网络产品评论挖掘研究①

单晓红, 杨 柳

(北京工业大学 经济与管理学院, 北京 100124)

摘 要: 以有效分析和挖掘网络产品评论中的用户观点从而为消费者和商家均提供有价值的信息为目的, 提出了网络产品评论挖掘的步骤和方法, 并在用户产品评论分析的基础上, 进一步对产品特征词的关注度和极性进行分析, 实现了更加全面地产品评论挖掘. 最后以 iphone 4s 为例对所提出的方法进行了实验, 验证了该方法的可行性.

关键词: 网络产品评论; 语料; 特征词; 极性; 极性强度

Research on Online Product Review Mining

SHAN Xiao-Hong, Yang Liu

(The College of Economics and Management, Beijing University of Technology, Beijing 100124, China)

Abstract: To analyze and mining online product review effectively and provide valuable information to both consumers and companies, this article gives the steps and methods of online product review mining. Besides the traditional review analysis, the attention degree and polarity of characteristic words are given further to achieve the more comprehensive product review mining. At last take iphone 4s mobile for example to do the experiment to verify the given method, the result is consistent with the reality.

Key words: online product review; corpora; characteristic words; polarity; polarity web

1 引言

网络生活的普及与深入改变了人们的生活方式, 特别是进入 Web2.0 时代, 各种网络社区、微博、博客、评测网站平台的兴起, 使得网民的参与热情大大提高, 利用网络平台发表自己的观点, 分享信息和资源, 互联网成为人们表达情感和观点的主要工具.

随着用户评论的激增, 商家和消费者对它的重视程度越来越高. 为了方便消费者之间交流购物心得, 以及消费者和商家之间的沟通, 很多电子商务网站设置了用户评论区, 第三方的评论网站也大量涌现. 网络产品评论的挖掘对消费者和企业都具有重要的参考价值. 一方面, 网络用户对产品的评论是基于用户的亲身体验, 其中包含对产品或服务的性能、质量、用户体验等评价, 是消费者获取产品和服务信息的一种新途径, 能够在一定程度上消除消费者购物时的心理不安全感和不信任感, 帮助潜在消费者判断产品或服务是否满足其需求, 进而做出合理的购买决策; 另一方面, 商家可以通过用户评论掌握和跟踪用户的需求、喜好及其变化趋势, 促进企业针对用户评论中产品或服务的不足有针对性的改进, 提高用户满意度, 也为企业产品研发、市场战略的制定提供参考.

网络产品评论数量庞大、内容复杂, 有些评论中包含了大量无用信息, 表达用户观点的信息只占一小部分. 如何能够采用自动或半自动化地方法对产品评论进行有效分析和处理, 挖掘出对消费者和企业都有用的信息称为学者们和业界普遍关注的问题.

2 网络产品评论挖掘研究现状

产品评论挖掘是基于文本挖掘技术的, 它的研究兴起于 2002 年, Turney 首先提出了使用语义倾向性将评论分为推荐和不推荐[1]. 目前的评论挖掘研究主要集中于两个方面: 一个方面是对产品评论挖掘各阶段任务所采

① 基金项目:北京市教委 2013 年度人文社科面上项目(SM201310005002);北京市自然科学基金项目(9112001)
收稿时间:2013-06-29;收到修改稿时间:2013-09-25

用的技术和方法进行研究, 如产品特征提取[2-4]、主观句定位[5,6]、用户态度提取[7,8]、用户态度极性分析[9-11]等等; 另一个方面是产品评论挖掘应用的研究, 如 2004 年 M. Hu 和 Bing Liu 提出网络用户的产品评论挖掘是指通过机器从大量的网络用户产品评论中自动地获取所关注的产品特征, 并利用该方法对于手机、数码相机等产品评论进行挖掘[12]. 之后也有学者采用机器学习的传统文本分类方法进行评论挖掘的研究, 如 Pang 等采用朴素贝叶斯分类、最大熵和支持向量机的方法对评论进行了文本倾向性分析[13], 徐琳宏等采取支持向量机和词频加权统计两种分类方法对网上搜索的 499 篇影评进行了文本倾向性的判别[14], 李实、叶强等针对中文的特点, 提出了基于改进关联规则算法的中文产品评论信息挖掘方法, 并针对手机、数码相机、书籍等 5 种产品的评论语料对该方法进行了数据实验[15], 吴丽华等采用情感分析技术, 提出基于客户感知价值的产品特征挖掘算法, 实现对于评论中 IT 产品特征及其情感倾向的语义分析、动态提取和综合信息挖掘, 并根据用户的关注权重将产品特征和情感倾向进行排列[16].

可以说目前的产品评论挖掘技术已经日渐成熟, 如何能够利用这些技术更深入细致地分析和挖掘出产品评论中隐藏的信息是我们亟需解决的问题. 本文正是从这个角度出发, 在用户产品评论情感分析的基础上, 进一步地对产品特征词的关注度和极性进行分析, 更加全面地对用户的产品评论进行分析, 从而为商家改进产品和服务提供更加有效的决策帮助.

3 网络产品评论挖掘步骤

网络产品评论挖掘分为用户产品评论语料库构建、数据预处理、产品评论挖掘和挖掘结果分析四个步骤, 如图 1 所示.

3.1 用户产品评论语料库构建

通过自动或半自动方式采集网上产品的评论数据, 构建网络产品评论语料库. 具体包括的评论属性有: 评论人、评论内容、发布时间、评论数量.

3.2 数据预处理

网络采集到的产品评论数据量大, 含有很多噪声或者是与观点无关的信息, 数据清理、删除停用词可以提高产品评论挖掘的效率和效果; 同时产品评论语料库中的信息无法直接采用情感分析的方法对其进行处理, 还需要对其进行一些必要的转换, 如分词、词性标注才能进行有效的评论挖掘. 因此数据预处理是保证产品评论挖掘有效性的前提.

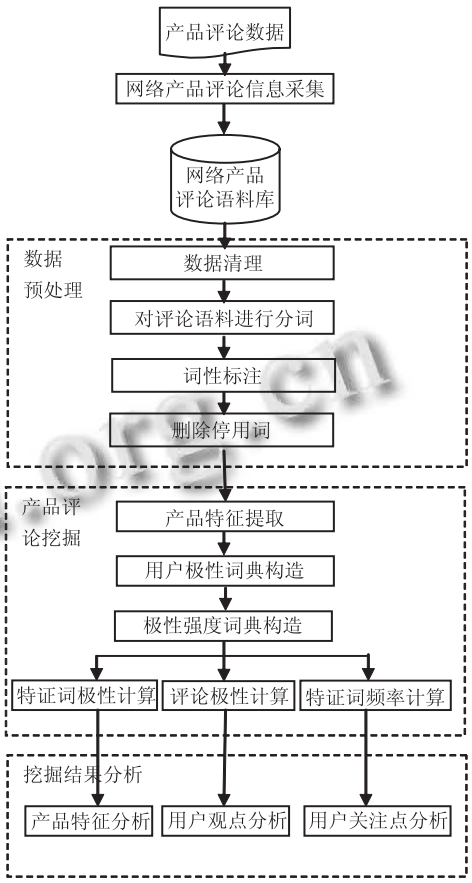

图 1　网上产品评论挖掘步骤

(1) 数据清理

数据清理工作一是去掉产品评论中的视频、音频、图片等非文本信息; 二是去掉非评论信息, 如广告等等. 因为产品评论挖掘主要是挖掘用户关于产品或者服务的观点, 所以与产品特征和用户观点无关的信息可以视为噪声删除.

(2) 对评论语料进行分词

词语是最小的能够独立使用的有意义的语言成分, 为了能够对用户的每条评论句子进行情感分析, 需要进行分词. 分词的好坏直接决定了计算机对文本语义分析的准确性. 本文采用了武汉大学 ROST 团队发布的 ROST WordParser 进行分词.

(3) 对分词后的语料进行词性标注

词性标注是产品评论挖掘的基础, 通过词性标注, 可以判断评论语料是属于特征词、观点词, 还是程度词, 从而帮助我们提取出产品特征和识别用户观点,

并对用户观点的极性进行判断.

(4) 删除停用词

评论存在了很多出现频率高但是实际意义不是很大的词, 如“他”、“了”等等, 称之为停用词, 删除停用词会提高产品特征分析的效果和效率.

3.3 产品评论挖掘

这一步骤是产品评论挖掘的核心, 其实质是对于用户的评论进行极性计算, 从而对用户的观点进行分析.

(1) 产品特征提取

产品特征提取包括两个部分, 一是产品自身特征, 即它的固有属性, 这部分特征由规格说明书中进行提取; 二是用户评论特征, 这部分是由网上的用户评论中提取. 网络产品评论的随意性很大, 完全依赖产品规格说明书中的产品特征是远远不够的, 必须将用户评论中的产品特征提取出来, 本文对于评论中产品特征的提取是在分词和词性标注的基础上, 首先提取出全部的名词, 找出出现频率比较高的名词作为候选词汇, 然后通过人工定义, 最后得到用户关注产品特征集合 F_2. 假设产品规格说明书中提取的产品特征集合定义为 F_1, 则最终产品特征集合为 $F=F_1 \cup F_2$.

(2) 用户极性词典构造

极性词是指句子中带有情感倾向的词语, 是用户表明用户自身观点和态度的词汇, 如“坏”、“好”等. 本文以《知网》的词库为基础, 选取与要分析的产品质量、外观、性能等主题相关的词汇构成用户极性词典. 这些词语是最基本的包含强烈褒贬感情色彩的词语, 是大量语料中总结出来最常用的极性词语.

本文中极性词典中极性词的极性计算如下: 从极性词典中选择 n 对基准词, 每对基准词含有一个正面评价词语和一个负面评价词语, 定义 P_j 为第 j 对基准词中的正面基准词, N_j 为第 j 对基准词中的负面基准词, 则 $T(pw)$为极性词 pw 的极性值, 极性值大于 0 为正面评价, 表示对产品持认可、赞扬的态度; 极性值小于 0 为负面评价, 表示对产品持否定、批评的态度. 具体计算如下:

$$T(pw)=\sum_{j=1}^{n}\left[Sim(P_j, pw)-Sim(N_j, pw)\right]$$

其中, $Sim(P_j, pw)$为极性词 pw 与第 j 对基准词中正面基准词 P_j 的语义相似度, $Sim(N_j, pw)$表示极性词 pw 与第 j 对基准词中正面基准词 N_j 的语义相似度, 并采用中科院提供的词语相似度的计算工具 Word Similarity 进行计算.

(3) 极性强度词典构造

极性强度词是指产品评论中用于加强语气的程度副词, 如“比较”、“非常”等和一些否定词, 如“不”、“没有”等, 这些词的修饰加强或者减弱, 甚至改变了原来词汇的极性, 在极性判断时一定要考虑这些极性强度词语的作用. 本文以《知网》的程度级别词语和汉语中常用否定词构造极性强度词典, 不同的程度副词和否定词赋予不同的权重, 其中程度副词的权重表明了程度副词对极性的增强(或者减弱)作用, 否定词的权重表明对用户持有的感情色彩的极性(褒义或者贬义)进行反转. 具体计算方法如下.

设 $T(pw)$为词语 pw 的极性, $T'(pw)$为极性强度词语 dg 修饰后的词语 pw 的极性, 设极性强度词语的权重为 $D(dg)$, 则极性强度词语 dg 修饰的词语 pw 的极性为 $T'(pw)=D(dg)\times T(pw)$. 其中, 当极性强度词为否定词时, $D(dg)\in(-1,0)$.

(4) 网络产品评论的情感分析

为了全面地分析网络产品评论的意义, 需要从特征词频率、特征词极性和评论极性三个方面进行产品评论的情感分析计算.

首先构造语法模式集合. 在李存青给出的 4 种汉语评论常用语法模式基础上[17], 我们给出了 9 种语法模式: 名词+形容词、名词+动词、名词+副词+形容词、名词+副词+动词、名词+副词+副词+形容词、名词+副词+副词+动词、动词+名词、副词+动词+名词、形容词+的+名词, 本文设定评论语法模式集合 $GP=\{gp_1,\cdots, gp_9\}=\{$(n+adj), (n+v), (n+adv+adj), (n+adv+v), (n+adv+adv+adj), (n+adv+adv+v), (v+n), (adv+v+n), (adj+de+n)$\}$在分词和词性标注的基础上, 对于每一条产品评论句子 R_i, 构造其语法模式集合为 $rgp_i=\{gp_{i1},\cdots gp_{ij},\cdots,gp_{in}\}$, 其中 $gp_{ij}\in GP$.

在定义好语法模式后, 就可以进行评论句子的清理. 这里主要清理两类句子: 第一, 如果第 i 个句子 Ri 的语法模式集合为空 $rgp_i=\phi$, 则删除掉句子 R_i; 第二, 对于第 i 个句子 R_i 的语法模式集合 rgp_i, 如果其所有元素中的名词都不属于产品特征集合, 则删除掉句子 R_i.

1) 特证词频率计算

第 k 个特征词的频率为所有产品评论句子中找到符合给定的 9 种语法模式的特征词 W_k 出现的次数, 记为 $f(W_k)$.

2) 特证词极性计算

每个特征词的极性为该特征词的极性平均值. 设第 k 个特征词的极性为包含第 k 个特征词的语法模式的极性之和除以该特征词的频率, 即 $T(W_k)=\sum_{j=1}^{f(W_k)}T(gp_{kj})\Big/f(W_k)$, 其中 T($gp_{kj}$)为第 k 个特征词 Wk 的第 j 个语法模式的极性.

3) 评论极性计算

在 3.2 中给出了每个极性词的极性计算方法, 评论极性计算包括两个部分, 一个是对每条网络产品评论的极性计算; 二是对所有网络产品评论的极性计算.

每条产品评论极性 $T(R_i)$为第 i 条评论 R_i 的每个语法模式的极性之和, 即 $T(R_i)=\sum_{j=1}^{n}T(gp_{ij})$, 其中 T($gp_{ii}$)为第 i 条评论 R_i 的第 j 个语法模式的极性.

所有网络产品评论的极性 $T(P)$为所采集的所有有效评论的极性之和, 即 $T(P)=\sum_{i=1}^{m}T(R_i)$, 其中 $T(R_i)$为第 i 条评论 R_i 的极性.

3.4 挖掘结果分析

网络产品评论主要是通过用户对网络产品的评价文本进行情感分析, 找到用户感兴趣的内容, 了解用户对产品的哪些特征持肯定态度、哪些特征持否定态度, 以及对产品的综合评价, 进而帮助商家有针对性地改进产品和服务, 以及为其他用户挑选产品和服务时提供参考信息. 所以本文的分析从三个方面进行:

(1) 用户关注点分析. 通过计算特征词频率可以看出用户关心的是产品或服务的哪些特征, 从而找到用户的关注点, 着力改进.

(2) 产品特征分析. 通过计算出每个特征词的相应得分, 能够看出用户对特定产品或服务某一特征评价的好坏程度, 从而判断产品在哪些方面做得好, 哪些方面还需要改进.

(3) 用户评论分析. 通过计算每个用户评论的得分, 可以看出每个用户对特定产品或服务的认可度.

4 实例

iphon4s 是 2012 年热卖的手机, 用户对这部手机众说纷纭, 褒贬不一, 因此本文选取这部手机作为产品评论的对象, 对 2012 年 3 月 15 日 14:30 采集到的太平洋电脑网上的用户关于 16GB iphone4s 产品评论数据进行分析, 从而了解用户对该款手机的评价.

4.1 数据预处理

在删除掉非评论信息后, 采用了武汉大学 ROST 团队发布的 ROST WordParser 5.8.1.9 内测版进行分词. 经过对名词的筛检, 选出有意义的, 可评论的特征词.

首先从 iphone4s 的官方网站上关于产品规格说明书中提取产品特征, 其产品特征词汇集合为 F1={屏幕、价格、游戏、功能、流畅、处理器、电池、软件、升级、外观 摄像头、信号、系统、设计、硬件、性能、操作、照相、音乐、体验、操作系统、版本、芯片、应用、界面、配置、内存、视频}.

然后在用户的评论中挖掘产品特征, 启动 ROST WordParser 5.8.1.9 软件进行特征词提取, 选择出现词频大于 2 的名词集合得到 F_2={屏幕、价格、手机、游戏、功能、流畅、处理器、苹果、电池、软件、升级、外观、速度、摄像头、问题、感觉、信号、系统、设计、硬件、性能、操作、时间、运行、语音、照相、体验、手感、产品、流量、做工、使用、用户、乔布斯、外形、实用、质量、人性化、机子、版本、芯片、效果、应用、联通、卡片、界面、电脑、三星、短信、配置、图形、助理、耳机、公司、程序、架构、电影、智能、节电、散热、显示、画质、智能机、打字、国内、性价比、兼容性、摄像、充电、内存、网页、视频、画面}.

最终产品特征集合为 $F'=F_1\cup F_2$ 后人工去掉与分析主题不相关的特征词, 再合并一个代表意思的词汇, 如: F_{12}={像素、拍照、照相、照片、画面}={像素}, 最后得到 F={屏幕、游戏、操作、功能、人性化、性能、画质、网速、性价比、电池、像素、信号、价格、做工、配置、外观、手感、体验、系统、质量、音乐}.

4.2 词典构造

(1) 用户极性词典构造

从《知网》的正面情感词语、正面评价词语、负面情感词语、负面评价词语中选取了与要分析的 iphone4s 相关的词汇以及手机评价常用词汇构成用户极性词典. 其中选择 10 对基准词{(清晰,模糊)、(流畅,卡)、(简便, 繁琐)、(实用,没用)、(强大,低级)、(便宜,昂贵)、(长,短)、(大,小)、(精细,粗糙)、(丰富,稀少)}.

(2) 极性强度词典构造

从《知网》的程度级别词语中选取了 212 个程度

副词和 15 个否定词, 并将程度副词分为 10 个级别及其权重分别为: 极其/最(2.0)、非常/太(1.9)、很(1.8)、更(1.7)、偏/较(1.6)、适当(1.5)、稍(1.4)、欠(1.2)、超(1.1)、有点(0.8).

4.3 产品评论分析

(1) 用户关注点分析

用户关注点分析是通过特征词频率来找到用户对 iphone 4s 手机的关注热点. 表 1 所示为出现频率在 9 次以上的特征词, 即: 屏幕、价格、手机、游戏、功能、流畅、处理器、苹果、电池、软件、升级、外观、速度、摄像头、信号、系统、设计、硬件、性能、操作, 说明这些产品特征是用户的主要关注点. 商家应该在这些方面给予足够的重视. 在这些方面的技术工作做好, 才能让产品更加迎合市场需求.

表 1 特征词出现的频率

特征词	频率(次数)	特征词	频率(次数)
屏幕	25	升级	14
价格	24	外观	12
手机	22	速度	11
游戏	18	摄像头	11
功能	15	信号	9
流畅	15	系统	9
处理器	15	设计	9
苹果	15	硬件	9
电池	15	性能	9
软件	14	操作	9

(2) 产品特征分析

产品特征分析是产品评论分析的关键内容, 通过对用户产品特征的分析, 可以看出用户对 iphone 4s 手机的哪些产品特征持肯定态度, 对哪些产品特征持否定态度. 经过计算, iphone 4s 手机的各特征词极性如图 2 所示.

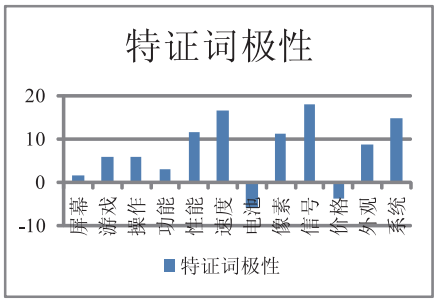

图 2 iphone4s 的特征词极性

从图 2 可以看出, 除了电池和价格外, 其他特征词极性均为正值. 其中信号、系统、性能、速度这些特征词的极性较高, 即用户对 iphone4s 这款手机的这些特征非常满意; 而电池、价格方面可能是 iphone4s 这部手机中令用户不满意的地方, 这与现实相符. 这就为苹果公司提供了一个改进的方向, 同时也为挑选手机的用户提供了一个参考, 如果想挑选价格便宜或者待机时间长的手机, 则 iphone4s 不是一个好的选择.

(3) 用户评论分析

以第 12 条评论为例: 优点: 比较喜欢 4S 外观, 看着干净大方, 虽然和 iphone 4 比没什么大的变化, 功能提升了很多, 还是很强大的. 缺点: 一是价格贵, 二是 siri 不支持中文 配置欠缺. 总结: 目前还没有需求, 再便宜点就更好了.

该评论得分为: 外观: 1.6*100=160, 功能: 1.8*100=180, 价格: －1*100=－100, 配置: －1.2*100=－120, 总得分: 120/4=30>0, 说明该用户对 iphone4s 持肯定态度, 大体上还是认同这部手机的.

依此类推, 在所采集的 43 条有效评论中, 35 条总分大于 0 的评论, 三条总分等于 0 的中立评论, 5 条小于 0 的评论, 积极评论占总体的 81.4%, 用户对 iphone4s 的评价还是很高的.

此外, 根据前述分析, “价格”是关注度高且产品特征得分最低的, 这说明价格是影响消费者购买行为的一个重要因素, 但是对于一些对手机部分特征, 如信号、速度等有偏好的用户来说, 当产品的某些特征令其非常满意时, 则可以忽略价格进行选择. 这也给手机制造商带来一个启示, 一定要有特色才能吸引消费者.

5 结论

网络产品评论挖掘是从用户庞杂的网上评论中找到用户对产品和服务的看法的一个有效途径. 本文给出了包括用户产品评论语料库构建、数据预处理、产品评论挖掘和挖掘结果分析四个步骤的网络产品评论挖掘过程, 及各个步骤的详细方法, 并以 iphone 4s 产品为例对其网络产品评论进行了实验和分析, 实验结果表明了该方法可以为商家和消费者提供一些有价值的参考.

参考文献

1 Turny P. Thumbs up or thumbs down? Semantic orientation

applied to unsupervised classification of reviews. Proc. of the Association of Computational Linguistics (ACL02). Philadelphia. 2002. 417–424.

2 Popescu AM, Etzioni O. Extracting product features and opinions from reviews. HLT/EMNLP 2005. 2005. 339–346.

3 Hu M, Liu B. Mining opinion features in customer reviews. Proc. of 19th National Conference on Artificial Intelligence(AAAI-2004). San Jose, USA. 2004. 755–760.

4 黄永文,何中市,伍星.产品特征的层次关系获取.计算机工程与应用,2009,45(22):236–240.

5 Kim SM, Hovy E. Identifying and analyzing judgment opinions. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics Proceedings. New York, USA. 2006. 200–207.

6 Bethard S, Yu H, Thornton A. Extracting opinion propositions and opinion holders using syntactic and lexical cues. Computing Attitude and Affect in Text: Theory and Applications, 2006, 20: 125–141.

7 Ramirez J, Segura JC, Gorriz JM. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. IEEE Trans. on Audio Speech and Language Proceeding, 2007, 15(8): 2177–2189.

8 Jong WS, Hyuk JK, Suk HJ. Voice activity detection based on conditional MAP criterion. IEEE Signal Proceedings Letters, 2008,15: 257–260.

9 徐琳宏,林鸿飞,杨志豪.基于语义理解的文本倾向性识别机制.中文信息学报,2007,21(1):96–100.

10 叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究.信息系统学报,2007,1(1):79–91.

11 建立,慈祥,黄剑雄.网络评论倾向性分析.计算机应用,2010,30(11):2937–2940.

12 Hu M, Liu B. Mining opinion features in customer reviews. American Association for Artificial Intelligence, 2004: 755–760.

13 Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002). 2002. 79-86.

14 徐琳宏,林鸿飞,杨志豪.基于语义理解的文本倾向性识别机制.中文信息学报,2007,21(1):96–100.

15 李实,叶强,李一军,Rob Law.中文网络客户评论的产品特征挖掘方法研究.管理科学学报,2009,12(2):142–151.

16 吴丽华,冯建平,曹均阔.中文网络评论的 IT 产品特征挖掘及情感倾向分析.计算机与数字工程,2012,40(11):52–54,131.

17 李存青.中文意见挖掘中的特征词提取以及情感倾向分析[硕士学位论文].重庆:重庆大学,2010.

(上接第 76 页)

4 Sadur CN, Moline N, Costa M, et al. Diabetes management in a health maintenance organization. Efficacy of care management using cluster visits.Diabetes Care,1999.

5 乌聪敏,幺莉,林济铿.基于改进 Web 三层结构的电力技术监督系统设计与实现.电力自动化设备,2010,30(2):118–122.

6 殷欣,卢广文.糖尿病病案数据库的系统设计.第一军医大学学报,2004,24:713–715.

7 咎风彪.在 ASP.NET 中给予角色的权限控制设计与实现.IT 技术论坛,2008,25(2):84–85.

8 刘辉,郝伟.一种基于.NET4.0Chart 的通用图形化统计模块的设计与实现.电脑知识与技术,2011,(7):7676–7680.