

Email 中基于英文文本的情感监测系统研究

于冬梅

(江苏技术师范学院电气信息工程学院, 江苏 常州 213001)

摘 要: 对情感计算应用方面研究进行拓展和延伸. Email 是 Internet 应用最广的服务之一, 在 Email 中嵌入基于文本内容的情感监测系统有利于构建和谐的人机交互环境. Email 中的语句集来源于数据库 Commonsense. 对数据库中的语句进行了筛选, 根据这些筛选出来的语句结构创建了四种语句结构模型, 最后通过程序语言实现了英文文本的情感监测. 在系统实现过程中, 用颜色去标识文本的情感倾向.

关键词: 电子邮件; 语句结构模型; 情感监测系统; 颜色心理学

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)12-0161-06

Research and Realization about the Emotion Monitoring System of English Texts in Email

YU Dong-mei

(School of Electricity & Information Engineering,
Jiangsu Teachers University of Technology, Changzhou 213001, China)

Abstract: This is to develop and extend the application of affective computing. Email is one of the most widely used applications on the Net. We design the emotion monitoring system based on Email texts, which is good for harmonious environment of HCI. The sentences in Email come from Commonsense. We chose some sentences from Commonsense. Four structure models are created according to these sentences. At last, we realized emotion monitoring of texts with program language. In the procedure of system implementation, we mark the contextual emotions with colors. This method is one novelty of the paper.

Key words: Email; sentence-structure model; emotion monitoring system; color psychology.

1 引言

电子邮件(Email)是 Internet 应用最广的服务之一^[1]. 尽管 Email 已经作为一种交流的媒介在全世界上被人们广泛的应用, 然而, 许多用户在发邮件的过程中仍然存在诸多的不满. Dow 等^[2]曾在美国消费者满意度索引的调查报告显示, 美国消费者在电子邮件的使用过程中的满意率仅为 65.5%, 这个数字低于大多数其他产品和服务. 如果在 Email 的用户界面中添加情感监测系统, 即用一个带有情感发现和情感表达的用户界面 Agent 来监测邮件

中的悲伤、愤怒、挫折、幸福等情绪, 那么就会使用户不需要再仅仅面对一个仅有文本的冰冷的计算机屏幕上的四方框.

本文以 OMCS 中的常识知识作为 Email 的情感词汇库, 研究了 Email 中英文语句的结构建模, 并设计了基于英文文本的 Email 情感监测系统. 该系统的描述结果不再是简单的曲线, 而是引入了颜色心理学的内容, 用不同的颜色描述 Email 文本的情感倾向. 如果将该系统嵌入到现有的邮件系统中, 将会具有更远大的应用前景.

2 Email 中基于文本的情感监测系统的研究背景

情感监测系统是人机交互的智能设计中更富有挑战性的一项工作,因为情感监测系统将使人机交互变得更加自然,更加符合交互的社会性.如果人们能够和一台具有识别人类情感的计算机进行交互的话,那么就相当于人与其他人之间的交流和互动^[3-4].

现有的分析文本中的情感倾向的方法大致分为四类:

① 关键词方法.这种方法容易实现,因此是很常用的一种方法.

② 词汇倾向性分析.很明显,词汇倾向性分析的方法比关键词方法更有效.这一方法不是仅仅监测到句子中明显情感词汇,而且通过某种特定情感词汇的概率的“倾向性”来进行识别.

③ 统计学方法.这也是一种在监测文本情感问题的常用的方法.这一方法是利用机器学习算法对带有情感解释的文本进行大量的训练.

④ 手动模型. Dyer^[5] 等认为监测文本中的情感类别是一个很深奥的问题. Dyer 的 DAYDREAM-ER 模型中的情感状态是按照心理学中讲述的人的需要、目标和愿望来手动建模的.

3 英文文本的情感词汇库的建立

3.1 Email 中语句集来源——Commonsense

Commonsense 的全称是 Open Mind Commonsense(OMCS)^[6], OMCS 是由 MIT 多媒体实验室创建的一个知识获取系统,最初是由人工智能专家来搜集日常生活知识,然后将这些知识作为公共的知识库,供研究者使用.

为了使本文下面提出的情感监测系统具有很好的鲁棒性,对模型进行验证的各个文本都是来源于 OMCS 的、与情感相关的生活常识的知识库.用英语句子陈述的生活常识比汉语更简单,更能减少汉语言的复杂性所带来的句子歧义.首先从 OMCS 中提取出一系列包含情感色彩的日常生活常识的句子,作为下面模型的验证文本.

3.2 Email 语句分词模块

对语句进行分词的目的是获取文档中的表达情感的单词.通常语句中存在的不只是单词,还有很多数字和符号,这些数字和符号并不是情感监测系统关注的的数据,应当将其过滤掉,仅留下词.因为本文

研究的仅仅是用英文书写的 Email,因此英文最小的语义单位为单词,在最大长度匹配出表示情感的词汇后,依靠空格就可以进行区分.

对于语句或者文档中无法表达 Email 书写者的情感的词汇称为停用词,例如冠词(a, an, the)、数词(one, two, third 等)、介词(in, up 等)、连词(and, or 等)、代词(this, your, his, her 等)、某些动词(be, do 等)和副词(upward 等),这些词汇通常无法表达邮件发送者的情感倾向,仅仅是英文语法的必要组成部分,因此没有统计的价值.

3.3 英文文本的情感词汇库的设计和建立

为了识别出 Email 中文本的情感倾向,需要建立一张情感词汇表来指导情感识别或者标识.影响文本情感识别研究的一个复杂因素是在海量数据中找到具有代表性的语句.对分析文本的语句的选择主要考虑以下两个方面:一是所选择的句子必须包含某一方面的情感倾向;二是对句子的选择上需要让情感的自由度越小越好,即该句子的情感不能过于模糊.

本文实验语句的选择是通过实验者来进行的.我们在 OMCS 中选择了 300 条语句,挑选了 20 名具有不同学历和在不同专业学习的实验者进行了实验,实验的过程为:首先,将 300 条语句任意分成 20 份问卷,每份问卷有 15 条语句;然后,由 20 名实验者随机抽取,实验者在每条语句中标识出他们认为该语句所具有的情感倾向.实验者把所有的问卷上交以后,我们认为情感特征明显的,是有效样本,最终采集到 257 个有效样本,即后面的实验将对这 257 条语句进行分析和实验.

为了分析构成 Email 的文本中的语句的情感值,首先就要找到大量的描述情感的词汇,然后将这些词汇按照六类基本情感进行预分类.这些经过预分类的表示情感的词汇将成为监测来自于 OMCS^[6] 中的语句的“情感依据”.用来描述人类情感的词汇成千上万,本文仅从 OMCS 的语句中找到的一些具有代表性的情感词汇,并按照前面提到的六类基本情感进行了分类

4 基于 Email 的语句结构模型与文本的情感监测示例

情感是人类认知的一部分,文献[4-5]等都解释了这一观点.心理学家指出,在语言上的情感认知依赖于个体所拥有的传统和文化.因此,人们总是无法理解其他文化背景下的人们表达文化的方式.对

文本的情感特征进行监测一般具有以下两个阶段:一是,对以段落为单位的文本中的每一句子进行情感分类,即将其划分到高兴、悲伤、害怕、厌恶、惊奇和愤怒六类基本情感中;二是,尽可能地优化算法,使上述情感分类过程在较大的文本单元中也同样有效。

本节将结合已有的方法提出一个新的结构——基于文本的自动情感感知机,该方法是在基于大规模的、人们对现实世界中的知识所表达的情感态度的基础上提出来的。如果能将这个情感感知机嵌入到 Email 浏览器中,就能实现对用户邮件中的文本所表达的情感进行实时监测,从而解决了现有电子邮件系统中的枯燥的用户界面。

4.1 Email 中的英文语句结构模型

计算机并不具有人类的智能,人在阅读文章后,根据自身的理解能力可以产生对文章内容的模糊认识,并能对文章的情感基调作以大致的判断,而计算机并不能轻易“读懂”文章,从根本上说,它只认识 0 和 1,所以必须将文本转换为计算机可以识别的格式。假定组成文档的单词在确定的 Email 中相互独立,这样,就可以使用文档中出现的单词集合代替文档。

在识别了一系列的具有情感倾向的、常识性的文本以后,通过下面的模型来监测文本的情感倾向。即 $e_i \in E (i = 1, 2, \dots, 6)$, 其中 $E = \{\text{高兴, 悲伤, 愤怒, 害怕, 厌恶, 惊奇}\}$ 。 e_i 用来标识文本的特定的情感倾向。在这里,我们仅仅采用 Ekman 的六类基本情感分类法,对于诸多的情感分类方法,该模型同样适用,对于其它的情感分类方法来说,只要更新集合 E 的元素即可。然而,并不是文本中的所有句子都带有情感倾向,因此,在 E 中,我们加入了一个无情感倾向的表示,即 $E^* = \{E, \text{Neutral}\}$, 其中,当 $\text{Neutral} = 1$ 时,可以得出该句子没有情感倾向。在下面的模型中用 V 来定义句子的情感倾向值,即 $V = \{v_1, v_2, v_3, v_4, v_5, v_6, n\}$, 其中 $v_1, v_2, v_3, v_4, v_5, v_6$ 分别表示句子中具有高兴、悲伤、愤怒、害怕、厌恶和惊奇等情感倾向的情感值,并且 $v_i \geq 0 (i = 1, 2, \dots, 6)$; n 为 Neutral 的值,即 $n = 1$ 或者 0。对于组成 Email 中的英文文本的句子大致可分为以下几种类型:

(1) 单个形容词类型(One-Word, OW)

该模型中一般是一个词形为形容词的词汇加上一个“!”构成的短句。型如句子“Awful!”。该句表达的就是一种害怕的情感,那么,词汇“awful”本身就

是该句的情感线索词,因此,该句的情感倾向值为 $V = \{0, 0, 0, 1, 0, 0, 0\}$ 。

(2) 动词短语类型(Verb-Phrase, VP)

在英文中,仅仅用单纯的动词短语就能够表达非常强烈的情感倾向,型如句子“Go to the devil.”, 其中,由这四个单词组成的短语就是该句的情感依据,即表达的是一种愤怒的情感,因此,该句的情感倾向值为 $V = \{0, 0, 0, 1, 0, 0, 0\}$ 。

(3) 主语-动词-宾语-宾语类型 (Subject-Verb-Object-Object, SVOO)

这个模型表示的是句子的框架,即该句子中包含主语、动词和宾语(英语中,宾语分为直接宾语和间接宾语)。比如句子,“I give him a book.”, 这个句子可以用下面的框架来表示: [$\langle \text{Subject} \rangle$: I, $\langle \text{Verb} \rangle$: give, $\langle \text{Object1} \rangle$: him, $\langle \text{Object2} \rangle$: a book], 在这个句子中,没有表示情感倾向的词汇,因此该句子按照 SVOO 模型,找不到标识情感倾向的词汇,因此该句的情感倾向值为 $V = \{0, 0, 0, 0, 0, 0, 1\}$ 。再比如句子“We get into a car crash on the way home.”该句可以用如下框架进行表示: [$\langle \text{Subject} \rangle$: we, $\langle \text{Verb} \rangle$: get into, $\langle \text{Object1} \rangle$: a car, $\langle \text{Object2} \rangle$: crash], 在该框架中并没有表示情感倾向的词汇,但是“crash”一词就是情感线索词,因为在英文中对“crash”的解释为“撞碎,猛使…破碎”等,而且一看到“crash”一词,大部分是指发生了车祸或者飞机坠毁等事故,因此,该句标识的情感倾向值为 $V = \{0, 0, 0, 1, 0, 0, 0\}$ 。

(4) 概念提取类型(Concept-Extraction, CE)

在这个模型中,对没有明显的表达情感倾向的词汇的句子将进行概念提取。尤其是带有否定词汇的句子进行概念提取。比如对于简单句,“I am not happy.” 在这个句子中,首先要提取出否定词“not”,然后才能够得出句子的情感倾向。CE 模型没有 SVOO 模型精确,因为 CE 模型中所谓的概念如果脱离了句子的文本内容就会变得毫无意义,但是 CE 模型比 SVOO 模型更加常用,因为在语言表达中,严格遵循 SVOO 模型结构的句子并不是很多。

对 Email 中的英文文本进行情感监测的时候,一般经过两个主要步骤:一是,首先要从 OMCS 中选择一些句子,这些句子中都包含大量的情感词汇,要将这些表达情感的词汇进行预分类,按照情感集合中的六类基本情感进行分类;二是平滑算法,使模型在大规模文本单元中更加有效,从而能够提高文本的情感倾向的监测效率。

4.2 Email 中文本的情感倾向的监测过程

从 OMCS 选取的大量文本可知,有些段落的第一句话带有非常强烈的感情色彩,然而,随着文本内容的深入,文本末尾大多变得毫无“情感”而言.比如,来自于 Tom 写给朋友 Mary 的 Email 内容如下(举例 1):

“Hi! Mary! I find that something is both happy and surprising. For example, playing basketball is exciting. However, playing basketball are typically found at basketball court.”

下面将通过上述模型对该 Email 的文本内容进行情感监测,只所以在本文中讨论的是英文文本,主要原因在于中文文本的切词是一个极其繁复的研究方向,对于本文的英文文本的切词相对来讲就容易多了.以标点符号“!”、“.”和“?”来确定段落中的每个句子,再以句子中的“空格”来确定标识句子情感倾向的词汇.根据模型(1)~(4)以及上面的切分句子和词汇的方法来手动进行 Email 文本的情感监测.监测结果如表 1 所示,

表 1 Email 中的手动情感监测表

编号	语句内容	情感判据词
①	Hi!	Hi
②	Mary!	—
③	I find that something is both happy and surprising.	Happy, Surprising
④	For example, playing basketball is exciting.	exciting
⑤	However, playing basketball is typically found at basketball court.”	—

对于句子①,在 OMCS 中,这类词属于 OW 模型,因此该句的情感值为 $V_1=\{1,0,0,0,0,0,0\}$;句子②的情感倾向值为 $V_2=\{1,0,0,0,0,0,0\}$;句子③中,情感线索词是很明显的两个情感词汇“Happy”和“Surprising”,由此,很容易得知该句的情感值为 $V_3=\{1,0,0,0,0,0,0\}$;同理,句子④的情感值为 $V_4=\{1,0,0,0,0,0,0\}$;句子⑤的情感值为 $V_5=\{1,0,0,0,0,0,0\}$. 根据句子①~⑤,可得文本中的情感词频矩阵为 $A=V_i^T(i=1,2,\cdots,5)$,对稀疏矩阵 A 应用上述 4.1 小节中的模型可知,将标识句子①③④(即将向量 V_1,V_3,V_4 两两做内积)的向量两两内积,结果表明这三个句子有着相同的情感倾向,如图 1 所以,图上的点表示句子中的情感元素.

如图 1 所示,举例 1 中的文本的主体情感倾向

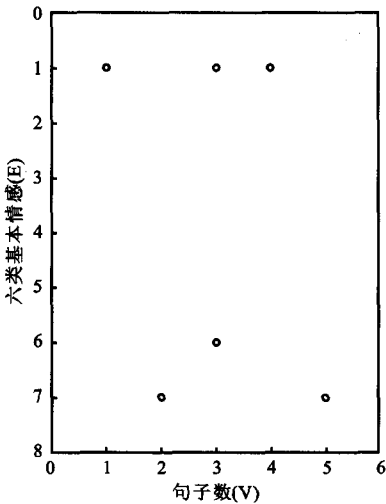


图 1 举例 1 中的情感元素的分布图

为“高兴”,虽然有若干句子没有情感倾向或者还具有其它的情感倾向,但是该 Email 的情感基调是“高兴”的.

5 Email 的英文文本情感监测系统的设计与实现

5.1 系统设计

基于 Email 文本的情感监测器的框图如 2 所示.

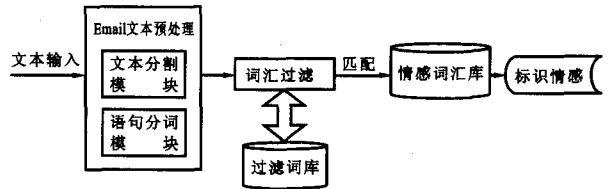


图 2 Email 中的情感监测系统

从图 2 可以看出,对于 Email 文本的情感监测一共分为几个模块:文本预处理模块、词汇过滤模块和文本中表达情感倾向的词汇的识别模块.

(1)文本预处理模块

文本预处理模块一共包含两个部分:文本分割模块和语句分词模块.文本分割模块的任务就是将大段 Email 文本分成一条条的语句,以标点符号“.”、“!”和“?”作为语句分割的结束符,而语句分词模块,则是将每条语句分成一个个单词,这个分割的标识符就是空格符号.

(2)词汇过滤模块

词汇过滤模块是将语句中的停用词过滤掉,即将无法表达情感的词汇过滤掉,比如人称代词、冠词、数词等等.过滤这些词汇也是要和停用词库中已经罗列出来的词汇相匹配,如果匹配成功,则被过

滤掉。

(3)关于情感词汇库的匹配模块

在情感词汇库的匹配模块中,将用最大长度匹配算法来实现这一模块.为了使计算机能够自动识别出表达情感的词汇.准确率是衡量系统性能的一个重要标准,由于情感库中的词表是无序结构,如果使用线性表查询,查询所花费的时间是,而使用 Hash 表进行查询的话,花费的时间为.因此,语句中的情感词汇与情感库中的表达情感的词汇的匹配将以 Hash 表的结构进行查询,这样一来,大大降低了查询时间.

所谓的最大长度匹配算法就是搜索来自于 Email 中组成语句的字符数最多的词汇,然后对这些词汇进行提取.比如,在某条语句中首先得到字符串参数为“abcdefg”,查询情感词汇库进行匹配,若不成功则去掉尾字符再查找“abcdef”,直到查到该单词,例如“abc”,然后去掉查到的字符取剩下的“defg”接着匹配,直到字符串为空结束.英文语句的基本单位是单词,在计算机中以单字节表示,而且英文字符串是以空格区分的单词为单位.而情感监测系统的情感标识模块将在系统实现部分进行研究.

感知.而且,颜色不但能够应用于大规模的分类,即使在很细微的分别上也能很明显的表示出来.在以往描述文本的情感中,多数情感结果用曲线来描述,而对于非常细微的表达,曲线就显得无能为力了.因此,在本文的基于 Email 文本的情感监测系统中,将采用六种不同的颜色代表六类基本情感,如表 2 所示.

表 2 用颜色表示情感

颜色	象征(含义)	表示的情感
橙色	热烈、喜悦、果敢和奋进	高兴(Happiness)
黑色	沉默、神秘、恐怖和纯粹	害怕(Fear)
灰色	和谐、浑厚、静止和悲伤	悲伤(Sadness)
蓝色	幽静、深远、冷郁和远离	厌恶(Disgust)
黄色	智慧、光辉、敏感和好奇	惊奇(surprise)
红色	喜庆、吉祥、绝望和愤怒	愤怒(Anger)
棕色	稳定和中立	中性情感

原型系统的开发主要使用了 SQLServer, JavaScript、C# 作为开发语言.图 3 所描述的程序运行结果就是对举例 1 的情感监测结果.

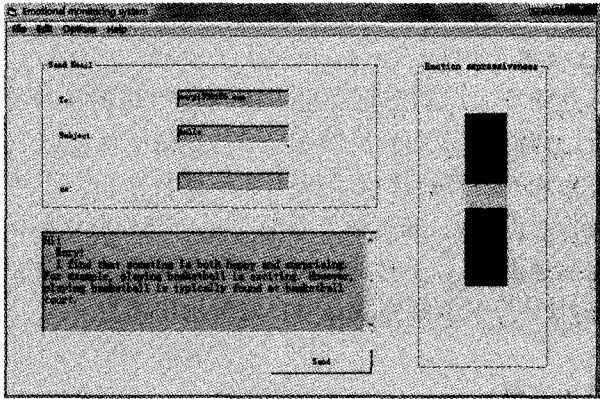


图 3 程序运行结果

5.2 系统实现

人机交互设计,主要是指用理论来指导设计,一方面防止出错,一方面用以提高工作效率.颜色心理学是致力于分析颜色在人类行为和感觉中的作用^[7].颜色对人的心理和生理的影响力却不容忽视,颜色不但能够影响人的视觉神经,进而还会影响心脏、内分泌机能和中枢神经系统的活动.广义上讲,颜色分为冷色和暖色两大类,颜色的冷暖从物理学的角度来讲,是指来自于人对光的体验.不用颜色的光的波长是不同的,紫光波长最短,红光波长最长.波长短即频率高,其能量不容易被物体吸收,所以让人觉得不容易通过高频光获取温暖.在心理学上,对于一种感觉兼有另一种感觉的心理现象,叫联觉现象.人们的颜色感觉容易引起联觉,因此,颜色容易对人的心理产生这样或那样的影响.

本节的系统实现部分将用颜色来标识 Email 中的文本所表达的情感倾向.之所以采用这样的方式,原因有二:一是颜色本身就带有唤起人类情感的自然特质,而且因为不同的文化对颜色有着不同的理解,因为这样的特征,对我们未来在研究文化感知方面奠定了坚实的基础;二是用颜色来标识 Email 的情感倾向,是因为颜色在虚拟环境中很容易被人眼

如图 3 所示,程序运行结果的左半部分和现有的邮件系统完全一致,而右半部分的彩色条形图即为 Email 中情感的颜色标识,即分别用橙色、黄色和棕色描述了 Email 中所出现的“高兴”、“惊奇”和“中性”情感,而且,从颜色标识图形上可以看出,橙色出现的次数最多,一共出现了三次(而其他颜色出现的次数均小于三次),而橙色表示“高兴”,由此可见,该邮件内容的情感基调为“高兴”.不难看出,该系统既增加了 Email 的情感可读性,又增加了趣味性,将具有很好的应用前景.

6 结束语

本文以 OMCS 中的常识知识作为 Email 的情感词汇库,研究了 Email 中英文语句的结构建模,并设计了基于英文文本的 Email 情感监测系统.通过

实例对情感监测系统的开发过程进行了说明。该系统的描述结果不再是简单的曲线,而是引入了颜色心理学的内容,用不同的颜色描述 Email 文本的情感倾向。因此,该系统虽然还处于研发阶段,但是其应用前景可见一斑。

参考文献:

- [1] 张红霞. 智能电子邮件检测技术的研究[J]. 微电子学与计算机, 2007(5): 143-145.
- [2] Dow K, Serenko A, Turel O, et al. Antecedents and consequences of user satisfaction with e-mail systems [J]. International Journal of e-Collaboration, 2006, 2 (2): 46-64.
- [3] 王培良, 吴小红, 许力. 一种人工情感及行为策略模型的有限自动机实现[J]. 微电子学与计算机, 2009, 26 (2): 36-39.
- [4] Thomas J, Ray D, Darren W. Reducing the effect of email interruptions on employees[J]. International Jour-

nal of Information Management, 2009(23): 55-65.

- [5] Dyer M G. A society of ideas on cognition: review of Marvin Minsky's The Society of Mind[J]. Artificial Intelligence, 1991, 3(48): 321-334.
- [6] Push Singh, Thomas Lin, Erik T Mueller, et al. Open mind common sense: knowledge acquisition from the general public[C]// Proceedings of AAAI Spring Symposium. Palo Alto, CA: AAAI, 2002.
- [7] Hugo Liu, Ted Selker, Henry Lieberman. Visualizing the affective structure of a text document[C]// Proceedings of the Conference on Human Factors in Computing Systems, Lauderdale, FL, USA: ACM, 2003: 740-741.

作者简介:

于冬梅 女, (1975-), 博士, 讲师. 研究方向为人工智能、模式识别和情感计算.

(上接第 160 页)

以让管理人员动态直观地了解当前网络的接收状况,而且可以及时地反馈客户端的控制信息,在不过多占用控制带宽前提下减小客户端发送 RTCP 控制信息的间隔,提高服务器端对客户端控制信息的响应速度,保证了合适的自适应视频码流的播放流畅,使 IPTV 直播系统具有良好的用户体验。

参考文献:

- [1] ITU-T FG IPTV. FG IPTV-ID-0042. Proposal on IPTV definition [S]. Geneva: Focus Group on IPTV, 2006.
- [2] Robins M. Delivering optimal quality of experience (QoE) for IPTV success [J]. Sprint Communications, 2006(1): 4-15.
- [3] Holbrook H, Cain B. Source-specific multicast for IP [S]. IETF draft-ietf-ssm-arch-07. txt. 2005..
- [4] 李燕灵, 马瑞芳, 左力. 基于 RTP/RTCP 的实时数据传

输模型与及实现[J]. 微电子学与计算机, 2005: 138-140.

- [5] Chesterfield J, Ott J, Schooler E M. RTCP Extensions for Single-Source Multicast Sessions with Unicast Feedback [S], Internet draft, IETF draft-ietf-avt-rtcpssm-13. txt, 2007.
- [6] Chesterfield J, Scholar E M. An extensible RTCP control framework for large multimedia distributions [S], 2003.
- [7] 方新勇, 马瑞芳, 刘萍芬, 等. 基于 RTP 的流媒体传输系统的设计与实现[J]. 微电子学与计算机, 2007, 24 (11): 183-185.

作者简介:

王玲玲 女, (1984-), 硕士研究生. 研究方向为计算机网络、多媒体网络.