开辟新的道路，为“互联网+”奉献自己的一份力。

2.2. 需求分析模型

企业的目的在于针对大量用户下载和评论的操作，经过数据挖掘来获取信息，从而为自己已有的产品提供改进思路并为新的产品提供构思。

第一阶段：我们对企业提供的命题介绍文档进行了需求分析，拟定需求分期初始文档。

第二阶段：项目组成员进行会议，汇总出命题中的模糊点，通过 E-mail 和打电话的方式与企业进行沟通，进而获得一个更明确地产品定位。

第三阶段：针对已确定的产品定位和需求分析，将需求分析细化，编撰需求分析总文档，并在实现的过程中根据项目进程对其进行恰当的增删改。

3. 项目价值

3.1. 项目难点分析

3.1.1. 冗杂评论数据的获取

每一款手机应用，都会在不同的平台做发布，因此不同地区的用户都会在各个平台评论不同应用的使用感受，而且评论数据每天都在更新，日积月累形成的数据总量将会非常庞大。本系统需要解决的问题就是如何去获取这些冗杂的评论数据，进而对数据进行预处理，并将其结构化地保存到数据库中以方便后续的统计分析。

3.1.2. 评论关键词语的提取

本系统的核心功能是从海量评论数据中为用户提取有用信息。因此关键词提取效果的好坏决定了该系统的成功与否。本系统需要解决的就是对评论进行分词，在分词结果中去获取对用户最有用的词语。

为不同需求的用户提供差异化的分析服务也是提取关键信息时的难点之一。不同的部门之间的需求既存在差异点又有共通之处，在提供关键词分类时需要权衡需求上的差异与共性。

3.1.3. 产品特征的用户情感分析

用户在发表评论时，一句评论中往往涉及多种产品特征的评价，且评论中普遍存在用户态度极性不突出的现象。如何从一句评论中分析出用户对不同产品特征的情感直接决定了本系统对用户反馈信息的正确与否。本系统需要解决的就是通过对评论进行情感分析，进而精确获取不同用户对不同产品特征的态度。

3.2. 项目优势分析

3.2.1. 技术优势

3.2.1.1. 实时的海量数据获取

对于来自不同平台的数据进行数据解析和结构化存储。同时采用分布式爬虫技术实现并行爬取操作，提高数据获取效率。强大的爬虫框架实现针对不同平台的自适应配置，保证管理员通过简单配置平台地址就可获取各大平台评论数据，提高操作的便捷性。

3.2.1.2. 准确可靠的评论数据分析算法

平台在评论文本分析的不同阶段根据数据分析的需求采用了最匹配最高效的算法，主要体现在以下几个方面：

1. 通过基于 Trie 树结构的词典扫描生成有向无环图和基于动态规划查找最大概率路径的分词算法实现评论文本的准确分词。

2. 采用去重和垃圾过滤算法解决评论数据分析的噪声问题。

3. 通过基于 TF-IDF 算法实现评论关键词提取算法，同时针对不同的用户需求改进了分类词典，保证关键词提取的针对性。

4. 使用模式匹配和机器学习算法相结合的方式完成评论中产品特征情感极性的自动判定，帮助用户区分好评差评。

3.2.1.3. 流畅友好的前端交互技术

本系统在前端交互技术的实现上主要有以下几个优势：

1. 采取结构统一、色彩均衡的界面设计提高用户的感官体验。

2. 使用响应式的页面布局保证用户能在不同尺寸的设备上获取到最合适的页面呈现效果。

3. 使用 AJAX 异步加载数据的方式避免重刷页面导致的用户体验下降。

4. 提供形式多样的条件筛选方式，满足不同用户希望获取不同粒度、不同方面信息的交互需求。

3.2.2. 应用优势

3.2.2.1. 个性化+动态化的管理平台

本系统在功能的设计上充分考虑不同用户的不同需求，在以下几个方面为用户提供个性化的服务：

1.用户可以打开关注内容模块对关注的 APP 进行查看和分组管理，也可新增需关注的 APP。

2.针对某一 APP，用户可根据自己的职能需求在原有分词词库的基础上添加或修改分词，从而获得贴合自身需求的关键词筛选结果，体现了评论分析的个性化和动态化。

3.管理员可对用户组别、用户查看权限、不同类别下的分词词库进行更改和编辑。

3.2.2.2. 精确的产品特征情感分析平台

利用关键词提取技术对评论中的产品特征进行提取，进而利用模式匹配和机器学习相结合的算法将评论自动分为好评和差评，用户可方便地查看产品特征的好评差评详情及其百分比，有助于用户更加直观地了解产品特征的用户情感变化。

3.2.2.3. 及时的统计分析反馈平台

为满足企业及时纠正产品问题，把握市场商机的需求，本系统根据用户所需，定时分析整理用户关注的内容，并将其关注 APP 每天新增下载量、新增评论数、产品特征用户态度变化、同类 APP 的比较结果以邮件的形式推送。

3.2.2.4. 直观的分析结果展示平台

本系统对爬取的大量评论数据进行分析处理，将处理结果以多类型的图表形式呈现，且在同一图表中反馈多元化的内容，我们采用的图表主要有如下几种形式：

1.词云图，该图集某一时间段内热词排行榜、某一热词词频随时间变化趋势、不同日期的热词分布这三种信息为一体，且用户可以通过点击相关热词查看包含该热词的好评与差评评论。

2.柱状图与曲线图，将下载量和评论量这两个信息结合在一张图中，呈现其按照时间的变化情况。

3.区域分布图，根据区域分布图中的颜色变化可以直观地了解 APP 在全球范围内的分布。

4. 项目解决思路

4.1. 总体思路

对于 Viiision—APP 评论数据分析系统，我们分四个阶段寻求系统的解决思路：

1) 分析阶段，通过对项目要求的详细分析和外包服务业务的背景学习，明确项目的基本目标。通过社会调研以及与发包方的沟通交流，深入了解需求。

2) 挖掘阶段，在需求分析的基础上，挖掘项目的核心价值，并对每一个价值点进行成本分析和风险评估以及可行性分析，进而设计出可行的技术路线。

3) 整合阶段，充分整合分析出具体的技术路线，提出宏观的系统架构。

4) 结题阶段，根据整合得到的成本分析，服务模型以及相应的技术路线设计系统最终完成解决方案的撰写。

提交解决方案，通过分析审核后进行项目的实施

4.2. 具体做法

Viiision—APP 评论数据分析系统主要面向各个企业内部，同时需要管理员定期管理后台，因此根据使用人群的不同将该系统分为用户平台和管理员平台，用户平台主要是将统计分析后的重要信息以多类型图表形式呈现给用户，管理员端主要是方便管理员对用户信息、分词详情、评论信息的管理，具体功能如下图所示：

图 1 系统功能图

用户端具体功能模块设计如下表所示：

表 1 用户端功能设计

用户端功能设计	
APP 概况浏览功能	用户可根据时间颗粒度查看关注的 APP 评论近况，评论近况包括 APP 下载量变化、评论数增长变化、喜好程度或态度变化等
APP 版本比较功能	将比较信息于同一图表中呈现，进而达到比较的目的
APP 地区比较功能	通过区域分布图，用户可通过查看颜色深浅了解 APP 分布情况
不同应用商店比较功能	用户可根据所需，了解 APP 在不同的应用商店中的反馈情况
自定义条件筛选评论功能	用户可按照时间、应用商店、关键词等搜索条件，查询相应的评论详情
评论分词统计功能	用户可通过词云图查看其关注分词的变化
评论分词编辑功能	用户可根据所需编辑评论的拆分分词内容，在下次查看时优先显示
个性化选择功能	用户可方便快捷地对关注的 APP 进行管理，并且可为不同的 APP 添加不同的分词
及时的问题反馈功能	定期整理分析用户关注的内容，并以邮件的形式推送
移动端适配功能	使网页适配各种屏幕大小，满足用户随时随地查看信息的需求并提高用户体验

管理员端具体功能模块设计如下表所示：

表 2 管理员端功能设计

管理员端功能设计	
编辑用户信息功能	管理员能够方便、快捷、高效地对用户进行增删改、分组
管理用户权限功能	管理员对不同的用户设置不同的权限，并分配可以查看的 APP 信息
评论本地导入功能	管理员可将本地整理好的评论详情方便地

	的导入到系统中
评论自动导入功能	管理员可通过输入 APP 名称并配置获取地址实现从网络平台自动获取数据
自动分词归类功能	能够自动对获取的评论内容进行解析，并实现分类
添加分词功能	管理员可对不同 APP、不同分词类别添加分词，以便下次查询优先展示
按颗粒展示功能	管理员可根据所需查看不同颗粒度下的评论详情