APP评论分析系统项目

--技术路线及实现方案--

创建日期：2017-10-27

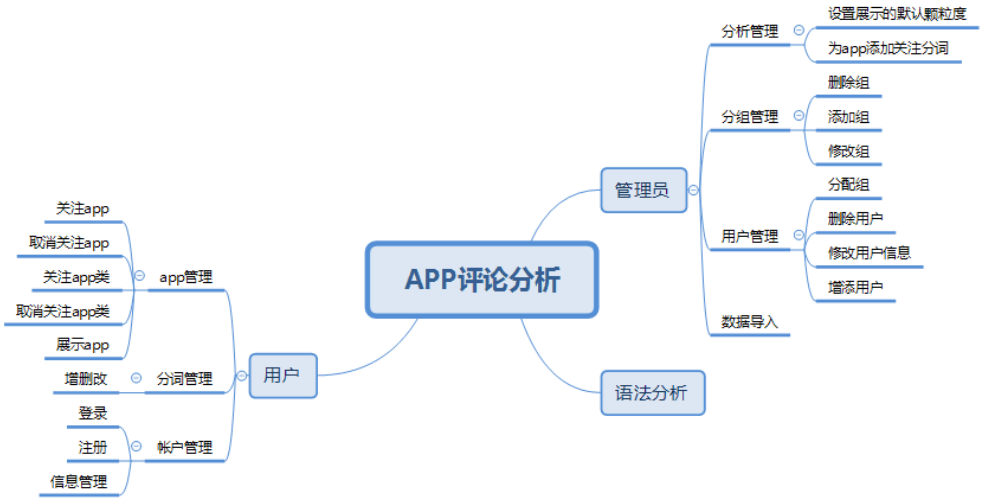
最后修订日期：2017-11-29

作 者：	技术负责人
日 期：	2017-11-5
审 核：	技术负责人
日期：	2017-11-29

目录

1. 功能模块

1.1. 功能模块架构



1.2. 功能模块设计

1.2.1.用户端

关注 APP 展示模块

本模块主要展示用户所关注 APP 的总体情况，在该模块中，用户查看评论数、评论增长、下载量等信息在不同时间段内的变化趋势

APP 搜索模块

本模块主要为用户提供 APP 搜索功能，用户可以对搜索得到的 APP 关注、查看详细信息。

APP 评论搜索模块

本模块主要为用户提供对某个特定的 APP 评论的搜索功能，用户可以根据系统提供的分词快速查询评论，并且用户可以对分词进行编辑。

版本比较模块

将同款 APP 不同版本间的相关信息放在一起进行比较，以图表的形式呈现给用户。用户还可以将图表下载下来。

分词展示模块

以三维词云图的方式展现一段时间内热词的分布情况、展现每一个分词词频数随时间的变化趋势、展现一天中不同词的词频比较。

选择查询模块

在该模块中，用户可以根据所需按时间段、按星级、按 APP 名称、按相关关键词、按平台名称对 APP 评论进行查询。

评论展示模块

将评论内容、评论星级、评论时间等评论详情以多种表格的形式呈现出来，且采用分页技术。

个人信息管理模块

在该模块中用户修改自己的相关信息，并且取消关注 APP。

分词管理模块

用户可以在管理员提供分词的基础上对分词做个性化的编辑。

1.2.2.管理员端

- **用户信息展示模块**

系统以列表形式展示用户信息。

- **权限管理模块**

本模块主要有管理员对用户进行分组管理，以实现用户可以关注的 APP 的权限

- **本地导入模块**

通过点击相应按钮，选择本地已经整理好的 Excel 内的评论内容，导入到数据库，并可以通过评论展示模块显示出来。

- **APP 详情模块**

该模块与用户功能描述的模块功能一致。

- **分词模块**

在该模块中，用户可以为不同的 APP 分配不同的分词词库，并可随时根据所需增加分词，为不同的分词规定优先级。

- **分类管理模块**

该模块是分词管理的基础，将不同类型的 APP 进行分类管理，也可以通过点击相关按钮进行增加或删除 APP。

2. 爬虫技术

2.1. 爬虫介绍

为了获取大批量的 APP 评论数据，方便管理员进行工作，我们为管理员提供爬虫技术支持，将数据过滤后存入数据库。

2.2. 页面解析

2.2.1.使用语言

我们选择 Python 作为爬虫语言。该语言利用正则表达式，能够简洁方便地进行爬虫。

2.2.2.数据来源分析

评论的数据来源主要可分为官方 API、静态页面中的数据、应用 AJAX 技术实现异步加载的动态数据。

2.3. 爬虫

采用 request 这个 HTTP 客户端库的相关函数对网页数据进行解析爬取，再采用 xlwt 这个库的换书对爬取的结构化数据进行 Excel 表格的写入 ，为方便 mysql 数据库的操作，对 emoji 表情进行了过滤。将过滤后的结构化数据存储 mysql 数据库中。

3. 算法技术

3.1. 分词算法

3.1.1. 分词器的选择

本项目德鲁伊的翡翠梦境—APP 评论数据分析系统采用结巴(Jie Ba)分词技术，结巴分词是国内程序员用 Python 开发的一个中文分词模块，它的特点有：

1) 支持三种分词模式：
   - 精确模式，试图将句子最精确地切开，适合文本分析；
   - 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
   - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。
2) 支持繁体分词
3) 支持自定义词典

3.1.2.分词过程

1) 加载字典，生成 Trie 树；
2) 给定待分词的句子，使用正则获取连续的中文字符和英文字符，切分成短语列表，对每个短语使用 DAG(查字典)和动态规划，得到最大概率路径，对 DAG 中那些没有

在字典中查到的字，组合成一个新的片段短语，使用 HMM 模型进行分词，也就是识别新词，即识别字典外的新词；

3) 使用 python 的 yield 语法生成一个词语生成器，逐词语返回。

3.1.3.分词过程算法

- **基于 Trie 树结构实现词典扫描并生成有向无环图 (DAG)**

结巴分词自带了一个叫做 dict.txt 的词典，里面有 2 万多条词，包含了词条出现的次数和词性。Trie 树结构实现的词图扫描，就是把这 2 万多条词语，放到一个 Trie 树中，而 Trie 树是有名的前缀树，也就是说一个词语的前面几个字一样，就表示他们具有相同的前缀，就可以使用 Trie 树来存储，具有查找速度快的优势。

DAG 有向无环图，就是后一句的生成句子中汉字所有可能成词情况所构成的有向无环图，即就是给定一个待分词的句子，对这个句子进行生成有向无环图。切分步骤如下所示：

1) 根据 dict.txt 生成 Trie 树

2) 对待分词句子，根据 dict.txt 生成的 Trie 树，生成 DAG，实际上通俗的说，就是对待分词句子，根据给定的词典进行查词典操作，生成几种可能的句子切分。

例如:{0:[1，2，3]} 这样一个简单的 DAG，就是表示 0 位置开始，在 1，2，3 位置都是词，就是说 0~1，0~2，0~3 这三个起始位置之间的字符，在 dict.txt 中是词语。

- **采用动态规划查找最大概率路径并找出最大切分组合**

字典在生成 Trie 树的同时，也把每个词的出现次数转换为了频率。对于频率和概率，按照定义，频率其实也是一个 0~1 之间的小数，是事件出现的次数/实验中的总次数，因此在试验次数足够大的情况下，频率约等于概率，或者说频率的极限就是概率。

动态规划中，先查找待分词句子中已经切分好的词语，对该词语查找该词语出现的频率(次数/总数)，如果没有该词(既然是基于词典查找，应该是有的)，就把词典中出现频率最小的那个词语的频率作为该词的频率，也就是说 P(某词语)=FREQ.get('某词语', min_freq)，然后根据动态规划查找最大概率路径的方法，对句子从右往左反向计算最大概率，因为汉语句子的重心经常落在后面，就是落在右边，因为通常情况下形容词太多，后面的才是主干。因此，从右往左计算，正确率要高于从左往右计算，这个类似于逆向最大匹配)，P(NodeN)=1.0，P(NodeN-1)=P(NodeN)*Max(P(倒数第一个词))...依次类推，最后得到最大概率路径，得到最大概率的切分组合。

3.1.4.分词器附带功能

- **词性标注**

对分词后的词语应用N-gram语言模型，比较该词在上下文中为某些词性的概率，从而提取最大概率者作为该词词性。该步骤对于后续关键词抽取和停用词过滤的效果也具有至关重要的影响。

- **自定义词典载入**

可以指定自己自定义的词典，以便包含jieba词库里没有的词。虽然jieba有新词识别能力，但是自行添加新词可以保证更高的正确率。用户可行编添加新词到自定义词典中，调整频率，从而满足用户希望获取近期流行新词的诉求。

3.2. 垃圾评论过滤算法

3.2.1. 垃圾评论分类

针对当前不同类型的垃圾评论，通常将垃圾评论分为内容型垃圾评论和欺诈型垃圾评论，依据本项目的实际倾向，本团队只考虑内容型垃圾评论。内容型垃圾评论通常指在评论内容中发布广告、评论内容与主题相关性低、在评论内容中发布 WEB SPAM 以及黄赌毒信息的评论。

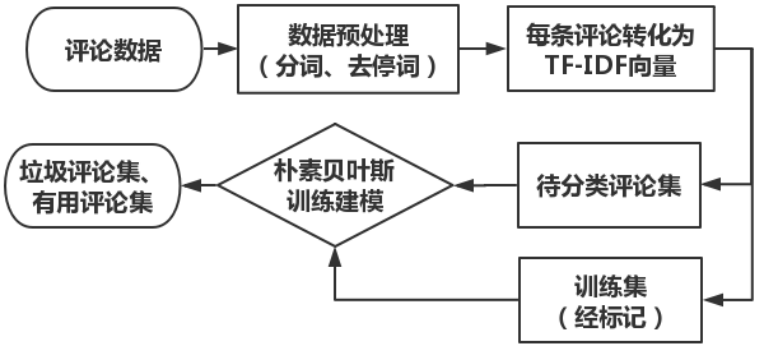
内容型垃圾评论的检测主要考虑评论的内容特征和评论质量特征，将经过分析的特征训练分类模型作为依据，进行内容型垃圾评论的检测。其中评论内容的质量特征包括如下几个方面：1.评论内容与主题的相关性；2.评论内容中是否包含广告；3.评论内容的结构完整性和可读性。综合国内外研究现状，本项目组对内容型垃圾评论的特征进行归纳整合，常用特征集合如下表所示：

3.2.2. 基于朴素贝叶斯的垃圾评论分类算法

通过使用有监督学习的朴素贝叶斯二分类器可区分有用评论与垃圾评论，进而过滤垃圾评论。该方法是在分布独立这个假设成立的情况下实现，而在该文本数据中，分布独立的假设基本成立。该方法过程简单速度快，分类效果好。

3.2.3. 垃圾评论检测框架设计

内容型垃圾评论的检测中，首先通过停用词库、情感词库和后缀词库对数据集进行预处理；然后将每一句评论使用 TF-IDF 算法进行特征的选择和提取，并转化为可计算的 TF-IDF 向量。将训练集使用朴素贝叶斯分类器进行训练建模。对于待检测的评论，将其置于模型中计算出分类概率，若为垃圾评论的概率大于 1/2，标记为垃圾评论。



4. 情感分析技术

4.1.1. 基本方法

要分析一句话是积极的还是消极的，最简单最基础的方法就是找出句子里面的情感词，积极的情感词比如：赞，好，顺手，华丽等，消极情感词比如：差，烂，坏，坑爹等。出现一个积极词就+1，出现一个消极词就-1。另外程度词，感叹号会加强语气，因此扫描情感词前后，如果有出现，则权值*3

5. 软件技术

5.1. Spring-boot 框架

5.1.1. Spring boot

该框架使用了特定的方式来进行配置，从而使开发人员不再需要定义样板化的配置。通过这种方式，Spring Boot 致力于在蓬勃发展的快速应用开发领域(rapid application development)成为领导者。

5.1.2. 模型(model)—视图(view)—控制器(controller)

由于一个应用被分离为三层，因此有时改变其中的一层就能满足应用的改变。一个应用的业务流程或者业务规则的改变只需改动 MVC 的模型层。

控制层的概念也很有效，由于它把不同的模型和不同的视图组合在一起完成不同的请求，因此，控制层可以说是包含了用户请求权限的概念。

最后，它还有利于软件工程化管理。由于不同的层各司其职，每一层不同的应用具有某些相同的特征，有利于通过工程化、工具化产生管理程序代码。

5.2. tomcat 服务器

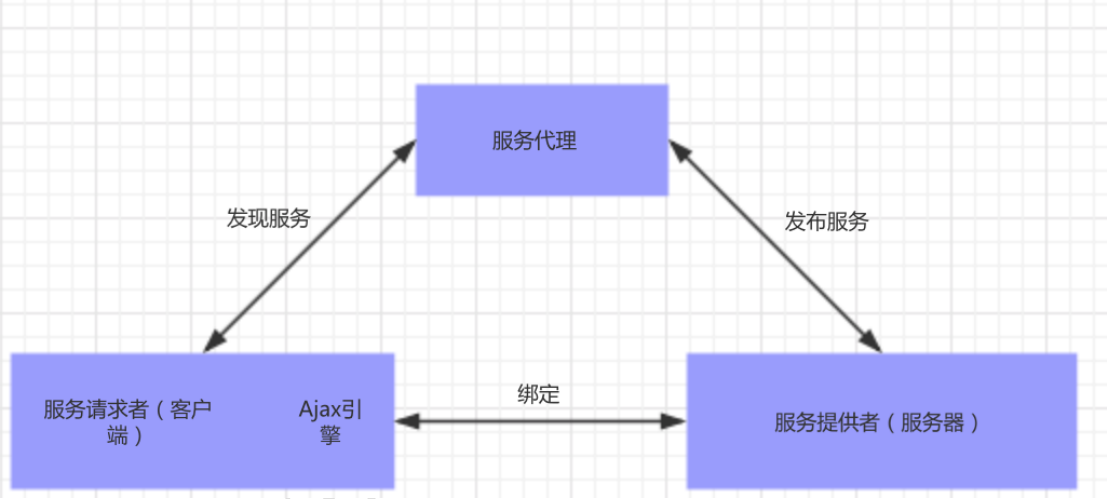
Nginx 服务器是一款轻量级的 Web 服务器和电子邮件代理服务器，并在一个 BSD-like 协议下发行其稳定性好、功能集丰富、系统资源低、占有内存少，并发能力强，且其并发能力在同类型的网页服务器中表现突出。

Spring boot 内置了 tomcat 服务器，tomcat 是一款轻量级的 Web 服务器，因其免费、开源、支持最新标准、更新快、支持跨平台等优点被广泛使用。

5.3. 基于 Ajax 技术的 Web 服务架构

在传统的 Web 服务模式下，用户和服务器之间是一种同步关系，服务器在处理请求的时候，用户多数时间只能等待，限制了交互性，用户体验较差。基于 Ajax 技术的 Web 服务架构为浏览器提供了与服务器端异步通信的能力，可以实现页面的局部刷新而不是加载整个页面，减少了用户等待的时间，更好的满足了用户需求，使得 Web 应用程序更加人性化。

Ajax 即“Asynchronous JavaScript and XML”，是一种创建交互式网页应用的网页开发技术。Ajax 技术实现过程是，Web 页面中的 JavaScript 脚本使用 XM-LHTTP Request 对象与服务器异步通信，服务器接收请求后返回业务数据；数据通过脚本程序处理后，经过数据可视化技术更新显示在 Web 页面中。这种异步数据读取方法使 Ajax 可以自主的发起 Web 请求，与远端服务器完成必要的数据交互，在构建 Web 页面时，无需中断交互流程即可重新加载和动态更新，既减轻了服务器负载又加快了响应速度，缩短了用户等待的时间。



图基于 Ajax 技术的 Web 服务架构

该项目在搜索 APP、搜索 APP 评论、展示用户关注的 APP 页面采用了 Ajax 技术，使用户不需要刷新页面而得到搜索结果。

6. 前端交互技术

6.1. thymeleaf 模板

德鲁伊的翡翠梦境—APP 评论数据分析系统前端设计框架采用的是 thymeleaf 模板 Thymeleaf 是一款用于渲染 XML/XHTML/HTML5 内容的模板引擎。类似 JSP，Velocity，FreeMaker 等，它也可以轻易的与 Spring MVC 等 Web 框架进行集成作为 Web 应用的模板引擎。与其它模板引擎相比，Thymeleaf 最大的特点是能够直接在浏览器中打开并正确显示模板页面，而不需要启动整个 Web 应用。Thymeleaf 是与众不同的，因为它使用了自然的模板技术。这意味着 Thymeleaf 的模板语法并不会破坏文档的结构，模板

依旧是有效的 XML 文档。模板还可以用作工作原型，Thymeleaf 会在运行期替换掉静态值。Velocity 与 FreeMarker 则是连续的文本处理器。

7. 数据可视化技术

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。

7.1. 基于几何的技术

基于几何的可视化技术包括 Scatter plots、Landscapes、Projection Pursuit、Parallel Coordinates 等等，是以几何画法或几何投影的方式来表示数据库中的数据。平行坐标法是最早提出的以二维形式表示 n 维数据的可视化技术。它的基本思想是将 n 维数据属性空间通过 n 条等距离的平行轴映射到二维平面上，每一条轴线代表一个属性维，轴线上的取值范围从对应属性的最小值到最大值的均匀分布。这样，每一个数据项都可以根据其属性值用一条折线段在 n 条平行轴上表示出来。

利用这个技术加上 d3.js 我们可以设计出符合要求的 x，y 轴，图表可以跨坐标系存在，例如折、柱、散点等图可以放在直角坐标系上，也可以放在极坐标系上，甚至可以放在地理坐标系中。以及合适的数据范围和比例尺。

7.2. 面向像素技术

面向像素技术的基本思想是将每一个数据点的数据值对应于一个带颜色的屏幕像素，对于不同的数据属性以不同的窗口分别表示（图 2）。面向像素技术的特点在于能在屏幕中尽可能多的显示出相关的数据项，对于高分辨率的显示起来说，可实现多达10^6数量级的数据。

本项目收集的评论数据以及下载量等数据少则几万条，多则几十万，利用此技术可以将数量级很大的数据清晰的显示在图表中，尽可能的保持数据的完整性。

7.3. 基于图标的技术

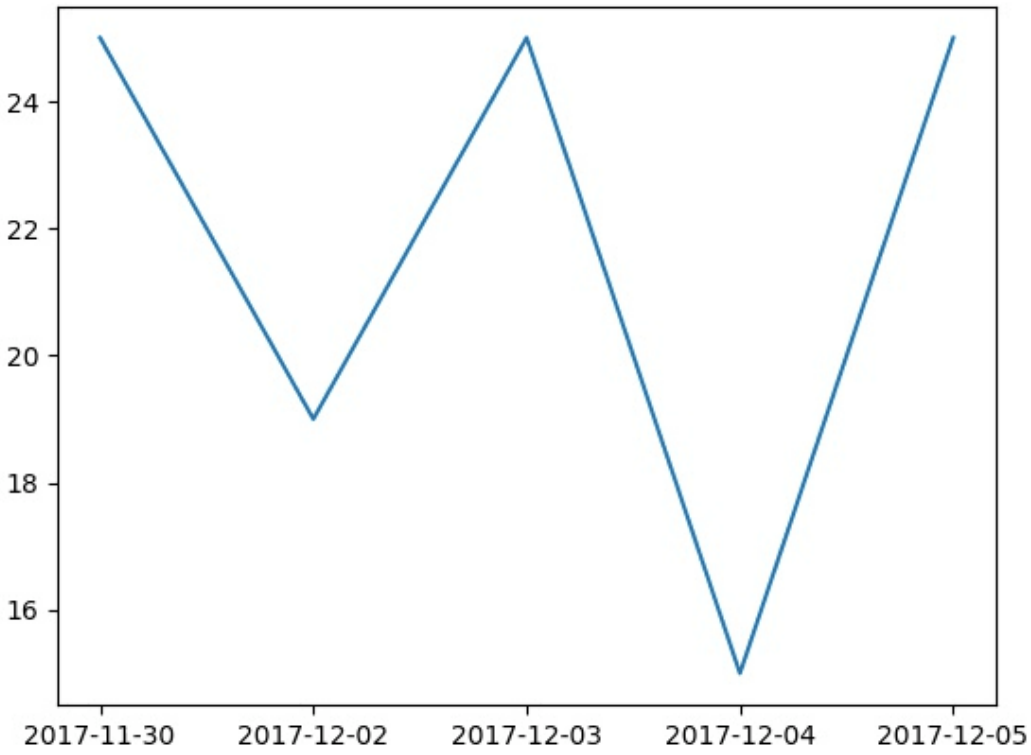
基于图标技术的基本思想是用一个简单图标的各个部分来表示 n 维数据属性。基于图标的可视化技术包括 Chernoff-face、Shape Coding、Stick Figures 等，这种技术适用于某些维值在二维平面上具有良好展开属性的数据集

枝形图方法是其中的基本方法之一。枝形图方法首先选取多维属性中的两种属性作为基本的 X-Y 平面轴，在此平面上利用小树枝的长度或角度的不同表示出其他属性值的变化。

本项目利用多维属性在一个图表中表示出一个或多个 app 的多个属性，可以使用户在短时间内有效快速的获得 app 的不同方面的信息。例如下图所示的两个数据点，它们对左边的二维属性含有相同的数据值，而右边的二维属性的数据值则不相同。

7.4. 独立坐标系

支持直角坐标系（catesian，同 grid）、极坐标系（polar）、地理坐标系（geo）。图表可以跨坐标系存在，例如折、柱、散点等图可以放在直角坐标系上，也可以放在极坐标系上，甚至可以放在地理坐标系中。



7.5. 多维数据支持和丰富视觉编码

除了加入了平行坐标等常见的多维数据可视化工具外，对于传统的散点图等，传入的数据也可以是多个维度的。配合视觉映射组件 Visual Map 提供的丰富的视觉编码，能够将不同维度的数据映射到颜色、大小、透明度、明暗度等不同的视觉通道。

我们可以在图表中加入视觉组件增加用户体验，当鼠标移动到相应位置时，会提示具体的数据，以及数据的权重大小会根据颜色的深浅更加直观的视觉输出。图表的控件可以将图表中的数据通过表格一键呈现，还可以将图表一键导出为图片，便于后续的整理以及邮件等的发送。


为什么
com
hang81
04闪退
新浪微博 熠言堂
完全不能开
熠言堂 新浪微博
可玩性高可以的啊
差评
04垃圾
n2
ios11
06iPhone
01iPhone
怎么办
怎么解决
谢谢
如题

图多维数据及视觉特效

8. 数据库技术

8.1. 关系型数据库 MySQL

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件，也是最适配于 PHP 框架 Laravel 的数据库。

MySQL 具有如下特性：

1) 使用 C 和 C++编写，并使用了多种编译器进行测试，保证源代码的可移植性。
2) 支持 AIX、BSDi、FreeBSD、HP-UX、Linux、Mac OS、Novell NetWare、NetBSD、OpenBSD、OS/2 Wrap、Solaris、Windows 等多种操作系统。
3) 为多种编程语言提供了 API。这些编程语言包括 C、C++、C#、VB.NET、Delphi、Eiffel、Java、Perl、PHP、Python、Ruby 和 Tcl 等。
4) 支持多线程，充分利用 CPU 资源，支持多用户。
5) 既能够作为一个单独的应用程序在客户端服务器网络环境中运行，也能够作为一个程序库而嵌入到其他的软件中。
6) 提供多语言支持，常见的编码如中文的 GB 2312、BIG5，日文的 Shift JIS 等都可以用作数据表名和数据列名。
7) 提供 TCP/IP、ODBC 和 JDBC 等多种数据库连接途径。
8) 提供用于管理、检查、优化数据库操作的管理工具。
9) 可以处理拥有上千万条记录的大型数据库。

8.2. 数据库表关系