数据挖掘在移动客户投诉分析中的应用研究

王　琴[1],张　炯[2]

(1.无锡机电高等职业技术学校,江苏无锡 214028;
2.中国联通无锡分公司,江苏无锡 214000)

【摘要】针对通信企业分析汇总海量客户投诉数据提取有效信息时工作量较大、定位不够精准的问题,文章提出了一种基于 Python 的客户投诉大数据挖掘的方法, 通过 Python 编程将离散的文本信息运用 TF-IDF 词频统计的方法精准定位出投诉热点,为企业降本增效、NPS 提升提供重要数据支撑。

【关键词】数据挖掘;Python;投诉热点;TF-IDF;NPS

【doi:10.3969/j.issn.2095-7661.2018.04.008】

【中图分类号】TP311.13;TN929.5　　【文献标识码】A　　【文章编号】2095-7661(2018)04-0025-03

Research on application of data mining in mobile customer complaint analysis

WANG Qin[1],ZHANG Jiong[2]

(1. Wuxi Machinery and Electron Higher Professional and Technical School, Wuxi, Jiangsu, China 214028; 2. China Unicom Wuxi Branch, Wuxi, Jiangsu, China 214000)

Abstract:The customer complaint database of telecommunication companies is massive. The analysis and extraction of valid data from the database is time or resource consuming and inaccurate. This paper proposes data mining of customer complaint database using Python. Python program processes text-based discrete information using TF-IDF word frequency statistics to accurately identify top customer complaints. It helps to reduce cost, improve efficiency and increase NPS by providing critical data support.

Keywords:data mining; Python; top customer complaints; TF-IDF; NPS

随着 4G 业务的飞速发展,通信业务从原有的语音、短信、低速数据上网等比较单一的业务,发展成为高速率、多媒体的综合性通信。移动支付、共享单车、支付宝等改变了人们的生活方式,广大用户对通信业务的需求越加迫切。在满足人们日益增长的通信需求的同时,有效降低设备投资,实现绿色节能减排,成为电信行业亟待解决的问题。因而通过大数据挖掘海量客户投诉信息,分析通信网络中存在的缺陷,获取网络建设基础数据已经成为运营商的必然选择。本文选取了某通信运营商一万多条网络类投诉单进行大数据分析。

1 设计思路

本文旨在研究出一种基于中文文本挖掘的大数据分析方法,通过对海量数据自动分词、词频统计,实现投诉热点的自动分拣、归类以及自动统计,以期降低大量的手工重复劳动工作量。

1.1 中文分词技术概述

英语分词相对比较简单,每个单词会通过空格来进行自动区分,而中文继承古汉语的书写习惯,没有明确的分隔符, 单个汉字一般无法表达完整的含义,需要 2、3 个或者更多的汉字组成一个词组才能表达准确的含义,中文分词与英文分词相比难度更大。中文分词技术属于自然语言处理技术类别,目前比较主流的分词算法有三类: 基于字符串匹配的分词方法、

[收稿日期] 2018-07-03

[作者简介] 王琴(1975-),女,江苏无锡人,无锡机电高等职业技术学校副教授,硕士,研究方向:电子与通信。

[基金项目] 2016 年江苏省职业教育教学改革研究项目立项课题“《通信电子电路》课程群‘做学教合一’教学模式的研究与实践”(课题编号:ZYB148)。

基于理解的分词方法和基于统计的分词方法。

1.2 中文分词技术比较

1)基于字符串匹配的分词方法。将分词结果与事先设定的字符串的字典进行匹配,根据不同的匹配原则可以分为最大匹配、最小匹配,或正向匹配、逆向匹配等。这种分词方法依赖于已知的词组库,对于专业文献、具有地方特色的文本的匹配度比较差,但它的优势是速度较快,无效文本比较少。

2)基于理解的分词方法。是使计算机模拟人脑的思考方式根据词义、句意的理解自动分词。目前比较主流的分词工具有 Jieba、SnowNLP、THULAC、NLPIR 等,基于理解的分词方法是未来中文分词的方向,随着人工智能的深入发展会逐步成熟,但是由于中文语言结构的复杂性,目前此项技术尚处于实验阶段,还不能实际投入使用。

3)基于统计的分词方法。根据事先设定的分词规则,利用计算机的高速运算性能,机械统计分词词频数。这一分词方法可以确保中文分词内容的完整、准确,缺点是统计的中文分词数量较多,无效数据需要人工干预处理。

1.3 中文分词技术的选择

三种中文分词技术各有利弊,需要根据不同的应用场景选择使用。由于客户投诉文本中有较多地名、人名等不属于常见的词组,选用基于理解或者基于字符串的分词技术容易导致漏词,所以本文主要采用基于统计的分词技术,同时结合部分基于字符串的技术进行二次筛查。

2 编程语言的选择

传统的编程语言主要是 Java、C 语言、C++、Visual Basic.NET 等,随着技术的演进,一些新型的编程语言开始崭露头角,其中比较有代表性的语言包括 Python、R 语言、MATLAB 等,本文使用的编程语言是 Python。Python 是一种新型的脚本语言,它是一种解释型语言,类似于 PHP 和 Perl 语言,无需编译环节。[1] Python 是面向对象语言,即支持面向对象的风格或将代码封装在对象的编程技术。Python 具有简洁、易于维护、便于移植、拥有丰富的资源库、具有较强的扩展性等优点。

3 文本挖掘的主要步骤

3.1 中文编码格式规整

英文、法语或者其他西方国家文字,只有几十个或上百个的字符构成,1 个字节的空间即可代表所有的字母组合,但汉字属于图形化的语言,常见的汉字较多,再加上一些生僻字等,需要一定的编码规则来完成汉字与 ASCII 码的对应关系。出于容量的考虑,一个汉字需要占用多个字节,才能满足所有汉字的翻译需求。目前常见的中文编码主要有:GB2312、BIG5、GBK 和 UTF-8。UTF-8 保留了 ASCII 字符的编码作为它的一部分,可以同屏显示多种语言,考虑到文本阅读的兼容性,本文选用 UTF-8 对文本字符进行格式规整。

3.2 除去文本中的非中文部分

原始文本中存在多种字符,有中文字符、英文、数字、html 的标签符号等,这些都属于垃圾数据,需要通过数据清洗将他们去掉。Python 中的正则表达式(re)可以完成此项任务,从 1.5 版本开始,Python 增加了 re 模块,它提供类似 Perl 风格的正则表达式模式,可以实现查找、比较、替换、匹配等功能。用 UTF-8 格式规整后的文本为双字节字符,可以通过[^\x00-\xff]语句来进行字符匹配,从而根据特征值清洗掉垃圾数据,并完成中文字符与英文字符的分离。

3.3 去除停用词

为了有效提高检索效率压缩存储空间,完成数据清洗工作,去除对检索结果无用的停用词也是文本检索的重要环节。网上目前存在各种停用词库,本文结合客户投诉的实际情况,对停用词库进行了归集,最终采用了一千多个停用词,对原始的投诉文本进行了数据清洗。

3.4 中文分词

本文主要采用的是基于统计的中文分词,运用 r.find()函数并结合[^\x80-\xff]语句匹配中文字符。然后将连续的 2-12 个 ASCII 码翻译成 2-6 个字组成词组,完成文本特征的提取工作,最终将词组统一存入字典,完成中文分词过程。

3.5 词频统计

完成字典录入后,需要进行词频统计,词频(term frequency, TF)指的是某个特定的词组在整个文件中出现的次数,它体现了特定词组在整个文件中的出现概率高低,通常出现概率大的文本被认为更为重要。[2]目前用于信息检索及文本挖掘的主流加权技术是 TF-IDF (term frequency inverse document frequency),其中 TF 刻画的是词语 t 对某篇文档的重要性,用词语 t 在文件中的出现次数除以文件中所有字词的出现次数之和,其表达式为:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

IDF(inverse document frequency)逆向文件频率,是一个词组普遍重要程度的考量,刻画的是词语对整个文档集的重要性。某一特定词语的 IDF,可以由所有文件数目除以包含该词语之文件的数目,再将得到的商取对数得到,其表达式为:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_i\}|}$$

将 TF 与 IDF 相乘,就得到了 TF-IDF 值:

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

文件内的高词组频率和该词组在整个文档集合中的低文件频率产生出了高权重的 TF-IDF, 从而过滤掉常见的词语, 保留下重要的词语。Python 关于 TF-IDF 的程序代码如下:

```
def tf(wd, jishu):
return jishu[wd] / sum(jishu.values())
def tongji(wd, jishu_lst):
return sum (1 for jishu in jishu_lst if wd in jishu)
def idf(wd, jishu_lst):
return math.log (len (jishu_lst)) / (1 +tongji(wd, jishu_lst))
def tfidf(wd, jishu, jishu_lst):
return tf(wd, jishu) * idf(wd, jishu_lst)
```

4 数据呈现

根据一万多条的投诉文本,在删除了客服标准回复语言、常见的停用词以及标点符号等无效字符后,共获得词频大于 6 次的各类词组近 4 万个。最后把实际有意义的高频词组一百多个进行分类,分为投诉业务类型、受理机构名称、投诉客户分布区域等,提供了多个维度的分析数据。统计结果精准揭示出了投诉用户的分布区域和痛点业务的种类, 对下一步网络规划、提升 NPS(净推荐值),提升客户感知提供了有价值的参考数据[3]。

4.1 投诉业务类型分布

目前通信业务日益复杂,投诉热点由原来的语音为主,逐步转化为上网业务为主,同时由于运营商提供了差异化的流量优惠减免策略,增加了计费的复杂程度。从图 1 统计的各类投诉业务数量显示,计费争议、上网速度、国际漫游是位列前三的投诉热点。寻找网络痛点业务,发现网络中的短板,可以为 NPS 的提升找到突破口,重点业务重点突破,对上网速率开展专题优化,对计费套餐进行针对性的梳理可以降低运营商的投诉量,还可以有效提升用户感知[4]。

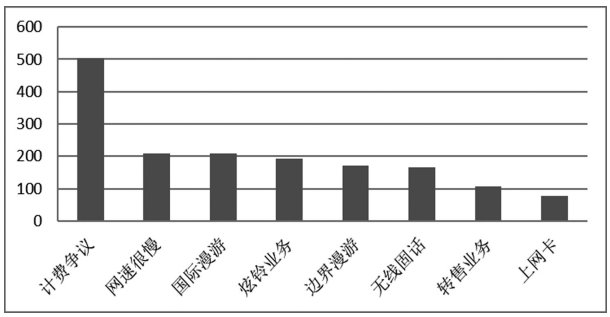

图 1　各类投诉业务数量分布图

4.2 投诉客户区域分布

投诉种类的分布情况使用传统的投诉汇总分析方法就能有效了解,但对于运营商最重要的投诉区域分布则由于技术条件限制,无法有效定位,只能统计到区县级大颗粒度,具体细分到更加微观的层级则无法实现。反映到网络建设方面就体现出建站针对性不够强,有投诉的区域网络优化周期较长,而新建设的基站话务量却不够高。采用数据挖掘技术细分到任意层级的区块能够准确迅速地确定站点位置,在一定程度上减少运营商的成本投入,从而实现资源的有效整合。[5]下面就从区县、乡镇、道路、密集人群四个维度展示细分层级的统计结果。

1)区县投诉量分布:从图 2 显示的统计结果来看,在当地 7 个行政区中,区域的面积较大的 JY 区和 YX 区投诉量最大,作为业务发展、网络覆盖的重点区域, 后续应该加大这两个区域的网络建设投资,但是具体在哪里建站,就需要根据后续更小的层级统计结果来确定。

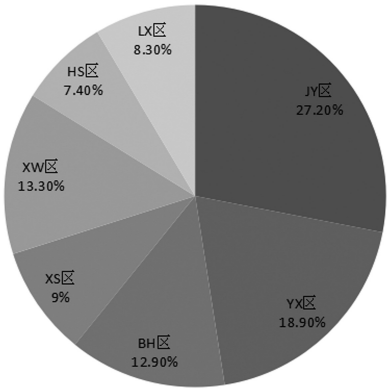

图 2　县区投诉量分布图

2)乡镇投诉量分布:当地共有五十几个乡镇,投诉量排名前三的分别是 MC 镇、SF 镇和 DT 镇, 究其原因主要是这三个乡镇的发展用户较多。因而该地的网络亟待提升,把网络资源投向有用户的区域显然是下一步网络扩容的重点。

3)道路投诉量分布:运营商每年都会有第三方的无线网络评测,测试方法一般是路测,即沿着城市的主干道路进行语言、上网业务测试,来衡量网络的实际质量。道路投诉的分布情况无疑是运营商需要掌控的, 传统意义上的拉网式路测不仅浪费人力物力,而且存在覆盖面较小的问题。从投诉统计分析中寻找数据是一种比较新的思路。从统计结果来看,在当地两百多条主干道路中, 投诉量最高的三条道路分别为: Q 路、B 路和 R 路。这些道路基本都属于城市中心区域, 热点区域的投诉解决可以明显提升网络考核指标、提升用户感知。

4)密集人群区域分布:密集人群区域主要是指公众场所、住宅小区。在全市四十多个密集人群区域中,

(下转第 42 页)

为:alter table table_name move tablespace tablespace_name。方法二,通过 expdp/impdp 操作来进行碎片整理。方法三,整理行迁移(dba_tables.chain_cnt)数目较大的表。

识别行迁移:SQL>analyze table TABLE compute statistics (会覆盖统计信息);SQL>@?/rdbms/admin/utlchain.sql;SQL>analyze table table_name list chained rows into chained_rows。

消除行迁移可通过 move、expdp/impdp、将迁移行插入临时表再删除后回填的方法(不适用于有外键引用的表)。

2.3 增强架构变更管理

由于企业营销手段的多样化、业务变更频繁,信息支撑系统会经常进行版本发布,而版本发布就会有数据库结构的变更。因此在版本发布前一定要加强架构变更管理,保证数据结构的变更不会破坏系统整体架构。同时加强对新增或变更 SQL 语句的审核,要在测试平台进行测试,同时加强这些 SQL 语句执行总效率的评估和上线后的性能跟踪。

2.4 实施日常性能监控

通过 AWR、ADDM 对数据库进行性能监控,及时发现消耗资源的 SQL 语句,通过业务及技术手段来优化这些 SQL 语句,以保证系统持续健康的运行。

1)定期检查消耗资源过高的 SQL。通过操作系统监控及时发现 CPU 的异常增长情况,对比近期 CPU 消耗情况,消耗资源过高的 SQL 可以通过 ORACLE 的 AWR 报告分析出来,及时优化[5]。

2)定期分析特殊的业务表。虽然 oracle 10g 以上版本是数据库后台自动进行表分析的,但是只有在数据表的数据量改变在 10%以上才会自动分析,对于大表,自动分析有时需要较长时间才会执行。带有时间索引的 SQL,如果不及时分析,将产生错误的执行计划,影响系统稳定。因此,根据业务需求,增加分析特殊业务表的定时任务是必要的。

3 总结

通过数据库架构优化和数据治理,基本可以达到 ORACLE 数据库系统所需容量有效控制、系统物理模型有效精简、系统碎片有效整理的效果,使得整个数据库系统性能提升、系统响应速度提高。同时系统运行效率较高,维护费用就会相对降低,一定程度上可节省系统维护成本。

【参考文献】

[1]吴锋珍.高校图书管理系统的数据库设计[J].湖南邮电职业技术学院学报,2016(2):36-39.

[2]韦壹刚.Oracle 数据库日常维护与优化建议[J].信息与电脑(理论版),2018(8):115,117.

[3]常正青.基于 Oracle 数据库系统的优化与性能调整研究[J].电子世界,2017(17):195.

[4]韦德昌.Oracle 数据库应用系统的性能优化[J].数码世界,2017(4):14-15.

[5]孟津平.浅谈 Oracle 数据库 SQL 调整与优化[J].数字技术与应用,2018(3):227,229.

(上接第 27 页)

投诉量最高的为:J 社区、火车站、R 医院。公众场所的投诉量居高不下,处理一个类似的热点区域往往能够同时处理几十个甚至上百个同类用户的投诉,此类投诉对 NPS 影响较大。因此加强公众场所、住宅小区的建设应该是今后建站重点考虑的方面。

5 结束语

进行基于大数据挖掘的客户投诉热点分析,通过挖掘投诉大数据价值,可以对运营商在网络、服务等多个环节进行预警,从客观角度给运营商提供了一个新的视角,审视运营商的服务是否达到客户的需求。运营商可以结合网管告警、话务统计、客户投诉等多维数据来源,实现网络的维护闭环处理,一方面完成投诉、NPS 等多种考核指标的提升,另外一方面可以提高网络投资效益,做到精准投资、有效投资,具备较大的推广价值。

【参考文献】

[1]聂晶.Python 在大数据挖掘和分析中的应用优势[J].广西民族大学学报(自然科学版),2018(1):77-79.

[2]卢中宁,张保威.一种基于改进 TF-IDF 函数的文本分类方法[J].河南师范大学学报,2012(6):158-160.

[3]曾雄辉.MR 大数据分析支撑 LTE 800M 精准规划建设研究[J].湖南邮电职业技术学院学报,2017(4):14-17.

[4]蔡志成.基于 4G 环境的数据挖掘在移动通信网络优化中的运用研究[J].通讯世界,2017(8):81-82.

[5]王森.4G 环境下移动通信网络优化中数据挖掘的应用研究[J].网络安全技术与应用,2016(12):96-98.