基于用户评论挖掘的商品分析系统的设计与实现

王淑军 刘 成 崔富超 李鹏飞 梁鑫月

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘 要：随着我国互联网普及率和人民生活水平的提高，网上购物已成为很多人愿意选择的购物方式。据调查显示，近几年来网购所占购物数额的比例逐年大幅度提升[1]。而与此同时，电子商务平台上商品的评论数量也呈几何式上升，使得想要购买商品的顾客很难在纷繁的评论中提取到自己所需要的信息。即使能得到信息也过于片面，不能系统合理评测商品。笔者站在用户的角度上开发本系统，目的是希望能降低顾客网络购物风险。本系统主要采用爬虫技术、Java 技术、数据挖掘中的一些算法和一些当前比较先进的开源接口和框架来实现，用户可参考系统给出的评测结果从而选择是否购买该产品[2]。

关键词：电子商务；评论挖掘；Python；推荐结果

中图分类号：TP391.1 **文献标识码：**A **文章编号：**1003-9767（2017）10-131-02

Design and Implementation of Commodity Analysis System based on User Review Mining

Wang Shujun, Liu Cheng, Cui Fuchao, Li Pengfei, Liang Xinyue

(School of Computer Science, Inner Mongolia University, Hohhot Inner Mongolia 010021, China)

Abstract: With the penetration of China´s Internet and the improvement of people´s living standards, shopping on the Internet has become a choice for many people. According to the survey, in recent years the proportion of online shopping accounted for a substantial increase in the proportion of shopping. At the same time, the number of reviews on the e-commerce platform also showed a geometric rise, making it difficult for customers who want to buy goods to extract the information they need in numerous comments. Even if you can get information, it´s too one-sided to evaluate the product properly. The author develops this system on the user´s point of view, the aim is to reduce the risk of shopping. The system mainly uses crawler technology, Java technology, some algorithms in data mining and some advanced open source interface and framework to achieve, the user can refer to the evaluation results given by the system to choose whether to purchase the product.

Key words: e-commerce; comment mining; Python; recommendation results

1 数据获取

主要是使用爬虫技术对商品信息页面上的商品评论进行爬取，并保存到数据库，但这中间存在一些技术问题。

以天猫为例，它使用了 Ajax 加密方式，它会从你看不见的其他页面读取数据，也就是说你想要通过直接右键查看源代码的方式查看评论的网页结构来做正则表达式匹配是不能够做到的[3]。但是这并意味着就没有办法获取到网页评论，可以通过一些流量监控工具如浏览器自带的开发者工具，启动网络流量捕获并留意类型为“text/html”或者“application/json”的网址，经过分析测试之后就可以分析获得你需要的正则表达式。

获取正则表达式之后运行爬虫抓取数据储存到数据库中，数据表包含如下几个属性列：用户名、用户评论时间、用户评论的内容以及标志位。

2 评论清洗

（1）根据获取的用户名和用户评论时间来进行初步评论清洗，因为目前商家普遍存在刷好评的现象，而这样刷的好评是没有任何参考价值的，所以要进行初步过滤。

具体实现流程如下。从第一条记录到第 N-1 条记录循环向下对比：如果两条记录的用户名相似度大于等于 0.8，并且两个用户之间评论时间差小于等于 24 小时，则将两条记录标志为 1。循环比较结束之后删除所有标志位为 1 的评论。

作者简介：王淑军（1996-），男，山东临沂人，本科。研究方向：用户评论分析、数据分析。

（2）删除系统默认好评记录。因为其不具备分析价值。

（3）使用 K-means 算法 [4] 将所有评论中按照以下几个属性，即评论长度、是否有图、是否匿名评价等进行聚类算法聚类商品评论 [5]。挑选出其中比较值得信任的聚类进行分析。下面对 K-means 进行解释。

K-means 的基本思想是初始随机给定 K 个簇中心，按照最邻近原则把待分类样本点分到各个簇。然后按平均法重新计算各个簇的质心（这个点可以不是样本点），从而确定新的簇心。一直迭代，直到簇心的移动距离小于某个给定的值。

在这里 K 值取 3，将所有评论分成三个种类。

具体实现步骤如下：

（1）为待聚类的点寻找聚类中心；

（2）计算每个点到聚类中心的距离，将每个点聚类到离该点最近的聚类；

（3）计算每个聚类中所有点的坐标平均值，并将这个平均值作为新的中心。

反复执行步骤（2）、（3），直到聚类中心不再进行大范围移动或者聚类次数达到要求为止。

3　评论内容处理

（1）利用 word2vec 对同义词进行聚类。如物流、快递、发货速度等统一归结到物流服务中，颜色、大小统一归结到款式中。

使用 word2vec 技术对同义词进行聚类，获得商品特征词和商品观点词的常用同义词，如特征词物流的同义词表、服务的同义特征词表以及观点词非常好、一般、不好的同义词表。

具体实现流程如下：

①网上下载中文维基百科数据作为训练数据在以后使用；

②将下载的 xml 文件转换成 text 文件；

③使用开源项目 open cc 将数据中的繁体字转换成为简体字；

④使用 jieba 分词对文档进行分词处理，方便以后使用；

⑤使用 iconv 去除中间的非 utf-8 字符。

Word2vec 使用已处理好的文档进行数据训练时间比较久，需要等候一段时间。

训练结束以后，输入之前需要的特征词和观点词，获取它们的同义词并保存到数据库中。

（2）利用 Stanford Parser 进行中文观点抽取。

所谓的观点抽取就是从文本中获取关于某个特征词的观点词语。特征词从词性上看一般为名词，而观点词通常为带有情感色彩的形容词或者副词。观点词的抽取在用户评价分析产品中非常有用。

例如，在句子“卖家的服务态度不错，快递也很迅速”这个句子中，“服务”和“快递”是两个描述卖家的特征词，而“不错”和“迅速”则是这两个词的观点词。

Stanford Parser 是由斯坦福大学自然语言处理小组开发的开源句法分析器，是基于概率统计句法分析的一个 Java 实现。

之所以选择 Stanford Parser 是因为它具有一系列优点，如既是一个高度优化的概率上下文无关文法和词汇化依存分析器，也是一个词汇化上下文无关文法分析器；提供了多样化的分析输出形式，除句法分析树输出外，还支持分词和词性标注文本输出、短语结构树输出等；通过设置不同的运行参数，可实现句法分析模型选择、自定义词性标记集、文本编码设置和转换、语法关系导入和导出等功能的定制。

具体的实现流程如下：

①首先将数据库中选出的用户评论进行语法分析；

②按照距离优先原则将之前第二步中获取的特征词同义词词组与评论中的情感词进行匹配，形成键值对，如物流 - 好、服务 - 不好等。

4　结果处理

经过前面的处理之后已能够获取一些关于商品评论的键值对了，接下来只需要再执行一次对键值对的聚类并且记录好坏数量推荐给用户即可。

5　结　语

产品评论挖掘是一个充满机遇和挑战的研究领域，尽管取得了一些研究成果，但是许多问题还有待进一步探索和研究。由于时间和条件的限制，本文的研究仍然存在不足，主要包括开发方式不集中、不能集中操作，以及不能自动抓取评论进行顺序分析等。在以后需要进一步优化和完善，为买家提供更好的服务。

参考文献

[1] 艾瑞网 .2014 年电子商务核心数据发布 [EB/OL].(2015-02-09)[2017-05-10].http://news.iresearch.cn/zt/246308.shtml.

[2]Hu N, Liu L, Zhang J. Analyst Forecast Revision and Market Sales Discovery of Online Word of Mouth[C]// Hawaii International Conference on System Sciences. 2007.

[3] 曾伟辉 . 支持 AJAX 的网络爬虫系统设计与实现 [D]. 安徽 : 中国科学技术大学 ,2009.

[4]SS Khan, A Ahmad. Cluster center initialization algorithm for K -means clustering[J]. Expert Systems with Applications, 2014, 40(18): 7444-7456.

[5] 孙吉贵 , 刘杰 , 赵连宇 . 聚类算法研究 [J]. 软件学报 ,2008,19(1):48-61.