基于文本挖掘的投诉热点智能分类

夏海峰, 陈军华*

(上海师范大学 信息与机电工程学院,上海 200234)

摘 要: 投诉识别系统在保证热点投诉正确分类、提高电信行业的服务质量中起到很重要的作用. 由于电信行业的客户投诉有其特殊性,所有的投诉必须在很短的时间内分类完成,从而往往会发生导航分类错误的现象. 提出了一套基于文本挖掘的模型,该模型能够智能地将热点投诉分类到正确的投诉导航上去. 实验表明: 该模型能够有效地进行投诉文本分类.

关键词: 文本挖掘; 智能分类; 投诉

中图分类号: TP 391.4 **文献标识码:** A **文章编号:** 1000-5137(2013)05-0470-06

手机通话、短信、网络 GPRS 等服务作为电信行业的基本服务,时刻与用户紧密联系着,提高服务质量任重而道远. 首先从管理流程上来讲,目前客户投诉分析面临很多挑战: 投诉内容难分析. 信息量大、非结构化,文本内容难分析,需要人工逐条查阅,工作繁琐且效率低下; 投诉点多难聚焦. 仓库管理系统(WMS)中对投诉分类固定粗放投诉散点多,投诉管理人员无法对投诉进行统一归类集中分析,只能逐条分析. 优化抓手难获取. 对投诉原因分析少,无法及时了解客户对服务和产品的不满意原因,造成对产品服务优化工作抓手获取难等. 这些问题都影响了客户投诉分析的质量.

为了解决这些问题,引入文本挖掘的理念和方法,探索了一套基于投诉文本的数据挖掘模型,提出了投诉热点智能分类的概念,在原有的导航分类的基础上,利用投诉文本数据,根据文本挖掘[1]的相关概念,采用 SVM 算法[2]、统计学知识,最终创建投诉导航树. 因为投诉分类种类过多,分词部分以"费用"相关投诉文本为例,进行相应的研究工作.

1 相关概念

文本数据挖掘(Text Mining)[3]是指从文本数据中抽取有价值的信息和知识的计算机处理技术. 顾名思义,文本数据挖掘是从文本中进行数据挖掘(Data Mining). 从这个意义上讲,文本数据挖掘是数据挖掘的一个分支. 文本数据挖掘是一个边缘学科,由机器学习、数理统计、自然语言处理等多种学科交叉形成. 文本挖掘的关键技术主要包括以下几点:

(1) 信息抽取. 信息抽取是从自然语言文本中抽取预先指定的实体、关系、事件等信息,形成结构化的数据并填入数据库的过程. 信息抽取常用于改善信息检索,帮助用户直接定位所需的信息,无需阅读文档的全部内容.

(2) 文本分类. 文本分类是利用计算机对文本集(或其他实体或物件)按照预先定义的分类体系或标准进行自动分类标记. 文本分类是采用基于主题对文档按主题进行自动归类. 投诉热点模型是基于主

收稿日期: 2013-09-04

基金项目: 上海市教委项目基金(12ZT05)

作者简介: 夏海峰(1989-),男,上海师范大学信息与机电工程学院硕士研究生; 陈军华(1967-),男,上海师范大学信息与机电工程学院副教授.

* 通信作者

题的应用.

(3) 文本聚类. 文本聚类是基于“同类的文档相似度较大,而不同类的文档相似度较小”理论,假设对文档集合进行有效地组织、摘要和导航,方便人们从文档集中发现相关的信息.

(4) 关联规则. 关联规则是描述一个事物中某些属性同时出现的规律和模式. 它的核心是将各种信息载体中的共现信息定量化的分析方法,以揭示信息的内容关联和特征项所隐含的寓意,藉此可以发现研究对象之间的亲疏关系,挖掘隐含的或潜在的有用的信息.

2　基于文本挖掘的文本分类过程和关键技术

2.1　投诉热点智能分类整体流程

选取最近 1000 条投诉分类文本作为模型的基础,按顺序进行 3 个阶段(图 1)的操作: 预处理阶段、文本表达阶段、知识挖掘阶段,经过这 3 个阶段的处理之后,形成最终的导航参考分类模型.

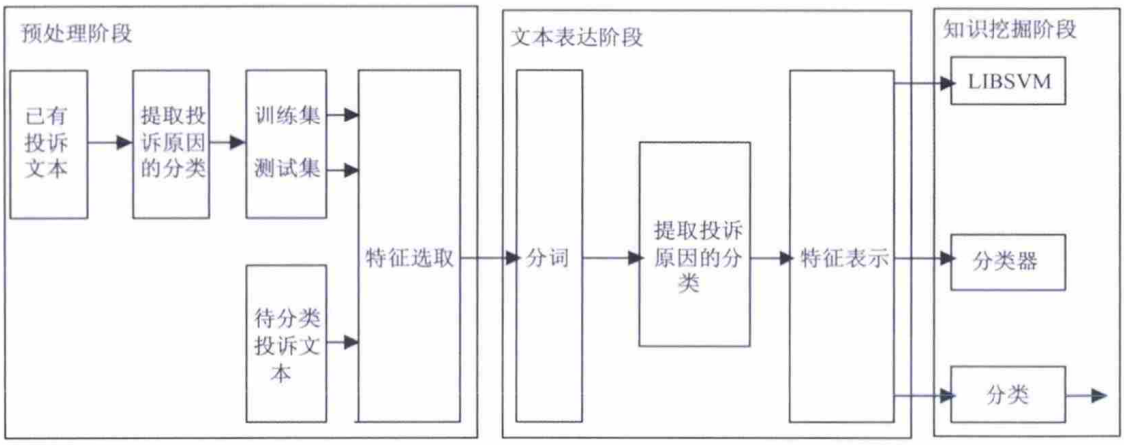

图 1　投诉监控热点智能分类操作流程图

2.2　文本预处理

文本预处理的文档来源于客户投诉文本内容. 投诉文本内容是典型的非结构化或半结构化的数据,缺乏关系数据库中数据的结构化和组织性. 因此,要对原始文档进行预处理,使其转化为较为规整且能反应文档内容的特征表示. 做简单的文档说明: (a) 整理已有的投诉文本内容; (b) 提取投诉原因的分类; (c) 信息训练集和测试集; (d) 进行特征选取,信息待分类投诉文本.

2.3　文本表达

文本表达的过程主要是对预处理出来的文档进行词法的分割、划分,最终提取出关键词字段,具体的过程主要包含以下 4 个方面:

2.3.1　中文分词

中文分词指的是将一个汉字序列切分成一个一个单独的词. 中文分词是文本挖掘的基础,其处理过程就是分词算法,对于输入的一段中文,进过分词之后,能够达到被电脑自动识别语义的效果. 在中文分词阶段,作者采用的是中国科学技术研究所研制的汉语词法分析系统(ICTCLAS)[4],具体的操作由以下几个部分构成:

(a) 词典配置. 配置用户字典文件 userdict. txt 和系统配置文件 Configure. xml.

(b) 结果验证. 通过分割一段文字,来实际检验效果. “用户来电反映,之前通过上海市世纪联华服务充值 50 元,但现用户发现未到账 .”,最后可划分为 “用户/来电/反映,之前/通过/上海市/世纪联华/服务/充值/50 元,但/现用户/发现/未到账”.

(c) 模型演练. 通过对 1000 条投诉文本的演练,将经过分词处理的文本,进行统计、汇总. 去除其中

部分特殊”高频“;主要包括常见的结构助词等等;同时去除词频很小的一些划分词. 通过以上方法获得了 2000 多个关键词,再通过人工的干预,将具有相同意义的词语进行组合、合并,最终得到了 897 个的关键词.

2.3.2 权重赋值

TF-IDF(term frequency-inverse document frequency)[5-7]是一种用于信息搜索和信息挖掘的常用加权技术. TF-IDF 的主要思想是:如果某个词或短语在一篇文章中出现的频率 TF 高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类. TF 词频(Term Frequency)指的是某一个给定的词语在该文件中出现的次数. IDF 反文档频率(Inverse Document Frequency)的主要思想是:如果包含词条的文档越少,IDF 越大,则说明词条具有很好的类别区分能力.

(a) 计算词频. 通过对 897 个关键词的词频的统计(公式 1).

$$\text{词频(TF)} = \frac{\text{某个词在文中出现的次数}}{\text{文章的总词数}} . \tag{1}$$

得到了每个关键词的 TF 值(如图 2),TF 值越大的话,也就说明了该词在文本中出现的次数越多,也就意味着该词更加能代表文本所要表达的意思.

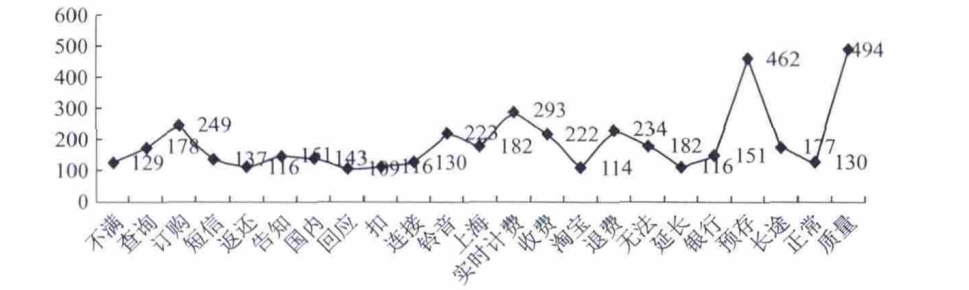

图 2 关键词对应 TF 的值

(b) 计算逆文档频率. 选取 1000 条投诉文本的记录内容作为语料库(corpus),来模拟出语言的使用环境,采用如下公式进行计算(公式 2):

$$\text{逆文本频率(IDF)} = \log \frac{\text{语料库的文档总数}}{\text{含关键词的文档数}(+)} . \tag{2}$$

如果 1 个词语越常见,那么分母就越大,逆文档频率就越小越接近 0. 通过对 897 个关键词的词频的统计(图 3),计算出对应的 IDF 的值作为统计的依据.

(c) 计算 TF-IDF. TF-IDF 与一个词在文档中的出现次数成正比,与该词在整个语言中的出现次数成反比(公式 3).

$$\text{TF} - \text{IDF} = \text{TF} \times \text{IDF} . \tag{3}$$

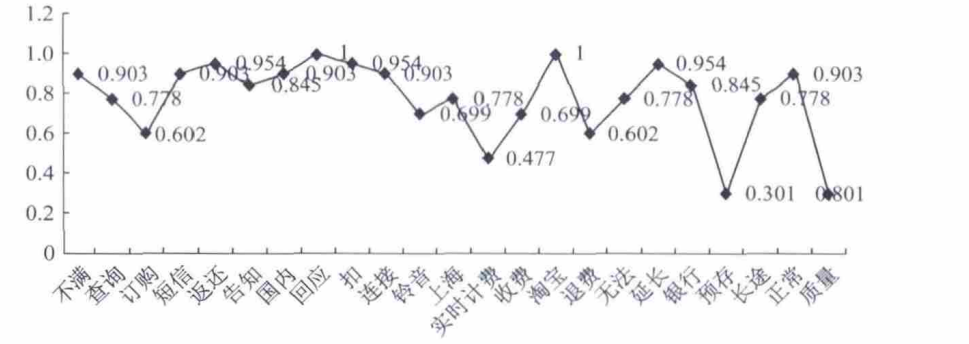

图 3 关键词对应的 IDF 的值

所以,自动提取关键词的算法就是计算出文档的每个词的 TF-IDF 值,然后按降序排列,取排在最前面的几个词(图 4) .

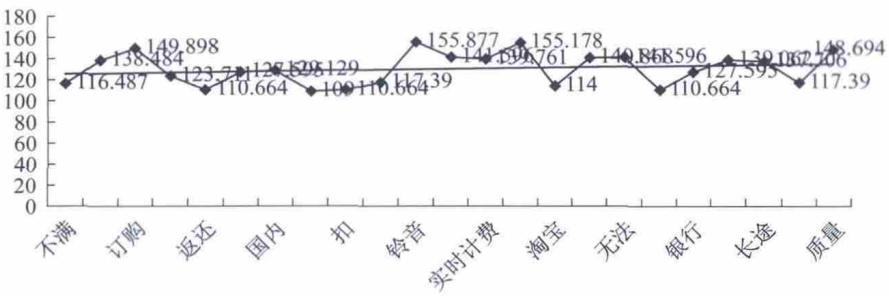

图 4　关键词对应的 TF-IDF 的值

2.3.3　特征选取

根据生成的 TF-IDF 排列倒序表,用尽可能少的词语表示文本,要求尽可能减少语义丢失,能够将不同语义的文本区分开来. 从训练集中初步梳理出各类别对应的特征词,合并同义词. 通过训练模型,测试集测试结果,调整特征词. 在调整的过程中主要注意以下两点: (a) 消除没有区分度的特征词; (b) 调整容易混淆的类.

最终给每个特定的类别定义其关键字,来加以区分(表 1) .

表 1　不同类别的区分

投诉类别 1	投诉类别 2	投诉类别 3	关键词	同义词
基础通信	G3 上网本上网问题	G3 上网本无法上网	上网本/无法/慢	失败,不可,不行使用,连接,登陆
基础通信	G3 上网本上网问题	G3 上网本上网速度慢	上网本/上网/慢	速度,上网,页面很卡,困难

2.4　知识挖掘

知识挖掘的过程主要进行的是对分词出来的结果进行分类,形成具有区分度的不同投诉类别.

2.4.1　LIBSVM 模型训练

LIBSVM [8] 是台湾大学林智仁副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包. 利用开源分类工具 – LIBSVM,核函数采用 RBF 函数: $\exp(-r|u-v|^2)$,进行模型训练,最后训练得到的模型文件* . range、* . model 文件,并且创建类别代码维表(表 2) ,由于篇幅问题,只取其中的 2 条分类路径加以说明.

表 2　类别代码维表

投诉类别 1	投诉类别 2	投诉类别 3	预测码
基础通信	G3 上网本上网问题	G3 上网本上网信号差	464
基础通信	G3 上网本上网问题	G3 上网本无法上网	456

2.4.2　对新增投诉文本进行权值赋值和特征表示

新增的投诉文本(格式: 投诉编号 – 投诉内容) 进行相应的分词、权值赋值特征表示之后,输入到 LIBSVM 软件之中进行比对,进行模型的预测和类别的输出(表 3) .

表 3 测试文本输出结果

预测编码	编号	待分类投诉文本
464	2012 - 06 - 17 - 2184	用户反应在浦东南泉北路 588 号地区信号差问题. 交涉后对方承认有问题,但无法解决,只同意延长一个月使用时间. 消费者不予接受,求助尽快解决.
456	2012 - 06 - 29 - 1517	用户来电反映所在地段无法连接上网,要求上门检测,请查证并处理,谢谢!

2.4.3 预测类别与文本对应

将类别编码维表和测试文本输出结果表按照预测结果进行对应,将测试投诉文本对应到 3 层投诉类别(表 4).

表 4 测试文本分类结果

预测码	编号	投诉内容	第一层	第二层	第三层
464	2012 - 06 - 17 - 2184	用户反应在浦东南泉北路 588 号地区信号差问题. 交涉后对方承认有问题,但无法解决,只同意延长一个月使用时间. 消费者不予接受,求助尽快解决.	基础通信	G3 上网本上网问题	G3 上网本上网问题
456	2012 - 06 - 29 - 1517	用户来电反映所在地段无法连接上网,要求上门检测,请查证并处理,谢谢!	基础通信	G3 上网本上网问题	G3 上网本无法上网

3 实验结果检测与分析

3.1 评估指标选择

目前有多种方法来评估文本挖掘,下面列出几种比较公认的评估方法和指标(表 5).

表 5 检测指标

指标	计算方法
分类正确率	计算文本样本与待分类文本的概率得出分类正确率
查准率	正确分类的对象所占对象集的大小
查全率	集合中所含指定类别的对象数占实际目标类中对象数的比例
F - score	查准率和查全率的调和均值(查全率* 查准率) / [查全率 + 查准率) /2]

分类正确率主要针对分词技术,投诉热点智能分类采用的是目前普及率和好评率较高的开源分词系统,因此不考虑分类正确率指标. 同时,投诉热点模型主要创造并演进了分类算法,在分类算法中不考虑查全率指标(查全率默认为 100%),因此模型的评估主要采用了查准率的指标(公式 4).

$$P_i = \frac{A_i}{A_I + B_i} \quad . \tag{4}$$

P_i 为正确分类的导航量,B_i 为错误分类的导航量.

3.2 统计结果

采用本文作者所阐述的方法对不同类型的投诉文本进行处理,形成的三级导航路径,将其与人工分类导航进行比对,进行准确率的统计(表 6).

表 6 准确率统计

分类器	涵盖样本/%	准确率/%
宽带	8	82
信号	15	80
地址	9	89
上网	15	78
订购	20	76
费用	15	90
其他样本	26	75
总计	98.8	81.4

4 结 语

投诉热点智能监控模型主要应用了文本挖掘中的两类核心技术: 文本分词技术和分类技术. 通过文本挖掘技术,以达到将投诉文本智能分类的目的. 通过一个类型的投诉导航文

本的计算,已完成模型的生成,又随机取出几个投诉文本来进行检测,说明了该方法的有效性,最后统计出了不同类型分类下的准备率.

常见的分类方法包括: 最邻近分类(KNN) 、特征选择方法、贝叶斯分类、支持向量机(SVM) 和基于关联的分类. 着重讲述了基于 SVM 的文本分类的方法,在后期的实践过程中,将综合其余的几种常见文本分类方法进行相关性的研究,不断地优化当前设计模型,以达到更好的分类效果.

参考文献:

[1] 范明,孟小峰. 数据挖掘概念与技术 [M] . 北京: 机械工业出版社,2001.

[2] 方辉,王倩. 支持向量机的算法研究 [J] . 长春师范大学学报: 自然科学版,2007,26(3) : 90 - 91.

[3] 王兴起,王维才,谢宗晓等. 文本挖掘技术在信息安全风险评估系统中的应用研究 [J] . 情报理论与实践,2013,36(4) : 107 - 110.

[4] 夏天,樊孝忠. 利用 JNI 实现 ICTCLAS 系统的 Java 调用 [J] . 计算机应用,2004,24(2) : 178 - 182.

[5] 徐凤亚,罗振声. 文本自动分类中特征权重算法的改进研究 [J] . 计算机工程与应用,2005,41(1) : 181 - 184.

[6] 景丽萍,黄厚宽,石洪波. 用于文本挖掘的特征选择方法 TF-IDF 及其改进 [J] . 广西师范大学学报: 自然科学版,2003,21(1) : 142 - 146.

[7] 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: 自然科学版,2012,40(6) : 158 - 160.

[8] 吴其叶. 科技查新的查准度和查全度与文献检索的查全率和查准率的差异 [J] . 现代情报,2003,23(9) : 8 - 9.

[9] 朱培根,梅卫江,石秀锋等. 基于 LIBSVM 代用燃料有效功率增量预测方法的研究 [J] . 石河子大学学报: 自然科学版,2012,30(5) : 657 - 660.

Hot complaint intelligent classification based on text mining

XIA Haifeng, CHEN Junhua

(College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China)

Abstract: The complaint recognizer system plays an important role in making sure the correct classification of the hot complaint, improving the service quantity of telecommunications industry. The customers' complaint in telecommunications industry has its special particularity which should be done in limited time, which cause the error in classification of hot complaint. The paper presents a model of complaint hot intelligent classification based on text mining, which can classify the hot complaint in the correct level of the complaint navigation. The examples show that the model can be efficient to classify the text of the complaint.

Key words: text mining; intelligent classification; complaint

(责任编辑: 包震宇)