摘要

在互联网飞速发展、电子商务技术日趋成熟的今天，随着移动终端——手机的日益普及，手机短信作为无线数据通讯的一个基本业务，为人们相互间交流提供了新的手段，并且受到了手机用户的青睐。作为一种新型的广告宣传的方式，手机短信有很大的潜力和前景，然而在实际运营中仍有不可回避的问题：垃圾短信。从目前的市场情况来看，手机短信广告最需要面对的就是如何解决扰民问题。这就需要广告发布者采取有效的方法以获得广告受众的相关信息，从而定向投放具有针对性、应变性的短信广告。

为实现这一目的，我们可以运用文本挖掘技术，从海量的电信业务数据进行分析，在此基础上有针对性地投放短信广告，从而形成广告投放者、广告受众、电信运营商三赢的局面。本文将文本挖掘方法应用于电信业务数据模型上，应用数据约简技术以解决数据量大的问题，应用数据除噪技术以解决数据含噪声的问题，综合运用多种模式识别技术以从电信业务数据中分析出用户的兴趣点所在、实现短信广告定向投放功能，这些都是本研究的重点，也是本研究的创新之处

本文的主要工作包括：（1）提出了一种针对短信文本的聚类方法，具有高准确度和高效率的特点。（2）提出了一种预加窗的中文文本校对技术，用于文本规范和校对，同时该算法具有较小的计算复杂度。（3）提出了一种自适应的重复特征选择技术，该方法能够最终得到最优的低维特征空间，同时也有效的解决了训练集含有噪声训练元素情况下的最优特征提取问题。（4）提出了一种基于关键词表的特征权重调整技术，进一步地突出了短信中的关键词成分，提高了关键点的识别率。

试验表明，本文中提出的方法是有效的，此外本研究还有一定的现实意义，对数据挖掘在电信行业中的应用有一定的促进作用。

关键词：短信文本，文本挖掘，预处理，特征提取

Abstract

In these days, Internet develops fast and e-commerce technology has become more and more mature. And with the increasing popularity of mobile phones – Mobile Terminal, SMS (Short Messaging Service) provides new tools for people's mutual exchange as a basic wireless data communications business, and it is accepted by the mobile phone users of all ages. As a new type of advertising, mobile phone messages have a great potential and prospects. However, in actual operations it remains an unavoidable problem: garbage messages. Judging from the current market situation, SMS advertising needs to be addressed is how to solve the most disturbing problem. This requires that the Advertising publishers take effective way to obtain relevant information from the audience, so that they can put the SMS response advertising targetedly.

To achieve this purpose, we can use text mining technology to analysis Massive Data of the telecommunications business. On this basis we can put the SMS response advertising targetedly, thereby it can create a situation of Advertising, advertising audiences, telecommunications operators-win. In this paper, text mining method has been applied to model data telecommunications business, we apply the data reduction techniques to solve the problem of large volume data and apply the data Eliminating-noise technology to solve the noise problem. Meanwhile, we use a combination of pattern recognition techniques to analyze data from the telecommunications business. So that we can gain what is the customer's interest and realize the function of putting advertising messages directedly. These are the focus of this study, as well as the innovation of the study.

The main work includes: (1) presents a text messaging cluster, with high accuracy and high efficiency characteristics. (2) Presents a window of the Chinese version of Pre-checking technology, and standards for proofreading text. Meanwhile the computational complexity of the algorithm is smaller. (3) Presents an adaptive feature selection technology, which could ultimately achieve the best low-dimensional feature space and is also an effective solution to the noisy training elements of the training set the optimal feature extraction. (4) Presents a technology based on the Keywords list to

adjust feature weights which further highlights the Keywords ingredients of the message and improves the recognition rate.

The experiments show that this paper presents the method is effective. In addition, the study has the practical significance still. Lastly, the study is a certain role on the application of data mining in the telecommunications industry.

Key words: text messaging, text mining, preprocess, teature extraction

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的资料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

签名：　　日期：　2007 年 03 月 28 日

关于论文使用授权的说明

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行查阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

签名：　　导师签名：

日期：2007 年 03 月 28

第一章 绪论

1.1 研究背景及意义

随着市场经济的蓬勃发展和以市场需求为导向的现代营销战略的推动，广告业获得了极大的发展，越来越多的企业开始重视并充分发挥广告营销的作用，宣传产品品牌、树立企业形象。与此同时，广告的发布方式也是日新月异。由于手机用户数量庞大，短信发送便捷、成本低廉且有强制阅读的优势，因此手机短信作为一种广告宣传的方式具有很大的潜力和前景。

手机短信广告具有不可逆的单向传播性质，并有着绝对低成本和无障碍直达两大优势。然而，尽管手机短信广告有着先天优势，在实际运营中仍有不可回避的问题：滥发的垃圾短信充斥着手机，造成了用户的普遍反感。从目前的市场情况来看，手机短信广告最需要面对的就是如何解决扰民问题。这就需要广告发布者采取有效的方法以获得广告受众的相关信息，从而定向投放具有针对性、应变性的短信广告[1]。

由于短信文本是一种半结构化的数据，并且，短信文本中主要包括的是文本信息，因此本文将文本挖掘的相关技术和方法引入到短信文本的处理领域，实现对短信的分类，从而实现具有针对性、应变性地投放短信广告的目的。此外，本文还针对短信文本本身的特点提出了一些改进方法，进一步提高了试验效果。

1.2 垃圾短信的危害及当前的状况

1.2.1 垃圾短信的种类

(1) 垃圾短信的定义

垃圾短信是指批量发送的内容违法或者违规的短信，或者违背收集用户主观意志接收到的并且客观上对用户造成骚扰的短信。垃圾短信有以下 4 个明显的特点：1，批量发送；2，内容违法、违规或涉及广告宣传；3，违背用户主观意志；4，客观上造成对用户骚扰或其它权益的侵害[2]。

我国一年的短信息总量约为 3000 多亿条[3]，其中有不少是垃圾短信，这不仅

占用了有限的网络资源，造成网络拥塞，使运营商耗费更多的资源对其进行处理、过滤，手机用户也要花费大量的时间来处理这些短信，同时那些以欺诈为目的的短信可能使很多分辨力差的手机用户损失大量的金钱。垃圾短信的经济成本无疑是一个惊人的数字，已成为一种社会公害。

(2) 垃圾短信的分类

从短信的发送者来区分，垃圾短信可以被分为四种形式[4]：

1. 违法短信

违法短信主要是指由不法分子发出的短信，不法之徒利用手机短信作案主要有如下三种形式：

(1). 通过发送手机短信进行诈骗，骗取、偷盗他人钱财。

(2). 利用手机短信从事制作假证件、假公章、假学历、考前出卖试题等违法经营活动。

(3). 通过手机散布破坏民族团结、影响社会稳定的政治谣言或有害信息，如果放任有害信息传播，有可能严重影响到国家安全。

2. 短信陷阱

不良服务提供商制造“短信陷阱”一般通过两个渠道：一个是通过电信运营商的网络平台向消费者发送“诱惑短信”，用户一旦回复短信就被当作确认定购服务；另一个是在互联网上设置“陷阱”，“诱惑”用户发回短信确认。无论服务提供商通过何种途径，都必须获得消费者的回复确认才可产生费用。为什么不良服务提供商要千方百计得到用户的回复呢？原因在于以往运营商在三方之中充当代收费的角色，服务提供商需要得到用户的回复确认与运营商进行费用结算，于是一些不良服务提供商千方百计设置“短信陷阱”让用户不自觉“回复”。

3. 不良短信

这类短信息一般不构成直接的利益侵犯，但却给接收者带来了身心的烦躁甚至伤害。此类短信多以整人为主要目的，加之内容低俗，格调低下，招致了不少用户的反感，被称为“精神污染”。当前有不少人靠编写不良短信谋生，更有新浪、网易这样的门户网站以不良短信牟利，更不用说许多利欲熏心的大小服务提供商们了。

4. 广告短信

此类短信多为各短信广告公司，主要具有以下四个特点：

1. 覆盖面广，据业内人士称：只要事先选择好手机或小灵通的一个号段，把开始的号码和最后一个号码输入软件，并输入相关软件群发即可。

2. 费用低，目前，市场上通行的短信广告的服务价格为三至四分钱一条，这意味着，即便是发送给一百万人，也只需花费三到四万元甚至更低。

3. “广告效果”好，由于用户在阅读该种短信前无法知晓该短信的内容，即使删除也是在阅读了该短信之后，类似被强迫洗脑。

4. 具有较强的隐蔽性，相对于传统媒体广告，利用短信发送的广告具有较强的私密性，一般人在收到广告消息之后多为阅读后即行删除，很少和周围的人分享。

1.2.2 垃圾短信的危害

垃圾短信的产生和存在很大程度上是由于商业的原因。不可否认，使用得当，如使用用户订阅的方式，短信是相当经济有效的广告方式，是开拓迅速增长的直销市场的有力工具。遗憾的是，很多商家并没有遵守游戏规则，采用了狂轰滥炸的方式，最终导致全民对垃圾短信行为的批评和抵制。

在讨论这个问题之前，让我们看一下为什么很多人都采用短信这种方式。总结一下，大致归于两点原因：低成本和易于匿名。据调查，发送短信的成本几乎为零，只要投资几百元就可以获得专门的短信群发器和所在地的手机号码，每小时可以发送上万条，而且手机短信具有强制阅读性，这样短信内容就可以保证被注意到。由于低廉的成本，即使只有很少很少的部分得到反馈，就足以支付这些费用了，比起昂贵的其他方式的广告自然很划算了。此外，由于短信是由群发器发出的，所以发送者具有高度的隐蔽性，很不容易被追踪到。

另外，从整个通讯资源来看，目前通讯资源还是比较有限。垃圾短信里的信息几乎没有什么价值，每次发送成千上万份这样的短信，会占用大量的通讯资源，严重时甚至会造成拥塞，中断信息的通讯。这些都是运营商和用户所不愿意看到的。据专家统计，消除垃圾短信可以为运营商和手机用户每年节省相当的成本。

其次，从手机用户来看，垃圾短信浪费了人们的大量时间。一般人们需要至少 10 秒钟时间来判断是否为垃圾短信，如果每天收到几十份垃圾短信，就的花大约 10 分钟的时间来处理它们，实在是比较痛苦的事情。垃圾短信也威胁无线网络的安全，特别是那些个人用短信群发器发送欺骗短信的情况。大量的网络资源被占用，严重时正常运作被迫终止。

垃圾短信不仅带来了技术方面和经济方面的问题，同时也带来了一系列社会问题。一些不法分子利用短信传播一些色情、反动、暴力、迷信等不良信息和带有欺诈性质的内容。还有一些宗教政治团体的挥之不去的垃圾短信更是引起了人们的愤慨。

最后，正如媒体上报道的，垃圾短信也严重地损害了移动、联通等电信运营商的形象，影响正常业务。

1.2.3 我国垃圾短信的当前情况

(1)当前我国垃圾短信泛滥严重

目前我国垃圾短信泛滥，情况极为严重。通过专门的问卷调查，我们发现用户每周收到短信的数量集中在 5 条以内的占多数，约 42.7%；每周收到 5-10 条垃圾短信的用户占 34.95%；受到垃圾短信达 10-20 条的用户占 14.19%；另外 6.25%的用户每周收到多达 40 条以上的垃圾短信。根据数据分析我国的手机用户平均每周收到 8.29 条垃圾短信。[5]

我国目前拥有手机用户超过 4.43 亿，他们是垃圾短信的直接受害者。根据上面的数据计算，每个手机用户每周至少需要在垃圾短信上花费 1.38 分钟。这就意味着，全国的手机用户每年会浪费掉 5.32 亿小时的宝贵时间。[6]

(2)垃圾短信的特点分析

通过对我国当前的垃圾短信的分析可知，我国垃圾短信的特点包括一下几个方面：

1) 从内容上看，国内的垃圾短信主要是来自国内的产品和服务的推广内容，相当一部分公司和个人利用短信这种形式推广新产品以及特别的服务等。

2) 从来源上看，绝大部分垃圾短信都是来自国内，国内的大部分垃圾邮件来自于推销为目的的公司。

3) 从发展趋势上看，国内垃圾短信问题形式不容乐观。通过短信来传达，正在被越来越多的公司选中，在相关政策出台之前，相信垃圾短信会更加猖獗。

4) 手机病毒正在逐渐蔓延，如果由于病毒而引发垃圾短信，无论从数量上还是危害上，都需要引起足够的重视。

1.3 本论文的主要研究内容及论文的组织

1.3.1 本论文的主要研究内容

本文主要是针对垃圾广告短信，在基于数据挖掘技术进行的短信文本分类研究，论文根据短信可转化为文本这一特性，通过对短信文本相关特性和相关技术的研究了解，提出了将文本分类算法运用到短信处理技术之中。本论文的研究工作主要包括一下几个方面：

1) 本文重点研究了在短信预处理方面将结构化、半结构化的短信转化为结构化的文本数据方法，特别是在整合短信文本时的新方法。

2) 在对短信文本进行预处理的基础之上，本文提出了一种预加窗的中文文本校对技术，用于文本规范和校对，同时该算法具有较小的计算复杂度。同时还提出了一种自适应的重复特征选择技术，该方法能够最终得到最优的低维特征空间，同时也有效的解决了训练集含有噪声训练元素情况下的最优特征提取问题。本文还提出了一种基于关键词表的特征权重调整技术，进一步地突出了短信中的关键词成分，提高了关键点的识别率。

3) 最后，构建了一个主要基于内容的短信分类系统测试模型。

1.3.2 论文组织

本论文共分五章，具体安排如下：

第一章是全文绪论。该章介绍了垃圾短信产生的原因、危害以及当前的状况，进而对短信文本分类进行了概要的阐述，包括常用的技术和方法。最后给出了论文工作的主要贡献。

第二章讨论了文本挖掘技术。本章首先对文本挖掘技术的定义、过程和挖掘方法作了叙述，然后分析了短信文本的特征，并简要讨论了文本挖掘技术在短信文本中的应用。

第三章对短信文本的预处理方法和技术进行了详细的论述。首先对文本的表示给出了相关定义，然后对数据预处理的几个主要步骤进行了阐述，其包括文本特征格式分析、中文分词处理、对错字和同音异形词的校对、去噪预处理、短信文本特征选择研究、基于兴趣关键特征词的权重调整技术、特征规范化处理以及一些后期处理工作。

第四章讨论了短信的分类技术。本章首先对文本分类的过程和分类方法进行

叙述，然后对短信文本的分类方法进行了研究。文中提出了分类之前先对短信进行预处理整合的思想，大大地提高了分类速度。

第五章主要介绍了短信分类系统 SVMCLS 的研究与设计。本章结合前几章研究的方法和思想，从实际应用的角度出发，设计了一个短信分类系统模型，并给出了对一些关键问题的处理方法和关键算法。

第二章 文本挖掘技术

文本信息的挖掘就是在对大量训练样本处理的基础上，得到文本数据间的内在特征，并以此为依据对信息资源中进行有目的的信息提取。本文在对短信文本的格式进行探讨，并对其半结构化的文本格式进行预处理的前提下，将文本挖掘的主要技术和方法应用到短信的处理中。本章首先对文本挖掘技术的定义、过程和挖掘方法作了叙述，然后分析了短信文本的特证，并简要讨论了文本挖掘技术在短信中的应用，其关键技术问题在后续章节中详细论述

2.1 文本挖掘

2.1.1 文本挖掘的定义

文本挖掘的定义 文本挖掘可以定义为提取散布于文本中新的、合理的、对于未来行为有指导意义的知识的过程，通过组织和运用这些知识可以为未来提供有价值的参考信息

文本挖掘不同于数据挖掘，数据挖掘面对的是结构化数据，采用的方法大多是非常明确的定量方法。其过程包括数据取样、特征提取、模型选择、问题归纳和知识的发现。而文本挖掘由于它处理的是非结构化的文本，因此，决定它采用的方法与数据挖掘不同。它经常使用的方法来自与自然语言理解和文本处理领域，如文本摘要、文本分类、文本检索等技术，发现的知识往往不是精确的数据，而是定性的规则。对于中文文本的文本挖掘其难度较大，体现为汉语分词问题，建立完整的汉语概念体系的困难和汉语语法、语义和语用分析的困难。[7]

文本挖掘可以对大量文档集合的内容进行总结[8][9]、分类、聚类、关联分析，以及利用文档进行趋势预测等。

当前，文本分类已成为一个日益重要的研究领域。随着文本信息的快速增长，文本分类显得越来越重要。由于分类可以在较大程度解决文本信杂乱的现象，方便用户准确地定位所需的信息和分流信息。因此，文本自动分类已成为一项具有较大实用价值的关键技术，是组织和管理数据的有力手段，可被用于抽取符号知识、新闻分发、排序电子邮件以及学习用户兴趣。

文本分类是指按照预先定义的主题类别，为文档集合中的每个文档确定一个类别。这样，用户不但能够方便地浏览文档，而且可以通过限制搜索范围来使文档的查找更为容易。利月文本分类技术可以对大量文档进行快速、有效地自动分类。目前，文本分类的算法有很多种，包括神经网络、遗传算法、粗糙集在内的技术都被用来进行分类，不过，比较常用的还是 TF-IDF 方法和 NaiveBaves 等基于统计学的方法。[10]

文本聚类与分类的不同之处在于，聚类没有预先定义好的主题类别，它的目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能地大，而不同簇间的相似度尽可能地小。Hearst 等人的研究已经证明了“聚类假设”，即与用户查询相关的文档通常会聚类得比较靠近，而远离与用户查询不相关的文档。[11]

关联分析是指从文档集合中找出不同词语之间的关系。Brin 提出了一种从大量文档中发现一对词语出现模式的算法，并用来在 Web 上寻找作者和书名的出现模式，从而发现了数千本在 Amazon 网站上找不到的新书籍。Wang 等人以 Web 上的电影介绍作为测试文档，通过使用 OED 模型从这些半结构化的页面中抽取词语项，进而得到一些关于电影名称、导演、演员、编剧的出现模式。[12]

分布分析与趋势预测是指通过对文档的分析，得到特定数据在某个历史时刻的情况或将来的取值趋势。Feldman 等人使用多种分布模型对路透社的两万多篇新闻进行了挖掘，得到主题、国家、组织、人、股票交易之间的相对分布，揭示了一些有趣的趋势。Wdthrich 等人通过分析 Web 上出版的权威性经济文章，对每天的股票市场指数进行预测，取得了良好的效果。[13]

2.1.2 文本挖掘的过程

文本数据挖掘的一般过程可以用图 2.1 来概括描述[14]。首先对数据挖掘的文本进行分词处理，把文本切成词条。接着建立挖掘对象的特征表示，例如：在 Internet 上的文本数据挖掘对象通常是一组 html 格式的文档集，这样的文本挖掘对象缺乏象关系数据库中数据的组织规整性，因此要将这些文档转换成一种类似关系数据库中一记录的较规整且能反映文档内容特征的表示，一般采用特征向量。但在目前所采用的文档表示方法中，存在一个共同的不合人意的地方是文档特征向量具有惊人的维数，因而特征向量的约简处理成为文本挖掘处理过程中必不可少的一个环节。在完成特征向量维数的缩减后，便可以利用机器学习的方法提取面向特定应用目的的知识模式。最后对获取的知识模型进行质量评价，若评价的

结果满足一定的要求，则存储该知识模式，否则返回到以前的某个环节分析改进后进行新一轮的挖掘工作。

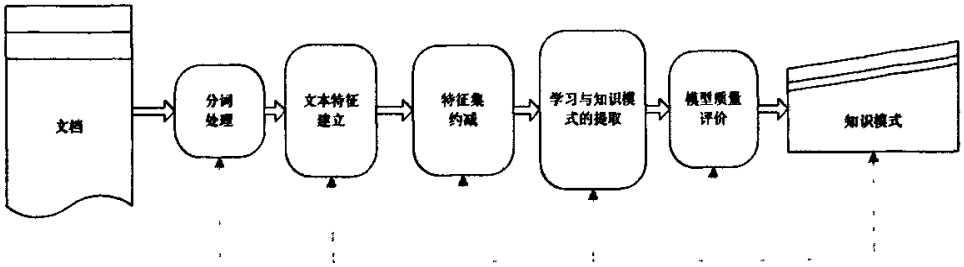


图 2.1 文本挖掘的一般过程

2.1.3 文本挖掘方法

在文本挖掘中，文本的特征表示是挖掘工作的基础，而文本分类和聚类是两种最重要、最基本的信息发现功能。由于在本文中要论述分类方法，所以此处只简单介绍分类。关于分类的详细介绍，我们在后面的章节详细研究。

(1) 文本的特征表示[15][16]

与数据库中的结构化数据相比，文档具有有限的结构，或者根本就没有结构。不同类型文档的结构也不一致。此外，文档的内容是人类所使用的自然语言，计算机很难处理其语义。文本信息源的这些特殊性使得现有的数据挖掘技术无法直接应用于其上。我们需要对文本进行预处理，抽取代表其特征的元数据。这些特征可以用结构化的形式保存，作为文档的中间表示形式。

文本特征指的是关于文本的元数据，分为描述性特征，例如文本的名称、日期、大小、类型等；以及语义性特征，例如文本的作者、机构、标题、内容等。描述性特征易于获得，而语义性特征则较难得到。W3C 近来制定的 XML，RDF 等规范提供了对 Web 文档资源进行描述的语言和框架。在此基础上，我们可以从半结构化的 Web 文档中抽取作者、机构等特征。

对于内容这个难以表示的特征，我们首先要找到一种能够被计算机所处理的表示方法。矢量空间模型(Vector Space Model VSM)是近几年来应用较多且效果较好的方法之一。在该模型中，空间文档被看作由一组正交词条所张成的矢量空间，每个文档 d 表示其中的一个规范化矢量 $V(d)=(t_1,w_1(d);\cdots;t_i,w_i(d);\cdots;t_n,w_n(d))$，其中 t_i 为词条项，$w_i(d)$ 为 t_i 在 d 中的权值。可以将 d 中出现的所有单词作为 t_i，也可以要求词 t_i 是 d 中出现的所有短语，从而提高内容特征表示的准确性，$w_i(d)$ 一般定义为 t_i 在 d 中出现频率 $tf_i(d)$ 的函

数，即$w_i(d)=\Psi(tf_i(d))$。常用的Ψ有：布尔函数$\psi=\begin{cases}1.....tf(d)\geq 1_i\\0.....tf(d)=0\end{cases}$；平方根函数$\psi=\sqrt{tf_i(d)}$；对数函数$\psi=\log(tf_i(d)+1)$；TFIDF函数$\psi=tf_i(d)\times\log(\frac{N}{n_i})$，其中$N$为所有文档的数目，$n_i$为含有词条$t_i$文档数目。

(2) 文本分类[17][18][19]

文本分类是一种典型的有监督的机器学习问题，一般分为训练和分类两个阶段，具体过程如下：

训练阶段：

(1) 定义类别集合$C=\{c_1,\cdots,c_i,\cdots,c_m\}$，这些类别可以是层次的，也可以是并列式的；

(2) 给出训练文档集合$S=\{s_1,\cdots,s_i,\cdots,s_n\}$，每个训练文档$s_j$被标上所属的类别标识$c_i$；

(3) 统计S中所有文档的特征矢量$V(s_j)$，确定代表C中每个类别的特征矢量$V(c_i)$；

分类阶段：

(1) 对于测试文档集合$T=\{d_1,\cdots,d_k,\cdots d_r\}$中的每个待分类文档$d_k$，计算其特征矢量$V(d_k)$与每个$V(c_i)$之间的相似度$sim(d_k,c_i)$；

(2) 取相似度最大的类别$\arg\max sim(c_i,d_k)$作为d_k的类别。

有时也可以为d_k指定多个类别$c_i\in C$，只要d_k与这些类别之间的相似度超过某个预定的阈值。如果d_k与所有类别的相似度均低于阈值，那么通常将该文档放在一边，由用户来做最终决定。对于类别与预定义类别不匹配的文档而言，这是合理的，也是必需的。如果这种情况经常发生，则说明需要修改预定义类别，然后重新进行上述训练与分类过程。

在计算$sim(d_k,c_i)$时，有多种方法可以选择。最简单的方法是仅考虑两个特征矢量中包含的词条的重叠程度，即：

$$sim(d_k,c_i)=\frac{n\cap(d_k,c_i)}{n\cup(d_k,c_i)}$$

其中$n\cap(d_k,c_i)$是$V(d_k)$和$V(c_i)$具有的相同词条数目，$n\cup(d_k,c_i)$是$V(d_k)$和$V(c_i)$具有的所有词条数目；最常用的方法是考虑两个特征矢量之间的夹角余弦，即

$$sim(d_k,c_i)=\frac{V(d_k)\bullet V(c_i)}{|V(d_k)|\times|V(c_i)|}$$

2.2 文本挖掘技术在垃圾短信方面的应用

2.2.1 短信文本的格式

短信文本通常分为普通文本和多媒体格式文本，后者比前者主要附加了彩铃和彩信功能。由于绝大部分垃圾短信都属于前者，所以这里我们主要讨论普通短信文本的格式。

普通短信文本的结构是相当简单的了，它含有一系列文本，每一部分有一个回车(CR)、换行(LF)以及内容组成。短信由收信人、短信内容、发信人和发送时间四部分组成，其中收件人、发件人和发送时间是必需的，而短信内容是可选的。下面看一个简单的例子：

From：13811111111

Content： 出租旺铺，门前客流大，适合开奶茶店 也可居住用 有院一楼 南向正房四全

To：13822222222

发送时间：11：02：26

2007.03.18

由上面的介绍可知，短信文本有一定的结构，而短信文本的内容部分大多是无结构的文本。文本主要研究内容之一就是如何将短信文本这一半结构化文本特征转化为结构化的数据形式，在转化为数据形式后，引入文本挖掘的一些技术方法对其进行处理，以便对其进行正确分类。

2.2.2 短信文本的分类过程

本文研究的主要目的就是把短信文本按内容的不同进行分离，即对短信文本进行分类。基于短信文本的半结构化文本特征，本文采用文本挖掘技术中的分类方法对短信文本进行处理。本文设计并提出的对短信文本的挖掘过程主要包括下面的步骤：首先对待挖掘的电子邮件文本进行分词处理，把文本切成词条；接着建立短信文本的特征表示，由于这样的文本挖掘对象缺乏像关系数据库中数据的组织规整性，因此需将这些文档转换成一种类似关系数据库中记录的较规整且能反映文档内容特征的表示，文本中采用的是特征向量；因为特征向量具有惊人的维数，因此接下来对特征向量进行约简处理：在完成特征向量维数的缩减后，利用机器学习的方法提取面向特定应用目的的知识模式；最后对获取的知识模型进

行质量评价，若评价的结果满足一定的要求，则存储该知识模式，否则返回到以前的某个环节分析改进后，进行新一轮的挖掘工作。关于短信的分类过程中的一些处理的详细介绍，将在本论文的其它章节介绍。

2.3 本章小结

本章简要介绍了文本挖掘，并着重阐述了文本挖掘的概念、处理过程及其主要处理方法。同时对短信文本的格式进行了阐述，提出将文本挖掘技术引入到短信文本处理中并给出其简单的实现过程，重点部分将在后续章节中阐述。

第三章 预处理技术研究

如前所述，短信文本不同于传统数据库中的结构化数据，有一定的结构，而其内容就没有结构。若想对短信文本这种半结构的数据施加信息处理技术，必须对其进行预处理，将以文本为主的短信文本表示为易于被计算机所处理的中间形式。

本章就短信广告定向投放技术中的短信文本预处理方面工作进行了详细的论述。文中首先对文本格式分析、中文分词处理、词性标注及无用词过滤等数据预处理方法进行综述。然后详细介绍了对错字和同音异形词的校对技术，并给出了结果与实验分析；接着介绍了重复型特征提取技术原理和算法，同样给出了相关实验结果分析；本章还提出了基于兴趣关键特征词表的特征权重提取技术，最后还对特征向量进行了规范化处理。本节的实例测试验证了本文提出的方法的有效性。

3.1 文本的相关定义

文档的表示是文本信息处理的最基本的前期工作[20]。目前这方面的研究工作已经取得了一定的进展。60 年代末由 Gerard Salton 等人提出的向量空间模型（Vector Space Model，VSM）[21][22]，因其简单及有效性，是近几年来应用较多且效果较好的方法之一。正是基于此，本研究中短信文本的表示选用了 VSM 模型表示。其基本定义有：

定义 3.1 文本

文本是短信分类系统处理的基本单位。泛指一个具有相对独立意义的自然语言片断(段落、句子组或句子)，一般指一篇文章。

定义 3.2 项

当文本的内容被简单地看成是它含有的基本语言单位(字、词、词组或短语等)所组成的集合时，这些基本的语言单位统称为项，也就是说文本 D 可以用项集来(Term List)来表示，即 $D(T_1,T_2,\cdots,T_n)$，其中 T_k 是项，$1 \le k \le n$。本文中，在不引起混淆的情况下，将使用“词”代替“项”这个术语。

定义 3.3 词的权重

对于含有 n 个词的文本 $D(T_1,T_2,\cdots,T_n)$，词 T_k 常常被赋予一定的权重 $W_k(1 \le k \le n)$，表示他们在文本中的重要程度，即 $D(T_i,W_i;T_2,W_2;\cdots;T_n,W_n)$。有时在特征词条确定时，常简记为 $D = D(W_1,W_2,\cdots,W_n)$。

定义 3.4 向量空间模型 (Vector Spaec Model，VSM)

给定一文本 $D = D(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$，由于 T_k 在文本中既可以重复出现又应该有先后次序的关系，分析起来仍有一定的难度。为了简化分析，可以暂不考虑 T_k 在文本中的先后顺序并要求 T_k 互异(即没有重复)。这时可以把 $T_1,T_2,\cdots,T_n$ 看成是一个 n 维的坐标系，而 $W_1,W_2,\cdots,W_n$ 为相应的坐标值，因而 $D(W_1,W_2,\cdots,W_n)$ 被看成是 n 维空间中的一个向量。我们称 $D(W_1,W_2,\cdots,W_3)$ 为文本 D 的向量表示。

定义 3.5 文本特征向量 (Feature Vector)

在 VSM 模型中，每一个文档都可以用一个向量来表示，向量的元素是由项(词条)及权重组成，该向量称之为文本的特征向量。特征向量是文档的一个特征表示，在某种意义上可以完全代表文档的特性。

在 VSM 中，每一篇文档都被映射成多维向量空间中的一个点，对于所有的文档类和未知文档，都可用此空间中的向量 $(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$ 来表示(其中 T_i 为词，W_i 为词对应的权值，用以刻画该词在描述此文档时的重要程度)，从而将文档信息的表示和匹配问题转化为向量空间中向量的表示和匹配问题来处理。

VSM 模型的不足之处在于它将文本表示成向量，作为向量空间的一个点，然后通过计算向量间的距离决定向量类别的归属时，一般不考虑向量中各个特征间的关系。这使得距离的计算不够准确，从而导致分类精度不够高。该问题会在后文中不断改善。

3.2 文本预处理

为了将短信文本内容表示成规范化的、易于为计算机所处理的数据，首先需要进行数据预处理，其主要步骤包括：文本特征格式分析及规范化、中文分词处理、中文文本校对、预处理去噪以及一些后期处理工作。

3.2.1 文本特征格式分析及规范化

分析文本格式可以有效地确定反映主体特征的文本区域，综合考虑词条权重，有助于特征词的高效抽取。[23]

对于普通文本而言，文本特征格式主要体现在文本的篇章段落构成形式上。

其中，文本标题是判断文本表达主题思想的一个特别值得重视的区域，标题很好地概括和总结了整篇文本内容；文本起始区域与终止区域也是与文本主题密切相关的，文本起始部分往往开宗明义，提出所要讨论的主题，而文本结束部分是对文本主题的再次强调。BaxendaleP.E.进行过统计，反映主题的句子 85%出现在首段，7%出现在尾段。[24]

于普通文本不同，短信文本是一种半结构化的数据。短信包含了 From、To、Subject 和 Date。一般对文档的分类比较有帮助的主要是 form 域，date 域和 body 域，我们可以把注意力主要用于这三部分。

如果忽视了这些信息，将所有的内容都同等对待，那么在向量空间模型中，那些决策特征的决策作用将会被众多非决策特征的影响所淹没掉。所以，本研究在设计中充分地利用了短信文本的格式信息，较好地提高了文本挖掘的性能。

文本规范过程主要清除所有的干扰字符和替换一些变异词。这里的干扰字符包括诸如标点、特殊字符等，比如*&^*，因为这些字符对于文本特征提取来说没有什么实际意义，而且还会造成对分类的干扰。所以在分词前，对短信文本进行检查，去除奇异符号。例如："我－－－－好想;）回家!!!"，将所有的非法字符去掉，得到"我好想回家"。

同时对于短信中的变异词，如 FQ(夫妻)，gf（女朋友）等，我们采用变异词表的方式进行处理。在每次分词处理后我们通过把分词过程所得词与变异词表中各词进行比较，如果该词在变异词表中，则用变异词表中相应正规词替换变异词。变异词表如下：

变异词	正规词
gf	女朋友
大虾	大侠
gg	哥哥
…….	……..

3.2.2 中文分词处理

众所周知，汉语的书写以汉字为基本单位，词与词之间没有明显的界限，要进行汉语的计算机处理，必须进行分词处理。分词是中文信息处理的核心和汉语自然语言理解的基础，很大程度上影响了中文处理系统的性能。目前主要有基于词典的分词算法和无词典的分词方法两种。

3.2.2.1 基于词典分词方法

基于词典方法的基本思想是：首先建立一个包含所有可能出现词的词库，然后对给定的待分词的汉字串 S，按照某种确定的算法切取 S 的子串。若该子串与词库中的某词条相匹配，则该子串是词，继续分割剩余的部分，直到剩余部分为空；否则，该子串不是词，转上重新切取 S 的子串进行匹配[25][26]。其主要包括以下方法：

1) 最大匹配法（Maximum Matching method, MM）：选取包含 6-8 个汉字的符号串作为最大符号串，把最大符号串与词典中的单词条目相匹配，如果不能匹配，就削掉一个汉字继续匹配，直到在词典中找到相应的单词为止。匹配的方向是从右向左。

2) 逆向最大匹配法（Reverse Maximum method, RMM）：匹配方向与 MM 法相反，是从左向右。实验表明：对于汉语来说，逆向最大匹配法比最大匹配法更有效。

3) 双向匹配法（Bi-direction Matching method, BM）：比较 MM 法与 RMM 法的分词结果，从而决定正确的分词。

4) 最佳匹配法（Optimum Matching method, OM）：将词典中的单词按它们在文本中的出现频度的大小排列，高频度的单词排在前，频度低的单词排在后，从而提高匹配的速度。

5) 联想-回溯法（Association-Backtracking method, AB）：采用联想和回溯的机制来进行匹配。使用该方法，词典的涵盖程度决定了词汇切分的准确率，要做到这一点很不容易。此外该方法无法正确切分出词表中未收录的新词，不具备自适应性。实际使用的分词系统，都是把基于词典的分词方法作为一种最初的切分手段，还需要通过利用各种其它语言信息来进一步提高切分的准确率。

3.2.2.2 基于统计的分词方法(无词典分词方法)

基于无词典的分词技术的基本思想是：基于词频的统计将原文中任意前后紧邻的两个字作为一个词进行出现频率的统计，出现的次数越高，成为一个词的可能性也就越大，在频率超过某个预先设定的阈值时，就将其作为一个词进行索引。这种方法能够有效地提取出未登录词。现有设计了一个基于无词典分词的算法，能比较准确地切分出文本中的新词。也有基于层次隐马模型，设计开发了“汉语词法分析系统”，将分词、词语排歧、未登录词的识别三个过程融合到一个相对统一的理论模型中[27][28]。

3.2.3 对错字和同音异型字校对

目前已经提出了多种校对方法与技术。这些方法按其采用的策略大体上可以分为两大类：一类是基于大规模语料库的统计数据，通过 N-Gram 模型进行校对。这种方法简便易行，有较高的召回率，但由于语料库覆盖的语言有限，所以它的查准率总是很低。另一类是通过构建字词混淆集，在校对时形成候选矩阵，运用动态规划或机器学习的方法从矩阵中找出一条最优路径作为校对候选字串，然后再与原句字串进行比较来实现校对。这类方法有较高的召回率和较好的查准率，但由于要对待校文本中的每一句进行矩阵空间搜索，算法复杂性太高，很难满足实用的需要。

在以上两类方法中基于 Markov 模型的文本校对方法有较好的效果，但是 Markov 模型只能处理语音的临近约束关系，不能用于处理较长的语言约束关系。为了弥补 Markov 模型这一缺点，本文提出了一种基于预加窗技术的校对方法。该方法利用词词连续性和词间字接续的方法来定位疑错窗口，然后再在各个疑错窗口内运用 Markov 模型进行定错和校对。由于此方法先用窗口技术初步定错，较大程度的减小了校对候选矩阵的维数，从而弥补了 Markov 模型只能处理语音的临近约束关系这一缺点。试验表明，我们提出的新方法是有效的。

3.2.3.1 疑错窗口方法原理

一般情况下，待校文本中的错误具有稀疏性的特点，所以无需用 Markov 模型对待校文本中的所有句子进行逐句构建候选校对矩阵和空间搜索最优，这样做多余的开销太大[29]。本文提出用预加窗技术来解决上述问题。在该技术中，窗口称为疑错窗口；疑错窗口由开始和结束位置确定；每个疑错窗口内包含了一或多个怀疑是错误的词，每个文本包含多个疑错窗口。本方法首先利用接续和散串方法来定位疑错窗口，把有可能出现错误的地方局限在疑错窗口内，然后最终的定错和纠错只要在窗口内进行就可以了。

假设 $W=\{w_1,w_2,\cdots,w_N\}$ 是词表中所有词的集合，$C=\{c_1,c_2,\cdots,c_M\}$ 是词表中所有汉字的集合；句子 $S=c_1c_2\cdots c_ic_{i+1}\cdots c_k=w_1w_2\cdots w_jw_{j+1}\cdots w_\phi$；$c_i$、$c_{i+1}$ 是相邻的汉字；w_j、w_{j+1} 是相邻的词；$i=1\cdots k, j=1\cdots\phi$；$w_j\in W$，$c_i\in C$。并设定一阈值 t_w，根据计算语言学中的二元语法模型理论[30]，可以利用字间互信息是否大于阈值 t_w 来判断词之间的前后连续关系。若大于 t_w 则认为上下文衔接合理，否则认为上下文衔接不合理。但是由于通常情况下两个相连词的相连字的连续性并不能表

征这两个词的连续性，同时由于语料库规模的限制，词词二元链接性计算存在一定误差，所以不失一般性，可用两个相连词的相连字的互信息和两个相连词的互信息的加权和作为词词连续性判据，针对待校文本是已经分词处理过的情况，其词间连续性判据可以表示如下：

$$MI(w_j,w_{j+1}) = \alpha I(w_j c_k, w_{j+1} c_1) + (1-\alpha) I(w_j, w_{j+1}) \tag{3.1}$$

其中：

$$I(w_j c_k, w_{j+1} c_1) = \log_2[P(w_j c_k, w_{j+1} c_1) / p(w_j c_k) p(w_{j+1} c_1)] \tag{3.2}$$

$$I(w_j, w_{j+1}) = \log_2[P(w_j, w_{j+1}) / p(w_j) p(w_{j+1})] \tag{3.3}$$

其中 $p(w_j c_k)$ 为 w_j 词的最后一个字 c_k 在语料库中的出现频率，$p(w_{j+1} c_1)$ 为词 w_{j+1} 的第一个汉字在语料库中出现的频率，$p(w_i c_k, w_{i+1} c_1)$ 为这两个汉字在语料库中连续出现的频率。$I(w_j c_k, w_{j+1} c_1)$ 为这两个汉字的互信息。$p(w_j)$ 为 w_j 词在预料库中的出现频率，$p(w_{j+1}.)$ 为词 w_{j+1} 在预料库中出现的频率，$p(w_i., w_{i+1})$ 为这两个词在预料库中连续出现的频率，$I(w_j., w_{j+1})$ 为这两个词的互信息。

词间连续函数定义为：

$$ZJ(w_i, w_{i+1}) = \begin{cases} 1 & \text{当} MI(w_j, w_{j+1}) \geq t_w \\ 0 & \text{其他} \end{cases} \tag{3.4}$$

但是，当两个字的互信息值 $I(c_i, c_{i+1}) \approx 0$ 时，此时 c_i、c_{i+1} 之间的连续关系不明确，根据统计理论[31]，Pearson 的 x^2-统计量可用于检测 c_i 和 c_{i+1} 的独立性。所以公式 3.4 应改为

$$ZJ(w_i, w_{i+1}) = \begin{cases} 1 & \text{当} MI(w_j, w_{j+1}) \geq t_w \ OR \ MX^2(w_j., w_{j+1}) \geq t_\theta \\ 0 & \text{其他} \end{cases} \tag{3.5}$$

这里 $MX^2(w_j, w_{j+1})$ 定义如下。

$$MX^2(w_j, w_{j+1}) = \alpha X^2(w_j c_k, w_{j+1} c_1) + (1-\alpha) X^2(w_j, w_{j+1}) \tag{3.6}$$

这里 $X^2(w_j c_k, w_{j+1} c_1)$ 为 w_j 词的最后一个字 c_k 与词 w_{j+1} 的第一个汉字的 $\chi 2$ 统计量；$X^2(w_j, w_{j+1})$ 为 $w_j c_k, w_{j+1} c_1$ 的 $\chi 2$ 统计量。

本文使用变长窗口模型作为疑错窗口模型，疑错窗口最小距离为 3 个词，各疑错窗口由开始位置和结束位置组成。例如给定一个语句 S 我们首先对该语句进行分词处理，得到词条集合 $w_1 w_2 \ldots w_j \ w_{j+1} \ldots \ w_\phi$，然后在此词条上定位疑错窗口。

定位过程从第一个词开始，此时标记第一个疑错窗口的开始位子为 1，然后计算随后的每相邻两个词的词间连续性。

这里可能有以下四种情况：

1. 如果在这搜索的 3 个词中没有词词连续性小于阈值的情况，同时第 3 个词和第 4 个词的词间连续性也不小于阈值，则把第 4 个词的位置记录为下一个疑错窗口的开始位置，并计算下一个疑错窗口内 3 个词的词间连续性。

2. 如果在这搜索的 3 个词中出现词间连续性小于阈值的情况，且出现词间连续性小于阈值的词在这 3 个词的中间位子，而第 3 个词与第 4 个词的连续性大于阈值，则定位该窗口的结束位置为第 3 个词的位置，并把该窗口的开始位置与结束位置记录到疑错窗口队列中。然后把第 4 个词的位置作为下一个疑错窗口的开始位置，并计算下一个疑错窗口内 3 个词的词间连续性。

3. 如果在这搜索的 3 个词中出现词间连续性小于阈值的情况，且出现词间连续性小于阈值的词在这 3 个词的末尾位子，即第 3 个词与第 4 个词的词间连续性小于阈值，则继续计算后续词的连续性，如果后续词的连续性仍然小于阈值，则继续计算，直到词间连续性大于阈值为止，并把该词间连续性大于阈值的两个词的前一个词的位置作为该疑错窗口的结束位置，加入到疑错窗口队列。然后把该疑错窗口的结束位子的下一个词作为下一个疑错窗口的开始位置继续搜索。

4. 如果在这搜索的 3 个词中出现词间连续性小于阈值的情况，且出现词间连续性小于阈值的词分别出现在这 3 个词间和 3 个词的末尾。则按第 3 种情况定位疑错窗口的结束位置。

疑错窗口的定位算法具体如下：

以下给出基本疑错窗口的数据结构及定位算法：

```
struct ErrorWindow//疑错窗口结构

{
interrorbegin;//疑错字串的开始位置
interrorend;//疑错字串的结束位置

}
ErrorWindoworientationerror(Sentence,LWSET,Ip)
//Sentence 为经校对预处理过的句子，LWSET 为低频

{int i=lp;
```

```
int n=length(Sentence,lp)//n 为 Sentence 句子，lp 位置后词的个数
ErrorWindow Ew;
Ew.errorbegin=lp;
Ew.errorend=0;
Bool error=false; Endlsee=false;
while(i<=n+lp)
{if ( w_i∈LWSET) //如果 wi 属于低频单字词
{ error=true;
  If(i=lp+3)
  Endless=true;
}
elseif (ZJ(i<n&&wi,wi+1)) //如果 wi 和 wi+1 的词间字接续小于阈值 tw
{error=true;
If(i=lp+3)
Endless=true;
}
else
{ If(i>lp+3)
{Ew.errorend=i;
Break;
}}
if(error==false&&i=lp+3&&Endless==false)
{ Ew.errorend=0;
Break;
}
elseif(error==ture&&i=lp+3&&Endless==false)
{ Ew.errorend=3;
Break;
}
elseif(error==ture&&i>=lp+3&&Endless==ture)
i=i+1;
```

```
i=i+1;
}
return Ew;
}
```

3.2.3.2 疑错窗口结合 Markov 模型的定错校对处理

本文将要校对的句子 $S = C_1C_2\cdots C_n$ 中的每个字 C_i 作为基字，以它的同音字集作为 C_i 的候选字集，从该候选集中依词频从高到低挑选出 m 个候选字，与基字 C_i 共同组成一列候选字向量 Z_i。因此，当句子 S 中的汉字个数为 n 时，句子 S 的字候选向量构成了 S 的字候选矩阵，$MATRIX(S) = Z_1Z_2\cdots Z_n$。寻找最佳路径需要借助于语言的统计特性本文，本文使用了计算语言学中著名的 Markov 统计语言模型。

1．马尔可夫模型

在 Markov 模型中，任何时刻的信源符号发生的概率只与前面已经发生的数个符号有关，而与更前面发生的符号无关[32]。当符号发生的概率只与前面 m 个符号有关时，则称 m 阶 Markov 模型。这 m 个信源符号所组成的符号序列看作是信源所处的状态，当信源符号的个数为 q 时，则模型中共有 qm 个不同的状态。

2．文本信息的统计特征

根据 Shannon 的信息理论，通常可把自然语言看成是一个离散的 Markov 模型。由于计算机的时间和空间的局限性，只能建立低阶 Markov 模型[33]。因此，这样的统计语言模型只能处理语言的近邻约束关系，对语言的远距离约束关系就无能为力了。若当前字 C_i 只与前 n-1 个字 C_{i-n}，C_{i-n+1}，…，C_{i-1} 相关时，则对于输出字串 $\langle C_1,C_2,\cdots,C_k\rangle$ 有：

$$p(C_1,C_2,...,C_k) = \prod_{i=1}^{k} p(C_i \mid C_{i-n},C_{i-n+1},....,C_{i-1}) \qquad 3.7$$

当应用二阶 Markov 模型时，即当前词的出现仅与它的前一个词有关，而与其它历史词无关时，则有：

$$p(C_1,C_2,...,C_k) = \prod_{i=1}^{k} p(C_i \mid C_{i-1}) \qquad 3.8$$

我们的任务是从语言结构元素格子图中寻找一条最佳路径(句子的最佳词序列 Wordilist*)，这条路径上节点中的词所构成的词序列，就是我们在候选矩阵中寻找的对应于要校对语句 S 的最佳语句。格子图中的路径搜索采用 Viterbi 搜索算法，

搜索中使用的路径评价函数为：

$$f(wordlist) = \lambda p(C_1, C_2, ..., C_i, ..., C_n) \quad 3.9$$

elei 是元素格子图中的元素节点，$elei = \langle C_{i,1}, C_{i,2}, \cdots, C_{i,N_i} \rangle$，其中 N_i 为 *elei* 中包含的词个数，对评价函数应用 Markov 模型时有：

$$f(Wordlist) = \prod_{i=1}^{n} \lambda_i p(C_i \mid C_{i-1}) \quad 3.10$$

公式中的 λ_i 为规则的权值调整系数，本研究中 $\lambda_i = 1 (i = 1, \cdots, n)$。

3．疑错窗口结合 Markov 模型中文定错校对

在得到疑错窗口集后，我们便可以用 Markov 模型算法对每个疑错窗口中的字词进行定错和校对[34]。此方法首先构造疑错窗口中的各词的字候选矩阵，并根据字候选矩阵构造格子图上的节点向量。然后应用 t 元规划构造格子图上的语音结构元素节点，并用 Viterbi 算法寻找最佳候选节点序列。搜索中使用路径评价函数为 3.9 式。最后将搜索到的最佳词序列与原词序列对照，不一致者作为发现的文本错误输出，并将最佳词序列中的对应字词作为第一候选并对错误进行改正。具体算法描述如下：

Step1：构造要校对疑错窗口中词组 S 的基字序列 $Baseword = \langle c_1, c_2, \cdots c_n \rangle$。

Step2：依据系统词典，对词组 S 中每个字选择 5 个高频同音字，构造 Baseword 中的字候选向量 Z_i，形成 S 的字候选矩阵 $Martix(S) = Z_1 Z_2 \cdots Z_n$，$Z_i = \langle c_{i,1}, c_{i,2}, \cdots, c_{i,m} \rangle$，$c_{i,j}$ 为基字 c_i 的字候选向量 Z_i 中 c_i 的同音字，m 为候选向量的长度。

Step3：依据公式 3.10，可利用 Viterbi 算法寻找最佳候选节点序列，对应于最佳候选节点序列的评价函数为 $f^*(wordlist) = \arg\max \prod_{i=1}^{n} \lambda ip(C_i \mid C_{i-1})$，同时保留前几个具有较高评价函数的字序列。

Step4：将搜索到的最佳词序列 Wordlist*与 Baseword 对照，不一致者作为发现的文本错误输出，并将 Wordlist*中的对应字词作为第一候选并对错误加以改正，在 Step5 中保留的具有较高评价函数的候选序列中的字词作为其余候选。

上述算法中，候选矩阵的构造很关键，因为能否发现文本错误以及给出的改正候选字词集中的正确字词是否被之覆盖均与之有关。根据文献[9]可采用两个原则来完成矩阵的构造：一是采用高频同音字，二是采用高频同音词。这样基本保证了查错率、第一候选的准确率及候选字词的覆盖率能满足实际的需要。所以本

研究中选择前5个高频同音字组成候选矩阵。

3.2.3.3 实验与结果分析

实验中使用约一千万字的新闻、报刊语料作为统计模型的参数训练语料。测试语料由800篇未校对的真实错误文本组成。各文本中主要包含错字和错词，共1320处。

为了评价该算法的性能，本实验采用召回率，查准率和修正率作为评价指标，以同音特征构造混淆集。同时在相同的条件下，对提出的新算法和基本疑错窗口算法作了对比测试。测试结果如表3.1所示。

表3.1 各算法测试结果对比

评价指标	改进型疑错窗口结合Markov	基本疑错窗口算法	基本Markov算法
召回率	82.24%	76.32%	79.83%
查准率	73.59%	65.73%	68.48%
订正率	54.14%	42.14%	49.44%

从表3.1可以看出，由于新算法对词间连续性判断进行了改进，有效的提高了疑错窗口的定位效果，同时由于新算法结合Markov模型作为定错和纠错算法，分别对各个疑错窗口中的词组进行定错和纠错，从而不仅较有效的弥补了Markov模型只能处理语音的临近约束关系这一缺点而且有效的提高了订正率。

3.2.4 去噪预处理

去噪处理模块主要完成去除稀有词，停用词工作[38]。本研究中的去除稀有词和停用词方法分别采用词频比较与停用词表方法。

（1）稀有词处理：

首先对分词后的各个词条频率进行统计，并设定一个词频阈值，只要是词条频度低于这个词频阈值的词就从词条空间中删除。

（2）停用词表

首先构造停用词表，在文本校对处理后把校对后的会话中的各词条分别和停用词表中词进行比较，如果该词条在停用词表中，则删除该词条。

3.3 短信文本特征选择研究

通过用向量空间法表示词条后，文本特征向量的维数往往达到数十万维[36]，即使经过删除停用词表中的停用词以及删除低频词，仍会有数万维特征留下。在这数万维特征中只有部分特征具有代表性，特征选择的目的就是从初始特征集中选出最具有代表性的特征集，从而不仅减少特征维数，降低了分类的空间和时间复杂性，而且去除掉了干扰特征或称为噪声特征，从而提高了分类精度。但是传统的特征提取算法都假设训练特征集是干净特征集，即训练特征集中不存在错误训练文本。但是在实际应用中，很难保证训练文档集的可靠性，特别是针对短信文本的情况[37]。因为训练时，需要大量的短信作为训练数据集对模型进行训练，但是由于训练文本数量巨大，很难人工的对每条训练数据进行精确标号，所以构造的训练集往往含有一些噪声数据，这些数据可能不应属于该类，而应属于其它类别或未知类别。因而，当训练集包含部分噪声数据的情况下，用传统的特征提取方法提取特征会造成特征集包含大量噪声特征，从而降低分类时的准确性[38]。

本文提出了一种新的重复特征选择方法。这种方法基于 EM（最大期望值）算法，通过自适应的方法对特征空间最优化，每次循环都除去一部分的噪声特征，从而构建了最优的特征空间，提高了分类器的分类能力。但是从试验中，发现当训练数据较少时，该算法可能产生过拟合。为了解决该算法的过拟合问题，本次研究对算法进行了改进，有效的控制了过拟合问题。具体特征提取模块理论及算法描述如下。

3.3.1 重复特征选择原理及算法

发现在文本分类应用中，分类和特征选择之间可以互补[39]。一方面更好的分类结果将提供更准确的训练数据集，从另一方面来说，更好的特征集将帮助分类器提高分类效果，以至于得到每一类的更好的训练数据集。所以根据这一事实，提出了重复特征选择（IFS）技术。该技术理论描述如下。

设 $D=\{d_1, d_2, \cdots., d_N\}$ 为训练文本集，d_i 是训练文本集中第 i 个文本；$C=\{c_1, c_2, \cdots., c_\beta\}$ 是类别集合，c_j 是第 j 类；$t=\{t_1, t_2, \cdots, t_L\}$ 是这些类的特征集合；$\theta=\{\theta_1, \theta_2, \cdots., \theta_c\}$ 是模型参数空间 θ_α 是第 α 类的模型参数。

分类系统中训练过程的目标是调整模型参数，从而 θ 使似然概率值 $p(D|\theta)$ 最大化。假设训练文本集中各文本相互独立的情况下，似然概率 $p(D|\theta)$ 可写为

下式：

$$p(D|\theta)=\prod_{i=1}^{N}p(d_i|\theta) \tag{3.11}$$

$$p(d_i|\theta)=\sum_{j=1}^{\beta}p(c_j|\theta)p(d_i|c_j,\theta) \tag{3.12}$$

公式 3.12 中 $p(c_j|\theta)$ 是类 j 的先验概率，$p(d_i|c_j,\theta)$ 是给定模型参数 θ 时，文本 i 在类 j 中的概率。进一步假设特征集中的特征也相互独立的情况下，似然函数可以重写为下式：

$$p(D|\theta)=\prod_{i=1}^{N}\sum_{j=1}^{\beta}p(c_j|\theta)\prod_{t_l\in d_i}p(t_l|c_j,\theta) \tag{3.13}$$

这里 $p(t_l|c_j,\theta)$ 是在给定模型参数 θ 时，类 j 中文本 d_i 中特征 t_l 的概率。并不是所有特征都与类有相同的相关度，所以全概率公式，$p(t_l|c_j,\theta)$ 可以看为相关分布和不相关分布的加权和，如下式所示。

$$\begin{aligned}p(t_l|c_j,\theta)=&z(t_l)p(t_l\ is\ \ relevant|c_j,\theta)+\\&(1-z(t_l))p(t_l\ \text{is}\ irrelevant|c_j,\theta)\end{aligned} \tag{3.14}$$

这里 $z(t_l)$ $p(t_l\ is\ relevant)$ 被定义为特征 t_l 是相关的概率。因此似然函数（19）可以重写为 3.15

$$\begin{aligned}p(D|\theta)=\prod_{i=1}^{N}\sum_{j=1}^{\beta}p(c_j|\theta)\prod_{t_l\in d_i}&z(t_l)p(t_l\ \ is\ \ relevant|c_j,\theta)+\\&(1-z(t_l))p(t_l\ \text{is}\ irrelevant|c_j,\theta)\end{aligned} \tag{3.15}$$

可以用 EM 算法通过循环以下两步来最大化似然函数值。

$$E-step:\hat{z}^{(k+1)}=E(z|D,\hat{\theta}^{(k)}) \tag{3.16}$$

$$M-step:\hat{\theta}^{(k+1)}=\arg\max_{\theta}p(D|\theta,\hat{z}^{(k)}) \tag{3.17}$$

这里 z={z(1), z(2), ….. z(u)}，u 是特征个数。

在 E-step 中，在给定第 k 次循环后的模型参数 θ^k 的情况下，计算期望的特征集 $\hat{z}^{(k+1)}$。在 M-step 中，根据 E-step 中的到的新的特征空间，计算新的模型参数 $\hat{\theta}^{(k+1)}$。这个模型参数使似然函数最大化。具体实现如下：

E-step 中，首先对上次循环得到的训练集进行重新分类，并且把错误分类结果从训练集中删除。然后对分类后真确的训练集中的每篇文档进行分词，预处理

工作并得到新的特征向量空间z'，并且用监督型的特征选择算法如 CHI 或 IG 计算新特征向量空间中每个特征的相关度分数。如果特征t_l的相关度分数大于一预设门限 T，则记$z'(t_l)=1$，否则记为$z'(t_l)=0$。如果$z'(t_l)=0$，则将该特征从新特征向量空间中删除。最后用z'更新 z 得到$\hat{z}^{(k+1)}$。所以在每次循环过程中，通过移除错误分类训练文档和相关度分数判断的方法去掉了噪声训练文本和噪声训练特征，从而自适应的最优化了特征集。

M-step 中，根据重构的训练文本集，使用任何的训练算法对模型进行重新训练，得到新模型参数$\hat{\theta}^{(k+1)}$。

完整的算法流程描述如下：

给定训练数据集 D

初始化 $z^1=\{1,1\cdots..1\}$

while (IEnd())

- 基于特征集$\hat{z}^{(k)}$和训练文本集训练模型参数$\hat{\theta}^{(k+1)}$；
- 用分类算法对训练文本进行重新分类，并得到正确和错误的分类结果集；
- 删除训练文本集中的错误分类结果；
- 更新 Z

  根据真确分类结果集构建新文本特征向量z'，然后计算新特征向量集中每个特征的相关度分数α_l

  if (α_l>T)

  $z'(t_l)=1$

  else

  $z'(t_l)=0$

  $\hat{z}^{(k+1)}=z'$

End while

end

算法流程图如下图所示：

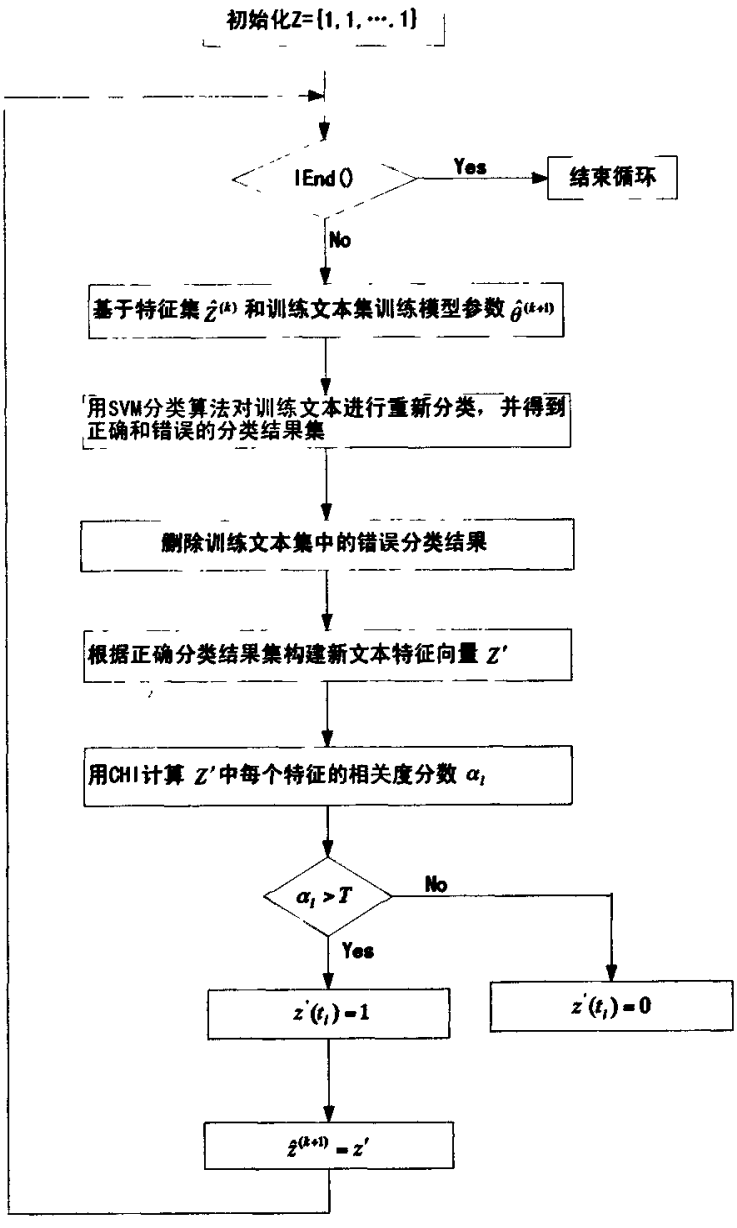


图 3.1 重复特征选择流程图

这里 IEnd()为循环结束判断函数。根据 EM 算法，循环结束函数可以写为下式。这里γ是一个很小的值，可以更具经验确定。

$$IEnd() = \begin{cases} 1 & (dis(\theta^{(k)},\theta^{(k+1)}) = \| \theta^{(k)} - \theta^{(k+1)} \|_2 < \gamma \\ 0 & otherwise \end{cases} \quad 3.18$$

通过试验表明：用上式作为循环结束函数可能会造成过渡匹配，因此会降低算法的性能。这主要是由于 3.18 的目标仅仅是使训练文本集最精确化，但是它忽略了训练文本的鲁棒性。

为了解决以上问题，对算法进行了进一步改进。通过把标准 $miroP > \eta$ 和传统的收敛标准相结合作为新标准用于判断循环结束。新的循环结束函数写为下式：

$$IEnd() = \begin{cases} 1 & (dis(\theta^{(k)}, \theta^{(k+1)}) < \gamma \ or \ miroP > \eta) \\ 0 & otherwise \end{cases} \tag{3.19}$$

$$miroP = \frac{\sum_{i=1}^{\beta} \alpha_i}{\sum_{i=1}^{\beta} \gamma_i} \tag{3.20}$$

这里 η 是预先设定的门限值，根据试验结果可以取 0.85~0.90 之间。α_i 是第 c_i 类中被正确分类的文档数；γ_i 是类 c_i 的文档总数。

采用 $miroP > \eta$ 作为循环结束判断标准主要有以下三个原因：

1) 当 $miroP$ 值足够大时，可以认为训练文本集中的噪声文本数较小。

2) 通过限定 η 的最大取值，可以保存一部分噪声文本，由此保证训练文本的广泛性和鲁棒性，从而控制过渡匹配。

3) 根据通常情况，$miroP > \eta$ 往往比传统的收敛性标准更早达到，所以新标准可以提高 IFS 算法的性能。

3.3.2 实验及结果分析

为了验证 IFS 算法的有效性，对 IFS 算法进行了测试。该实验使用 6 类文本数据（教育，经济，计算机，军事，环境，交通），每类数据的训练集由 2000 篇文档组成，测试集由 500 篇文档组成。各文本从 internal 网上获得。

试验中分别选用 SVM，ICTCLAS 和 IG 作为分类器、分词算法和基本的特征提取算法。特征 tl 的信息增益 IG 公式描述如下：

$$IG = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t_l)\sum_{i=1}^{m} P(c_i \mid t_l)\log P(c_i \mid t_l) + P(\bar{t}_l)\sum_{i=1}^{m} P(c_i \mid \bar{t}_l)\log P(c_i \mid \bar{t}_l)$$

为了评估新算法的有效性，选择基本的 CHI 和 IG 算法作为对比算法。同时选择精度，召回率和 miroF1 作为评估标准。这些标准定义为下式。

$$R_i=\frac{\alpha_i}{\mu_i},p_i=\frac{\alpha_i}{\gamma_i},miroR=\frac{\sum_{i=1}^{m}\alpha_i}{\sum_{i=1}^{m}\mu_i},miroP=\frac{\sum_{i=1}^{m}\alpha_i}{\sum_{i=1}^{m}\gamma_i},\ miroF1=\frac{2\times miroR\times miroP}{miroR+miroP}$$

这里是第c_i类中被正确分类的文档数；γ_i是类c_i的文档总数；μ_i是被分类到c_i类的文档总数。

试验中以不同的百分用噪声训练数据替换了原有的训练数据的方法构建了含噪训练数据。表 3.2，3.3，3.4 分别显示了采用 CHI、IG、IFS、改进型 IFS 作为特征选择算法的测试结果。其中表 3.2 是对干净训练集进行特征提取得到的测试结果，表 3.3 和表 3.4 分别是对含有 20%噪声和 35%噪声数据的训练集进行特征提取而得到的测试结果。

表 3.2 干净训练集条件下各算法测试结果

categories	precision				recall			
	CHI	IG	IFS	IIFS	CHI	IG	IFS	IIFS
computer	0.988	0.984	0.984	0.984	0.976	0.964	0.966	0.964
education	0.978	0.980	0.978	0.980	0.950	0.962	0.960	0.962
economy	0.968	0.972	0.970	0.972	0.918	0.927	0.929	0.927
environment	0.974	0.982	0.978	0.982	0.951	0.966	0.955	0.966
traffic	0.982	0.988	0.988	0.988	0.972	0.978	0.978	0.978
military	0.976	0.974	0.974	0.974	0.966	0.966	0.962	0.966
MiroF1	0.966	0.969	0.967	0.969				

表 3.3 训练数据含有 20%噪声数据情况下，各特征提取算法测试结果

categories	precision				recall			
	CHI	IG	IFS	IIFS	CHI	IG	IFS	IIFS
computer	0.964	0.964	0.966	0.980	0.914	0.924	0.939	0.958
education	0.962	0.968	0.972	0.980	0.928	0.934	0.945	0.962
economy	0.938	0.942	0.962	0.968	0.854	0.862	0.897	0.916
environment	0.948	0.962	0.974	0.978	0.911	0.928	0.951	0.964
traffic	0.944	0.940	0.972	0.982	0.914	0.910	0.952	0.974
military	0.952	0.958	0.968	0.974	0.924	0.931	0.952	0.966
MiroF1	0.928	0.934	0.948	0.965				

表 3.4 训练数据含有 35%噪声情况下，各特征提取算法测试结果

categories	precision				recall			
	CHI	IG	IFS	IIFS	CHI	IG	IFS	IIFS
computer	0.912	0.918	0.950	0.972	0.829	0.833	0.896	0.940
education	0.922	0.924	0.958	0.974	0.856	0.858	0.921	0.947
economy	0.904	0.898	0.952	0.962	0.758	0.761	0.839	0.902
environment	0.896	0.902	0.940	0.972	0.845	0.850	0.907	0.952
traffic	0.918	0.924	0.952	0.962	0.872	0.878	0.937	0.956
military	0.928	0.928	0.954	0.968	0.882	0.895	0.934	0.954
MiroF1	0.874	0.878	0.926	0.954				

从表 3.2 到表 3.4 我们可以看出：

1）在干净训练集的情况下，每个特征提取算法都有较高的分类效果。

2）随着噪声训练数据的增加使用 CHI 和 IG 作为特征选择算法的系统分类能力随之下降。这是因为 CHI 和 IG 算法不能很好的去除掉噪声训练文本带来的噪声特征。相比之下，使用 IFS 和 IIFS 算法作为特征选择算法的系统分类有更好的分类效果，同时系统的性能也校对稳定。这是由于 IFS 和改进型 IFS 算法自适应的重复的减少了噪声训练数据和噪声特征，从而得到了较精确的特征集。

除此之外，我们同样发现，相比基本的 IFS 算法，改进型 IFS 算法有更好的分类效果。这是由于改进型 IFS 使用了改进的循环结束判断函数对基本 IFS 算法的过渡匹配问题加以了控制。相比基本的 IFS 算法，改进型 IFS 算法不仅能够压缩噪声特征，同时还可以保证训练数据的鲁棒性。

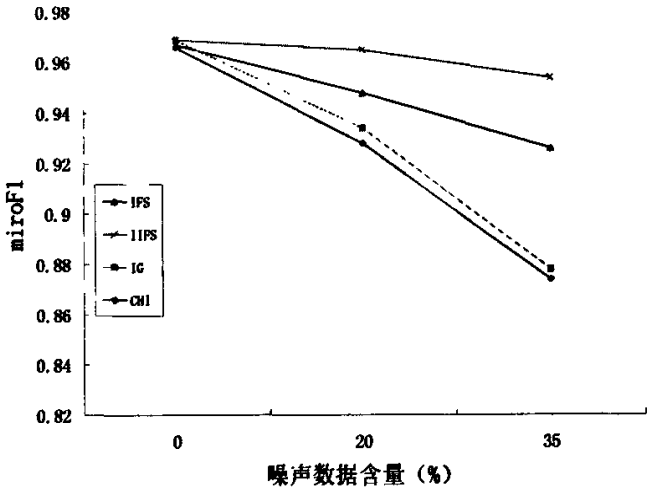


图 3.2 各模型在不同程度的噪声元素含量情况下 miroF1 值

图 3.2 为各模型在不同程度的噪声元素含量情况下 miroF1 值。综上所述，在训练数据集存在一些噪声数据的情况下，我们提出的重复特征选择方法相比 CHI 和 IG 方法有更好特征选择能力，因此有较好的分类效果。

3.4 基于兴趣关键特征词表的特征权重调整技术

文本训练和分类的目的分别是使 $p(\theta_i|D)$ 和 $p(D|\theta)$ 概率值最大[40]，其中 i 是 D 所属于的类。从 3.14 式和贝叶斯公式我们可以得出，$p(\theta_i|D)$ 和 $p(D|\theta)$ 的大小主要由 $p(t_l|c_j,\theta)$ 决定。当 t_l 为关键词表中的特征词时，并且 $p(t_l|c_j,\theta)$ 较小时，我们可以对 $p(t_l|c_j,\theta)$ 进行适当的调整，使 t_l 在模型中所起作用增大，从而突出兴趣关键特征词表中词的作用使特征集和模型更能反映用户的兴趣分布。从前面可以看出 $p(t_l|c_j,\theta)$ 的大小主要由权重的大小来决定，下面将描述基于兴趣关键特征词表的特征权重调解方法。

权重调节过程要实现两个目标：

1. 使调解后在兴趣关键特征词表中的词在文本中的权重较大
2. 调整权重后的在兴趣关键特征词表中的词的权重不能完全掩盖其它特征词在文本中所起的作用。

因此只当在兴趣关键特征词表中的词的权重较小时对该词权重进行调整，同时为了实现第 2 条目标，调整范围必须加以限制。以下我们对权重调整方法进行

描述。

由于均值可以反映随机量的分布情况，所以这里使用均值作为权重调整标准，当在兴趣关键特征词表中的词在该训练文本中的权重小于该训练文本中所有词的权重的均值时，用该均值取代该词原有的权重。

为了实现使调整后在兴趣关键特征词表中的词在文本中的权重较大和调整权重后的在兴趣关键特征词表中的词的权重不能完全掩盖其它特征词在文本中所起的作用两个目标。我们计算了除去在兴趣关键特征词表中词后的待分类文本属于各个类别的概率P_i。然后根据计算得到的P_i序列来判断是否对在兴趣关键特征词表中的特征词的权重进行调整。假设待分类文本的特征词集为C，其中属于兴趣关键特征词表中的词集为M，则不属于该兴趣关键特征词表的词集表示为C-M。

P_i计算公式如下：

$$P_i(d|c_i)=\prod_{k\in C-M} p(t_i|c_i) \qquad 3.21$$

根据计算得到的P_i序列，可以分以下几种情况对特征权重进行调整：

1. 当兴趣关键特征词集M中包含不止一个词时$\{t_1,\cdots,t_m\}$，各个特征词都属于一个类j，并且M的势远小于C-M的势时。则当$P_j-P_{max}=T<T_l$时，且该特征词的权重小于该文本中所有词的权重的均值时，用该均值取代该词原有的权重对该特征词的权重进行调整。这里T_l是预设定的门限值。$P_{max}=\max\{P_i|i=1,...,N\}$。

2. 当兴趣关键特征词集M中包含不止一个词时$\{t_1,\cdots,t_m\}$，各个特征词都属于一个类j，并且M的势远小于C-M的势时。则当$P_j-P_{max}=T>T_l$时，则表示该待分类文本不属于类j，为了满足第2个目标，所以对兴趣关键特征词表中的词的权重不作调整。

3. 当兴趣关键特征词集M中包含不止一个词时$\{t_1,\cdots,t_m\}$，各个特征词都属于一个类j，并且M的势接近于C-M的势时，则直接对兴趣关键特征词集M中特征词的权重小于该文本中所有词的权重的均值的特征词，用该均值取代该词原有的权重对该特征词的权重进行调整。

4. 当兴趣关键特征词集M中包含不止一个词时$\{t_1,\cdots,t_m\}$，M中特征词属于多个类$\{c_i,\cdots,c_{i+p}\}$，并且M的势远小于C-M的势时。同样首先计算$P_k-P_{max}=T_k$（k=i,…,i+p），如果存在$T_k<T_l$（k=1,..,0），且k类所对应的兴趣关键特征词的权重小于该文本中所有词的权重的均值时，用该均值取代该词原有的权重对该特征词的权重进行调整，且最大P_k的k类所对应的特征词再相应增大均值的20%。

5. 当兴趣关键特征词集M中包含不止一个词时$\{t_1,\cdots,t_m\}$，M中特征词属于

多个类$\{c_i, \cdots, c_{i+p}\}$，并且 M 的势接近于 C-M 的势时。则直接对兴趣关键特征词集 M 中特征词的权重小于该文本中所有词的权重的均值的特征词，用该均值取代该词原有的权重对该特征词的权重进行调整。

3.5 特征规范化处理

对于训练和识别任务要求特征越正交越好，所以在提取特征后，对特征进行了进一步正交化。通过正交化，不仅使特征集更具有区别性，而且通过正交变化，去除了各特征之间的冗余进一步对特征进行了压缩[41]。本文中使用离散余弦变换作为规整变换，算法描述如下。

假设给定 N 维特征矢量序列 x(n)，n=0, 1, …, N-1，其离散余弦变换定义为

$$X_c(0) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \tag{3.22}$$

$$X_c(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos \frac{(2n+1)k\pi}{2N}, k = 1,2,\ldots,N-1 \tag{3.23}$$

其中 $X_c(k)$ 为变换后标准化的第 k 维特征量。

3.6 试验及结果分析

为了测试本研究的有效性，对新系统和传统的文本特征提取系统进行了对比模拟测试。试验中使用了三类文本数据（计算机，汽车，教育）。因为真实短信测试数据很难获得，所以采用论坛上的短文本和人工构造的类似短信数据的短文本组成训练集和测试集。构造方法通过把长文本分成短文本，每一短文本作为一条短信数据。训练数据每类有 800 条文本，测试集每类有 500 条文本。长文本和论坛文本从 internet 网上获得。

试验中分别选用 SVM，ICTCLAS 和 IG 作为分类器、分词算法和基本的特征提取算法。传统文本分类系统的文本特征选择算法同样使用 IG 算法。同时选择信息熵和特征维数作为评估标准。表 3.5 为传统特征提取算法和新方法所的测试结果。

表 3.5 传统特征提取算法和新方法的测试结果

类别	Entropy		特征维数	
	传统方法	新方法	传统方法	新方法
计算机	2.305	1.870	320	243
汽车	2.02	1.794	284	197
教育	2.185	1.825	308	236

从表 3.5 可以看出新方法所提取的文本特征的信息熵要低于传统系统所提取的信息熵，这说明，新方法所提取的特征有更好的表征类型的能力。同时从表 3.5 可以看出，新方法相比传统方法有更好的特征压缩效果。

3.7 本章总结

短信文本的非结构化特征决定了数据预处理环节的重要性。数据预处理部分主要任务是对短信文本进行必要的规范化处理，并以适当的形式表示短信文本，提高后继处理效率。如果不对短信文本进行必要的预处理，使用再好的分类方法也不能得到满意的结果。

本章首先给出短信文本预处理的逻辑框架，并概述预处理阶段的各项主要工作。然后本文对其中的一些工作进行了扼要的描述，最后，文中论述了对错字、同音异形词校对，文本特征选择等重要的数据预处理方法。

本文中提出了一种新的自动中文校对技术。此方法有效地解决了短信文本中的错字、错词等对系统的训练和分类过程带来的影响。并且本文把中文文本规范化和新文本校对方法相结合，有效地消除了短信文本中的变异符号、变异文本，提高了分类的精确性。

针对含有噪声的训练数据，本文提出一种自适应的基于 EM 算法的最优化特征选择方法，该方法能够自适应的对特征空间进行改变，去除噪声特征，最终得到最优的低维特征空间，同时此算法也有效的解决了训练集含有噪声训练元素情况下的最优特征提取。

最后提出了一种基于兴趣关键特征词表的特征权重调整方法，有效的反映了用户兴趣点。最后实验结果验证了文中提出的方法的有效性。

第四章 短信文本分类技术的研究

文本分类是大规模文本处理重要的应用技术之一。长期以来，文本分类一直都是自然语言处理中一个重要的应用领域；80 年代末以前，在文本分类方面占主导地位的一直是基于知识工程的分类方法，即由专业人员手工编写分类规则来指导分类，其中最著名的系统是为路透社开发的 Construe 系统。90 年代以来，随着信息存储技术和通信技术的迅猛发展，大量的文字信息开始以计算机可读的形式存在，并且其数量每天仍在急剧增加。这一方面增加了对于快速、自动的文本分类的迫切需求，另一方面又为基于机器学习的文本分类方法准备了充分的资源。在这种情况下，基于机器学习的文本分类方法逐渐取代了基于知识工程的方法，成为文本分类的主流技术。基于机器学习的文本分类通常由训练和分类两个阶段组成，在训练接阶段，从训练文本中学习分类知识；在分类阶段，则根据分类器将输入文本分到最可能的类别中。

在本研究中，将文本分类技术用于短信文本的挖掘中，并根据短信文本的特点做了一定的改善，使其具有不错的效果。

4.1 文本分类方法

分类方法很多，比如采用自组织映射(Self-Organising Map, SOM)和 BP 网络混合的方法进行自动分类[42][43]。目前常用的文本分类方法主要有朴素贝叶斯分类算法(Naive Bayesian classifier)、K—最近邻参照分类算法(k—Nearest Neighbor)以及最近中心分类法。

(1) 利用贝叶斯分类算法进行文本分类

若文本采用 DF 向量表示方法，即文档向量的分量为一布尔值，0 表示相应的单字在文档中未出现，1 表示出现，则采用该表示方法的文档 Doc 属于 C 类文档的概率为：

$$P(C|Doc)=\frac{P(C)\prod_{F_j\in V}P(Doc(F_j)|C)}{\sum_i P(C_i)\prod_{F_i\in V}P(Doc(F_i)|C_i)} \qquad 4.1$$

$$P(Doc(F_j)|C)=\frac{1+N(Doc(F_j)|C)}{2+|D_c|} \tag{4.2}$$

其中 $P(Doc(F_j)|C)$ 是 C 类文档中特征 F_j 出现的条件概率的拉普拉斯概率估计，$N(Doc(F_j)|C)$ 是 C 类文档中特征 F_j 出现的文档数，$|Dc|$ 为 C 类文档所包含的文档数目。

若文档采用 Tf 向量表示法，即文档向量的分量为相应单词在文档中出现的频率，则采用该表示方法的文档 Doc 属于 C 类文档的概率为：

$$P(C|Doc)=\frac{P(C)\prod_{F_j\in V}P(F_j|C)^{TF(F_j,Doc)}}{\sum_i P(C_i)\prod_{F_i\in V}P(F_i|C_i)^{TF(F_i,Doc)}} \tag{4.3}$$

$$P(F_j|C)\frac{1+TF(F_j,C)}{|V|+\sum_i TF(F_i,C)} \tag{4.4}$$

其中，$P(C)$ 为一个文档属于 C 类的概率，$P(F_j|C)$ 使对 C 类文档中特征 F_j 出现的条件概率的拉普拉斯概率估计，$TF(F_j|C)$ 是 C 类文档中 F_j 出现的频度，$|V|$ 为单字词典的打小，等于文档表示中所包含的不同特征的总数目，$TF(F_j,Doc)$ 是文档 Doc 中特征 F_j 出现的频度。由于在单字辞典集 $|V|$ 中许多特征 F_j 均不出现在 Doc 中，从而 $TF(F_j,Doc)=0$，所以可将上述(4.3)中的第一公式变作：

$$P(C|Doc)=\frac{P(C)\prod_{F_j\in Doc}P(F_j|C)^{TF(F_j,Doc)}}{\sum_i P(C_i)\prod_{F_i\in Doc}P(F_i|C_i)^{TF(F_i,Doc)}} \tag{4.5}$$

(2) 利用 K 最近邻参照分类算法进行文本分类

K—最近邻参照分类算法将对一个文档的所属类别范畴的预测建立在与之最为相识的 k 个文档所属类别的概率分布上。文档 Doc 属于 C 类文档的概率为：

$$P(C|Doc)=\frac{\sum_{l=1}^{k}similarity(Doc,D_l)P(C|D_l)}{\sum_i\sum_{l=1}^{t}similiarity(Doc,D_l)P(C_i|D_l)} \tag{4.6}$$

其中 D_l 为文档 Doc 最邻近的 k 个文档之一，它既可以按照概率属于不同的类别，也可以属于唯一的一个类别 C_k（这时 $P(C_1|D_1)=1$，当 $i=k$ 时，否则，$P(C_1|D_1)=0$）。

(3) 最近中心分类法

把文本分为训练文本和测试文本两个部分，并对训练文本进行人工分类。设

训练文本集D共分为K类，即D_1、D_2、.....、D_k，每类分别有S_1、S_2、.....、S_k篇文档，总文档数$S=\sum_{i=1}^{k}S_i$，D_k类中包含的文档为d_{k1}、d_{k2}、.....、d_{ksk}。现分别计算每类文本的D_k的聚类中心C_k类(即该类别中所有的文档向量的几何中心)，则有：

$$C_k=(C_{k1},C_{k2},\cdots,C_{km})，其中C_{kj}=\frac{1}{S_k}\sum_{i=1}^{sk}w_{ij}，\quad 1\le k\le K,1\le j\le m \qquad 4.7$$

和文本相似度一样，我们可以利用聚类中心定义类间相似度$sim(D_i,D_j)$及文档d与类别D_k的相似度$sim(d,D_j)$。

4.2 短信文本分类方法研究

在研究过程中我们发现，短信文本有它自身的特点，适用现有固定的方法在分类过程中存在着一定的缺陷，而且在分类精度上很难再有很大程度的提高，为了在一定程度上弥补缺陷并获得较高的正确识别率，文本采用改善的分类方法进行研究。

4.2.1 短信预处理与整合

由于短信字数的限制（一般少于 70 个汉字），以及输入不便等各种原因，通过步骤 1 得到的短信数据存在数目极大，而且内容主题分散等问题，这些问题将可能造成后续的文本分类过程时间复杂度的急剧攀升，同时严重影响用户需求的准确挖掘[44]。此外，由于目前对点对点群发短信的监管存在一定困难，而这些群发短信往往并不反映手机用户的兴趣和需求，因此不必对这些号码发送的短信进行分析[45]。

本研究采取两个步骤解决上述问题。首先，剔除群发短信号码。根据数据采集时间设定一阈值 k，如果某个短信号码发送的短信条数超过该阈值，就判该号码为短信群发号码，需要将该号码发送的所有短信数据从短信数据库中删除。对于号码发送短信条数的判断可利用数据库管理系统的统计功能实现。阈值的选取一般应取明显异常的值，例如，如果短信采集时间为一天，则可取阈值 k=300；如果采集时间为一月，则可取阈值 k≈2000。

其次，针对短信字数较少，往往需要连续几条短信才能表达明确的内容，而且与不同的接收对象所交流的主题并不一定相同的问题，我们利用短信文本的时

间相关性和对象相关性，按照短信内容进行聚类。时间相关性和对象相关性可以通过对短信数据库排序得到，其中主关键字为短信发送手机号码，次关键字为短信接收号码。短信聚类的好处是使得短信文本的数量极大减少，同时使得文本主题相对集中，便于后续短信分类。

为降低聚类算法的复杂度，我们提出一种基于滑动窗口的文本整合方法。该方法根据系统的实际情况，确定一个合适的窗口 w，新文本仅需和最近的 w 个已整合短信文本进行相似度计算，对相似度高于阈值的最相似短信文本进行整合。通过适当调整 w 值，该算法在保证效果的同时使得时间复杂度可控。

图 4.1 所示为本研究实施例的短信预处理与整合流程图。该流程需要事先指定群发度阈值、相似度阈值和滑动窗口尺寸，这些参数可根据需要进行适当调整。更进一步的实现为系统根据已有经验自行学习并采取合适的值。然后本研究利用数据库管理系统提供的现有功能对短信数据排序。在短信排序时我们利用了短信数据流特有的时间相关性和对象相关性，因此相邻的短信内容大体相同或者近似，这就非常便于下面的短信聚合。

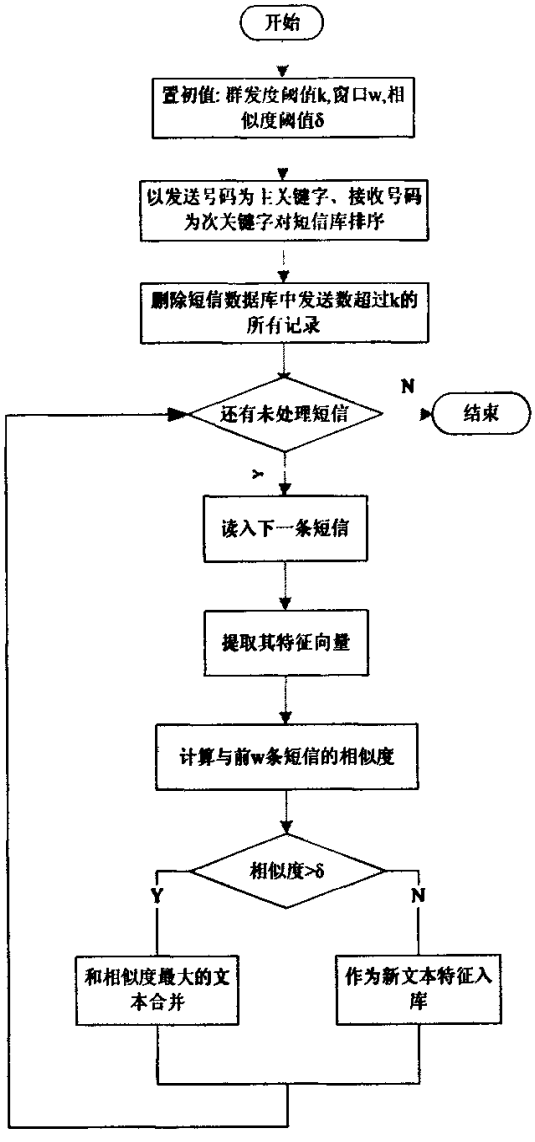


图 4.1 短信数据预处理与整合

系统随后逐条读入短信，判断是否为同一用户所发，如果是同一用户的短信，则判断该短信内容是否与前 w 条相似，如果相似就进行合并。否则作为新的文本，同时窗口向后滑动一格。

文本相似度的计算方法如下：对两个文本 S1 和 S2，令它们的所有特征词构成的向量空间为 V = { X1, X2, X3, ..., Xn }，其中 Xi 为特征项。设文本 S1 的特征向量$V_1 = (\omega_1, \omega_2, \cdots, \omega_n)$，其中$\omega_i$为特征词 Xi 在文本 S1 中的频度；文本 S2 的特征向量$V_2 = (\varphi_1, \varphi_2, \cdots, \varphi_n)$，其中$\varphi_i$为特征词 Xi 在文本 S2 中的频度。则两个文本的相

似度按照下式计算：

$$Sim(S_1,S_2)=\overrightarrow{V_1}\cdot\overrightarrow{V_2}=\frac{\sum_{i=1}^{n}\omega_i\cdot\varphi_i}{\sqrt{\sum_{i=1}^{n}\omega_i^2}*\sqrt{\sum_{i=1}^{n}\varphi_i^2}} \quad 4.8$$

文本的合并方法为直接按照特征值对应相加并规范化。设文本 S1 和文本 S2 的特征向量表示如上，将其各特征项所对应的特征向量相加，然后对其规范化。新文本的发送时间为新近被合并文本的发送时间。

合并后的文本向量规范化方法优选采用 z-score 规范化方法（也称为零-均值规范化）。设直接按照特征值对应相加后所得到的文本特征向量为$V=\{v_1,v_2,\cdots,v_n\}$，其中v_i为相加后文本的特征项 Xi 所对应的特征向量的频度。设规范化以后所得到的特征向量为$V=\{\phi_1,\phi_2,\cdots,\phi_n\}$，其中$\phi_i$为合并且规范化以后文本的特征项 Xi 所对应的特征向量的频度，其计算方法如下：

$$\phi_i=\frac{v_i-\bar{v}}{\sigma_v}$$

其中，$\bar{v}$与σ_v分别是属性v_i的平均值与标准差。

经过短信预处理及整合后的短信文本具有如下格式：

发送方 ID	发送日期、时间	短信文本向量

它可以存入数据库中，也可以存为文件或者其它形式。

由于利用了短信特有的时间相关性、对象相关性和内容相关性，经过整合的短信文本主题相对集中，同时极大的减少了短信数量，使得整合后的短信文本更易于后续挖掘任务。由于采用了“加窗的短信内容整合技术”，该技术能在可控的时间范围内取得很好的效果。

4.2.2 短信文本分类

短信文本分类用于将手机用户发送的短信文本分类到系统预先定义的类别中。中文文本分类技术主要包括多分类器集成学习的方法、支持向量机（SVM）、kNN 方法、朴素贝叶斯方法、决策树、神经网络、最大熵模型等[46][47]，它们都可用于本研究的分类过程。由于 SVM 的分隔面模式有效地克服了样本分布、冗余特征以及过拟合等因素的影响，具有很好的泛化能力，在效果和稳定性上的相对其

它方法具有优势，因此本研究优选 SVM 方法作为分类算法。此外，由于最大熵模型也具有类似优势，也是本方面优选的分类算法。

下面选择 SVM 方法的实现作为一个实施例，然而本研究显然不局限于使用 SVM 算法。在具体实现时，既可以直接实现 SVM 算法，也可以利用现有的软件工具进行实现。本研究采用后者的方法，即利用 LIBSVM 软件包来实现 SVM 分类操作。LIBSVM 是台湾大学林智仁(Lin Chih-Jen)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。该软件包可以从 http://www.csie.ntu.edu.tw/~cjlin/免费下载。

在分类之前，需要准备训练文档并利用训练文档构建 SVM 分类器模型，然后利用生成的分类器模型对短信文本进行分类。

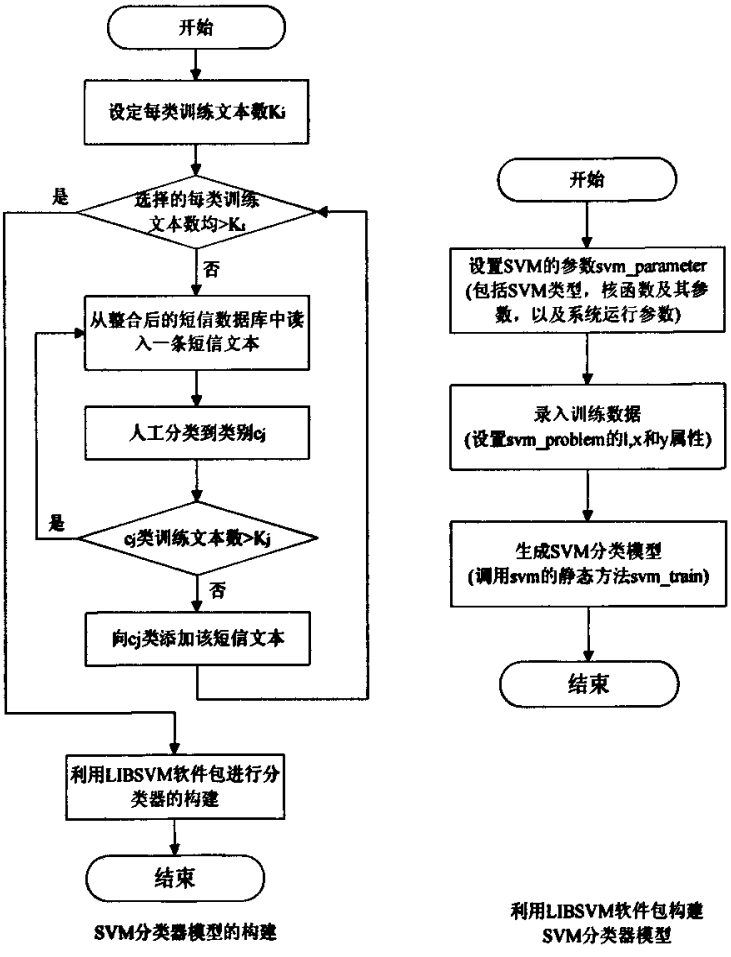


图 4.2 文本分类器构建

图 4.2 所示为本研究实施例的短信文本分类流程图。首先，从整合后的短信数据库中选择出若干条短信文本作为训练文本集，并对这些训练文本进行人工分类。训练文本的选择需要使得各个类别的文本数量差别不大。具体实现时，可以预先指定每类文本数目，例如 100，然后逐条从短信数据库中读入短信文本，对其进行人工分类。如果该类文本数量不足，则将该文本进行类别标记并放入训练集中；如果该类别文本数量已达到指定数目，则简单的丢弃该文本，重新从短信数据库中读入下一条文本。

得到训练文本后，需要提取该训练集文本的特征，并使用向量空间模型(Vector Space Model, VSM)将训练文本集中的文本表示为其对应的特征向量。特征向量的提取有多种方法，本研究实现时采用 tf×idf 方法，该方法的具体实现可以参考文献[52]

经过上述处理后，训练数据集可表示如下：

$$T = \{T_i \mid T_i = (W_i, c_i),\ \ c_i \in C\}$$

其中，Wi 为训练文本集中第 i 个训练文本的特征向量，C 为该特征向量的人工分类类别集。第 i 个文本的特征向量 Wi 表示如下：

$$W_i = (w_{i1}, w_{i2}, \cdots, w_{in})$$

其中 $w_{ik}(k = 1,2,\cdots d)$ 为特征项 k 对文本 i 的贡献程度，n 为特征向量的维数。人工分类类别集 C 表示如下：

$C = \{c_1, c_2, \cdots, c_m\}$

其中，m 为类别数。

接下来利用 LIBSVM 工具进行文本模型的训练，其步骤如下：

1) 设置系统参数。该参数可通过 LIBSVM 软件包提供的 svm_parameter 方法进行参数设定。本实施例中，选用 C_SVC 类型的支持向量机，其核函数(Kernel Function)使用径向基函数(Radical Base Function，RBF)：

$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \mid \mathbf{x}_i - \mathbf{x}_j \mid^2)$

设定 RBF 核函数的参数 γ 的初始值为 0.5；svm_type 属性，具有 C_SVC，NU_SVC，ONE_CLASS，EPSILON_SVR，NU_SVR 共五个可选值，本实施例选用 C_SVC；C 属性，表示分类的类别数，设置为类别集的元素个数，即为 m；kernel_type 属性，具有 LINEAR，POLY，RBF，SIGMOID，PRECOMPUTE 共五个可选值，本实例选择使用 RBF；shrinking 属性，本实例将其值设置为 1。此外，

从计算机运算的角度上出发，本实施例设置缓存大小为 40MB，运算的精度为 0.001，这些参数分别对应着 svm_parameter 的 cache_size，eps，shrinking 属性。综上所述，本实施例选择的参数为：

svm_type = C_SVC;

C = m;

kernel_type = RBF;

cache_size = 40;

eps = 0.001;

shrinking =1

2) 训练属性设置

设定 SVM 的参数后，将训练数据集做为 SVM 的输入，经过训练后产生 SVM 的分类器的分类模型。在 LIBSVM 软件包中，使用 svm_problem 来描述当前的分类问题。设置 svm_problem 的 l 属性为训练数据集 T 的元素个数，x 和 y 属性分别设置为训练数据集 T 的训练文本特征向量集和对应的训练文本的类别集。

在使用 LIBSVM 时，svm_problem 的 x 属性是一个二维的 svm_node 数组。将其第一维大小设置为训练数据集 T 的元素个数，第二维设置为训练数据集 T 中训练文本特征向量的维数。训练数据集 T 中的每一个元素对应着 x 中的一行。对于 svm_problem 的 x 属性中的第 i 行 j 列元素 x[i][j]，设置其 index 属性为 j+1，同时设置其 value 属性为训练数据集合中第 i 个训练文本的特征向量的第 j 维数值。

svm_problem 的 y 属性是一个一维数组，其大小为训练数据集 T 中的元素个数。对于 y 的第 i 维，设置其值为训练数据集 T 中的第 i 个训练文本的类别 ci。

3) 训练 SVM 分类器模型

在 LIBSVM 软件包中，调用 svm 的静态 svm_train 方法便可以完成 SVM 分类器的训练工作。该方法使用 svm_problem 和 svm_parameter 作为参数，这两个参数在前面的步骤中均已经设置完成。svm_train 方法的返回值为 svm_model 类型的对象，该对象即为 SVM 分类器模型。

4）短信分类

通过上述步骤就完成了 SVM 分类器的构造任务，接下来开始短信文本分类。在对未知文本分类之前，需要将文本 d 按照 VSM 模型表示为其特征向量：

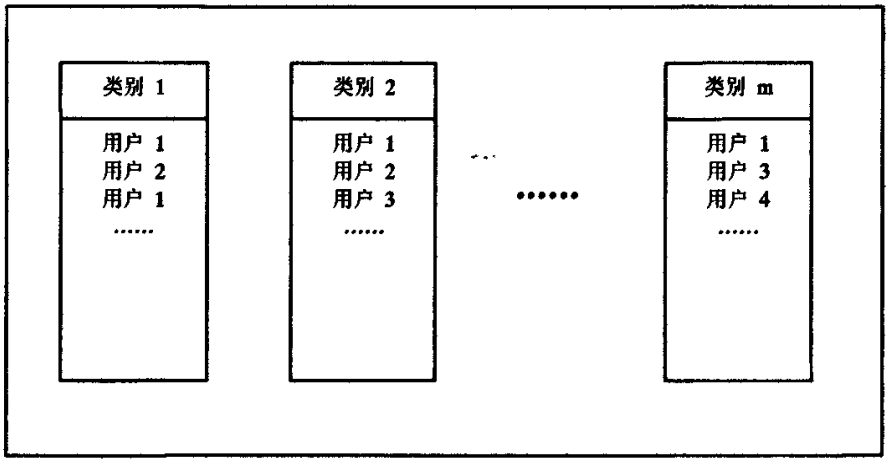
$$w_d = (w_{d1}, w_{d2}, \cdots, w_{dn})$$

LIBSVM 软件包提供了利用 SVM 分类器模型来预测未知文本类别的功能。对

于未知文本 d 的特征向量 wd，按照训练数据集中的训练数据的录入方式录入，只是不设置其对应 svm_node 的 value 属性。

录入待预测的文本的特征向量后，调用 svm 的静态 svm_predict 方法，便可以完成预测工作了。该方法使用 svm_model 和 svm_node 数组作为参数。svm_model 为步骤 1 中生成的 SVM 分类器模型，svm_node 数组则对应着待预测类别的文本的录入数据。svm_predict 方法将返回通过 svm_model 预测的文本的类别。

上述步骤将短信文本分类后，每一条短信文本就归属于某个特定的类别。在实现时通过事先建立类别文件，如果判断出某条短信的类别，则将该短信文本中的发送者号码记入分类类别中。类别文件如下所示：



4.2.3 分类器的评价指标

1.评价指标选取

本文中借用文本分类的相关指标对于短信文本的分类性能进行评价，本文选用两个指标来评判分类结果：正确率和错误率[48]。我们假设待测试的短信集中共有 N 条短信，测试结果的判定如下表 4.1 所示：

表 4.1 短信分类系统判定情况分布(单位：条)

	实际为汽车类	实际为手机类
系统判定为汽车类	A	B
系统判定为手机类	C	D

其中，$N=A+B+C+D=N_q+N_s$。($N_q=A+C$ 为汽车类短信数目，$N_s=B+D$ 为手机类短信的数目)。

(1) 正确率：$R=\frac{A}{A+B}$，即短信类别分正确的比例。正确反映了分类系统判断短信类别的能力，正确率越大，正确类别的短信被误判的数量越少。

(2) 错误率：$Err=\frac{B+C}{N}$，即对所有短信文本(包括汽车和手机)的判错率。

第五章 短信文本分类系统 SVMCLS 模型的设计

作为一种新兴的数据处理技术，短信文本研究已经吸引了国内外众多的科研工作者。应该说，短信研究一开始就是作为一种实用性很强的技术出现的，也只有奖这种技术转化为通用的、可操作的工具才能使这项技术得到更大的发展。

在前面的章节中，我们讨论了文本挖掘的概念、过程、及其在短信分类中的应用，尤其重点讨论了短信文本预处理过程中的一些方法技术和短信文本的分类技术。本章结合上述方法和思想，从实际应用的角度出发，分析并设计一个短信分类系统模型(SVMCLS)[49]。

5.1 短信分类系统模型 SVMCLS 的分析与设计

整个系统的概念化设计如图 5.1 所示，它是基于内容的分类方法为主，多种方法协作式的一个短信文本分类系统。整个分类系统分为训练和分类两个主要步骤。系统的具体工作包括以下几个方面：

1) 短信文本的训练模块

用户首先抽取一定量的短信文本进行人工训练，以便建立分类特征向量库。在训练模块中，首先将训练用短信文本进行预处理，将短信表示成规范的、易于为计算机所处理的数据。通过分词处理构造短信文本的初始向量，然后使用特征向量抽取算法以获得短信文本的特征向量，并将这些短信文本特征向量加入特征向量库中。

2) 短信文本的分类模块

在分类模块中，首先对待分类短信文本进行文本整合处理，然后再进行标准化预处理，得到其特征向量，然后进行分类，最后得到短信文本的类别。

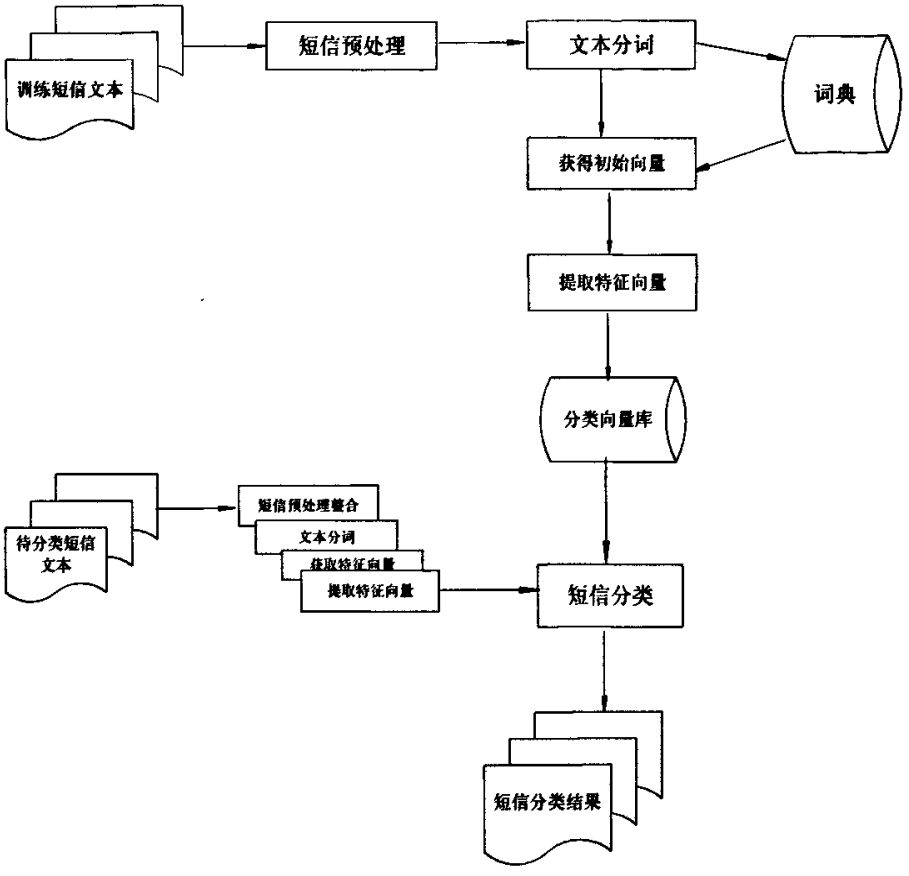


图 5.1 SVMCLS 系统模型

5.2 SVMCLS 分类系统模型中关键问题的处理

在前面的章节中，本文已详细介绍了一些关键技术的设计和实现方法，为了清楚起见，在本节，再把系统中设计的关键技术进行以下总结。

1. 短信文本的表示

在基于内容的短信文本分类器中，由于短信主要包含的是文本信息，系统中选择向量空间模型作为短信文本的表示方法[50]。

2. 分词的处理

采用分级词典，分级的依据是频度。同时，为了提高效率，建立索引文件方式，采用的查找方法是二分查找法。在存储上，把分词词典存储于扩展内存中，提高处理速度[51]。

3. 词典设置

为解决 VSM 模型中要求特征词条相互独立与自然语言多样性之间的矛盾，本设计在模型中建立了四个词典：中心词词典、同义词词典、近义词词典和蕴含词词典用于词频统计，其中主词典中的词条要求在含义上保持尽可能的独立。进行词频统计和特征提取时，以主词典中的主词条为表示词条进行处理[52]。

4. 权重计算

在计算词的权重时，对具有标志性的词，如“手机”、“汽车”、“住房”等，给与了较大的权重。而且，根据实际需要，我们可以人为地加大某些关键词的特征[53]。

5. 特征向量维数的选取

选出信息度最高的有限个特征项，便生成特征向量，这样可以大大提高系统的处理速度

5.3 系统中的主要算法

5.3.1. 特征向量的获取算法

基于内容的短信分类模型的一个重要问题就是反映了短信文本内容的特征向量的获取，也好似进行基于内容短信处理的最为关键的步骤[54]。系统中提出采用统计方法关键词集来提取的办法。其实现前提是短信文本已经过词的切分，变成了一个词的序列，提取的步骤如下：

算法 5.1 文本特征向量提取算法

1) 使用奇异符号表过滤，将文本中的各种对文本特征区分无多大作用的奇异符号去掉；然后进行中文分词处理，通过奇异词表进行奇异词规范化处理，消除文本中的奇异词；通过预加窗的文本校对技术去除到文本中的错字和同音异形字；同时也要消除停用词、稀有词等。

2) 对文本从头开始逐词顺序往下扫描，并按以下方法进行统计：每个词在其第一次出现时设一个相应的计数器，并置成 1，此后该词没出现一次就在其相应的计数器中加 1。

3) 归一化：将所有次的计数器积分相加得到和数 S，然后每个计数器的积分除以 S 再放入计数器。

4) 特征词选择：设定阈限 λ（一个选定在 $[0,1]$ 区间中的小数），进行“λ—过

滤操作”，即把该模糊关键字集中的隶属度小于 λ ($0 < \lambda \leq 1$) 的关键词滤掉，仅选取那些计数器的计分大于等于 λ 的词作为关键词。采用上述手段可把不够重要的关键词忽略掉，而最终得到一个可以近似描述原文语义的“模糊关键词集”。这里需要指出不同的场合选用多大的 λ 值来进行过滤要根据实际情况而定，不能一概而论。当然，为了比较两个文本的关联程度，应该喜爱用相同的产生“模糊关键词集”的方法和相同的过滤计数 λ。

5) 利用提取出的特征向量对训练文本进行分类，如果精确度满足要求则特征向量提取结束，将上一步骤中提取的特征向量作为分类向量，文本特征向量提取结束；如果精确度不满足要求则返回第 4 步骤，对提取出的特征向量进行再次提取如此反复迭代知道满足结束条件。

5.3.2. 训练模块的流程

训练阶段流程如图 5.2 所示：

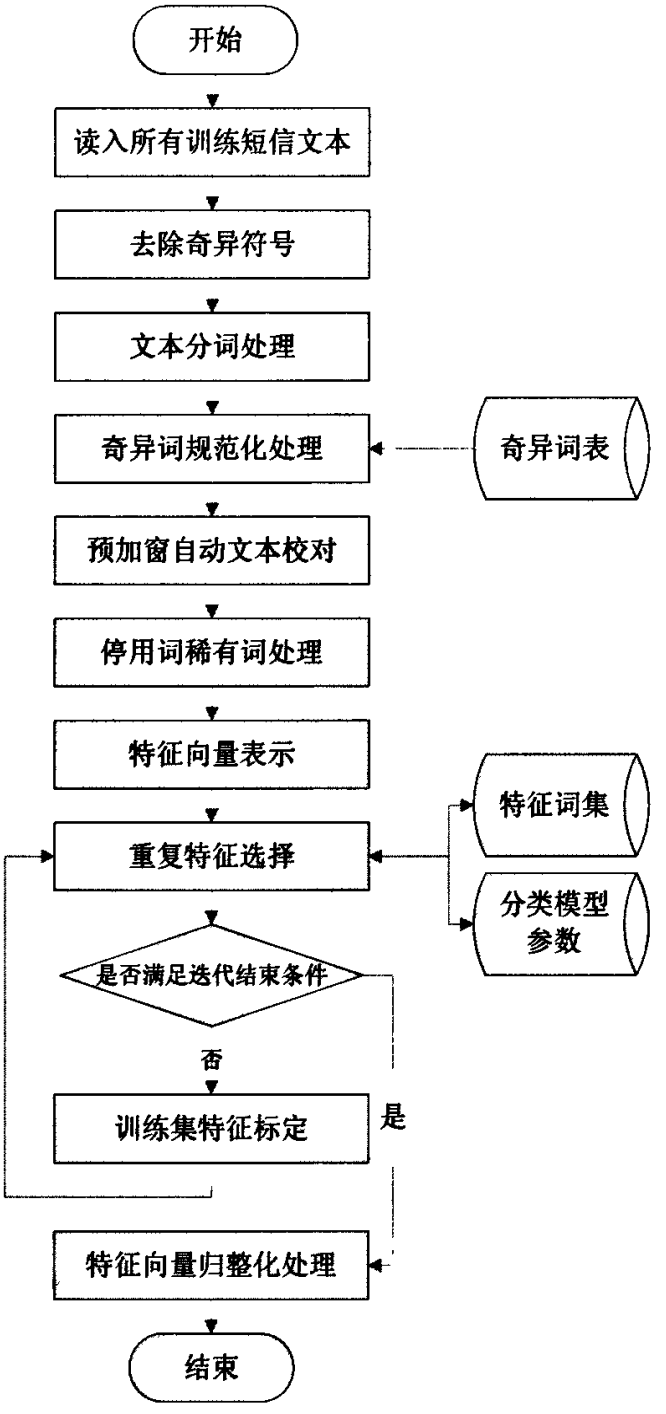


图 5.2 分类系统训练阶段流程图

5.4 实验

1. 实验中一些系统参数的设置

1) 特征向量的选取

在对短信文本进行了分词处理后，实验模型中只选取名词、动词和形容词作为初始向量，对于短信文本的特征向量则从这些初始向量中抽取。

2) 性能评价

在实验中我们找了四类模拟短信文本来进行测试，第一类测试类型为环境，其中训练样本数为 705，测试样本数为 69；第二类测试类型为军事，其中训练样本数为 741，测试样本数为 54；第三类为体育，其中训练样本数为 904，测试样本数为 63；第四类为医疗，其中训练样本数为 633，测试样本数为 42。

表 5.1 传统算法

实验编号	样本数量	正确识别数	正确率	错误识别数	错误率
1. 环境	69	56	0.8116	13	0.1884
2. 军事	54	46	0.8519	8	0.1481
3. 体育	63	60	0.9524	3	0.0476
4. 医疗	42	28	0.6667	14	0.3333

表 5.2 改进型算法

实验编号	样本数量	正确识别数	正确率	错误识别数	错误率
1. 环境	69	66	0.9565	3	0.0435
2. 军事	54	49	0.9074	5	0.0926
3. 体育	63	61	0.9683	2	0.0317
4. 医疗	42	32	0.7619	10	0.2381

通过表 5.1 和表 5.2 中数据的比较，我们可以看到，本文中提出的方法与传统方法相比在性能上有了较大进步。由此可以看到，本文中提出的方法是有效的。

第六章 总结和展望

6.1 总结

短信已成为人们生活中必不可少的一个工具，她给人们的生活带来了极大方便。然而，作为其发展的副产品——垃圾短信，却给广大用户和管理者带来了巨大的麻烦，在时间上，经济上造成巨大的损失，也带了一定的社会问题。如何帮助人们有效的反垃圾短信，营造一个健康、和谐、有序的环境，已成为一个新的研究热点。

本文主要针对垃圾广告短信的处理做了较为深刻的研究和较为详细的论述，对一些关键技术进行了深入的研究，针对其中的一些难点提出了适宜的解决方法，有一定的理论意义和使用价值。总结全文的研究工作，主要包括以下几点：

1. 将文本挖掘的分类技术和方法引入到短信文本信息的处理领域，实现了对短信的文本的有效分类。并结合当前一些文本分类算法，对短信文本分类方法和分类策略进行了研究，提出了在短信进行分类之前，先进行有效的整合的思想，进一步提高了分类的效率，改善了分类的精度。

2. 概要讨论了文本预处理工作的重要性和必要性，并在给出文本预处理逻辑框架的基础上，探讨非结构化、半结构化文本转化结构化数据的数据预处理方法，包括文本特征格式分析及规范化、中文分词处理、对错字和同音异型字的校对、短信文本特征选择、基于兴趣关键特征词表的特征权重调整、特征规范化处理等，并对各种方法进行了简单评述。特别是，文中提出的一种自适应的重复特征选择技术，该方法能够最终得到最优的低维特征空间，同时也有效的解决了训练集含有噪声训练元素情况下的最优特征提取问题。

3. 提出了一种预加窗的中文文本校对技术，用于文本规范和校对，同时该算法具有较小的计算复杂度。

4. 针对广告本身的特点，提出了一种基于关键词表的特征权重调整技术，进一步地突出了短信中的关键词成分，提高了关键点的识别率。

5. 根据软件工程思想和模块化设计方法，设计并实现了一个短信分类原型系统——SVMCLS，系统具有很好的可用性、可扩展性。

6.2 展望

短信文本分类作为一门新兴的数据处理技术，其发展非常迅速，吸引了众多的科研人员。本文也对该领域的研究做了一些贡献。但是，它毕竟是一门新技术，有很多不完善的地方，还有许多值得进一步深入研究的问题。

1. 本文中我们主要讨论了关于广告垃圾短信的处理，还有很多其他的非广告垃圾短信，本研究中并没有提到，但是这些垃圾短信确确实实也是需要被处理的，所以关于垃圾短信这方面需要研究的问题还很多。

2. 在短信分类器实现的过程中需要配置大量的系统参数，一般需要专业人士的手工调节。如何实现参数配置的自动化，增加系统的易用性，也是一个值得进一步研究的问题。

3. 在短信文本处理过程中涉及到数据预处理、特征抽取等多个方面，而每个环节都有若干实现方法，系统地性能与过程的实现方法及方法的组合方式息息相关。采用何种软件工程技术，使得整个系统易于理解、易于构造、易于优化、易于维护及易于扩展也是一个需要讨论的问题。

此外，在论文的完成过程中，我们发现对算法进行功能、性能方面的测试非常困难，需要编写需要辅助性的程序，费时费力。而将算法与其他已有的算法进行比较时，常常由于测试环境、实现方式不同等原因，使得同一算法可能得出不太相同的结论。因此，构造一个标准的测试平台是非常有意义的工作。同时，虽然国内外的一些科研单位已经着手进行中文标准文本测试集的标注和整理工作，但与实际应用的需要还有一定的差距。

总之，短信文本处理技术虽然刚刚起步，但作为实际应用的一门新兴的数据处理技术，已经引起了人们的广泛关注。这项随着垃圾短信产生而出现的技术，必将随着信息时代的发展而日臻完善，有效的阻止垃圾短信的发展和泛滥，将垃圾短信的危害降到最低。

参考文献

[1] 太平洋电脑网. 通信频道. http://www.pconline.com.cn/comm/sp/0703/979332.html 2007.3

[2] 搜狐网. IT 频道. http://www.pconline.com.cn/comm/sp/0703/979332.html 2005.8

[3] 易易域. 咨询频道. http://www.yiyicheng.com/news/newsck/2007-2/zx_v32156.html 2007.2

[4] qq 网. 教育频道. http://edu.qq.com/a/20061110/000244.htm 2006.11

[5] 石家庄新闻网. 燕赵晚报, http://www.sjzdaily.com.cn/yanzhao/2006-12/08/content_1009280.htm 2006.12

[6] 太平洋电脑网. 通信频道. http://arch.pconline.com.cn/comm/sp/0611/911264.html 2006.11

[7] 吴立德. 大规模中文文本处理[]. 上海:复旦大学出版社, 1997.23-24.

[8] Kunnger S, Spafrord E H.A pattern matching model for misuse intrusion detection. In:Proceeding of the 17th National Computer Security Conference. U.S.A., 1994.11-21

[9] 季矩, 罗振声. 基于概念统计和语义层次分析的英文自动文摘研究[J]. 中文信息学报, 2003, (2):14-20

[10] Li Y H, Jain A K. Classification of text document. The Computer Journal,1998, 41(8):537-546

[11] Tao Liu, Shengping Liu, Zheng Chen, et al.An evaluation on feature selection for text clustering. In: Proceeding of the 20th International Conference on Machine Learning(ICML-03), 2003, 488-495

[12] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977).Maximum likelihood from incomplete data via the EMalgorithm. Journal of the Royal Stat. Society,39,1-38.

[13] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An Evaluation on Feature Selection for Text Clustering," Proc. Int'l Conf.Machine Learning (ICML'03), pp. 488-495, 2003.

[14] 傅祖芸. 信息论基础[M]. 北京:电子工业出版社, 1989

[15] K. Daphne and M. Sahami, "Toward Optimal Feature Selection,"Proc. 13th Int'l Conf. Machine Learning, pp. 284-292, 1996

[16] 李建华, 王晓龙, 等. 多特征的中文文本校对算法的研究[J]. 计算机工程与科学

2001, 23(3):93-96.

[17] Jurafsky D., Martin J. H. 自然语言处理综论[M]，北京：电子工业出版社，2005, 472-473

[18] Agrawal R, Srikant R. Fast algorithm for mining association rules in large databases[A]. The International Conference on Very Large Data Bases[C]. 1994, 487-499

[19] 于津凯等. 一种基于N-Gram改进的文本特征提取算法. 图书情报工作，2004，48(8)：48-50

[20] 林鸿飞、战学刚、姚天顺，中文文本挖掘的特征导航机制，东北大学学报(自然科学版)，2000年6月，P240-243

[21] 陆玉昌, 鲁明羽, 李凡, 等. 向量空间法中单词权重函数的分析和构造[J]. 计算机研究与发展, 2002, 39(10):1205-1210.

[22] Ruger S. M. Feature Reduction for Information Retrieval [A]. In Voorhees, E. M. and Harman, D. K., editors, The Seventh Text REtrieval Conference (TREC-7), pages 409—412 Gaithersburg, Maryland. Department of Commerce, National Institute of Standards and Technology (NIST) Svecial Publication 500-242.1998

[23] 汪维家，陈笑蓉，秦进等，一种基于窗口技术的中文文本自动校对方法，贵州大学学报（自然科学版）Vol. 20 No. 2, 2003

[24] 胥桂仙, 苏筱蔚, 陈淑艳. 中文文本挖掘的无词典分词的算法及其应用[J]. 吉林工学院学报, 2002, 23(1):16-18

[25] HAN Jiawei, KAMBER Micheline. Data Mining Concepts and Techniques[M]. Toronto: Morgan Kaufmann Publishers, 2001, 35-117

[26] 黄昌宁，中文信息处理的主流技术是什么[EB/OL]。CTI 论坛，http://www.ctiforun.com/ 2002

[27] Kang Bo-Yeong, Lee Sang-Jo. Document indexing: a concept-based approach to term weight estimation[J]. Information Processing and Management, 2005, 41: 1065-1080

[28] 王科，高常波，翟雪峰等. 汉语分词主要技术及其应用[J]. 通信技术，2003(6)：12-15

[29] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[30] Y. Yang and J.O. Pedersen, "A Comparative Study on FeatureSelection in Text Categorization," Proc. 14th Int'l Conf. MachineLearning, pp. 412-420, 1997.

[31] 张仰森, 丁冰青. 基于二元接续关系检查的字词级自动查错方法[J]. 中文信息学 2001, 15(3):36-43.

[32] Lewis Frey, Douglas Fisher. Identifying Markov Blankets with Decision Tree Induction [C]

. Proceedings of the Third IEEE international Conference on Data Mining（ICDM03），2003

[33] JelinekF.Statistical Methods for Speech Recognition[M] HongKong:AscoTrade TypesettingLtd, 1997.19-23.

[34] 李凡，鲁明羽，陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报(自然科学版), 2001, 41(7):98-101.

[35] K. Daphne and M. Sahami, "Toward Optimal Feature Selection,"Proc. 13th Int'l Conf. Machine Learning, pp. 284-292, 1996.

[36] A.L. Blum and P. Langley, "Selection of Relevant Features andExamples in Machine Learning," Artificial Intelligence, vol. 97,nos. 1-2, pp. 245-271, 1997

[37] Zhu Liu, an Efficient Algorithm for clustering short spoken utterances, Acoustics,Speech,and Signal Processing,2005,Vol.1 pp:593-596

[38] 王晓龙, 王开铸. 声音语句输入的研究[J]. 计算机学报, 1994, 17(2):96-103.

[39] Peter Pirolli, Patricia Schank, Marti Hearst, Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In proceedings of the ACM SIGCHI Congercece on Human Factors in Computing Systems, 1996, 213-220

[40] 李晓黎、刘继敏、史忠植，概念推理网及其在文本分类中的应用，计算机研究与发展，2000年9月，P1032-1038

[41] 刘明吉，知识发现方法研究[D]，南开大学博士学位论文. 2001

[42] 李晓黎、史忠植，用数据挖掘方法获取汉语此行标注规则[J]，计算机研究与发展，37(12):1409-1414. 2000

[43] Quinlan J R. Induction of decision trees[J]. Machine Learning，1986，14(1)：81-l06

[44] ZhangZhaohuang.APilot Study on Automation Chinese Spelling Error Correction[J]. Communications of COLIPS,1994,14(2):143-149.

[45] 马志毅，姚天顺，基于情境的文本理解[J]。计算机科学，25(3)：26-29. 1998

[46] ArampatzisA., vander Weide T.P.,KosterC.H.A.,van Bommel P.An Evaluation of Linguistically-motivated Indexing Schemes [A]. In Proceedings of the BCS-IRSG' 2000.2000

[47] ArampatzisA.,vander weide T.p.,Koster C.H.A.,van Bommel P.Term Selection for Filtering based on Distribution of Terms over Time[A].RIAO'2000 Conference Proceedings,Vol.2,April 12-14,Paris,France 1221-1237.2000

[48] Agrawal R., ImielinskiT., SwamiA. Mining association rules between sets of items in large database[A]. Proceedings of the ACM SIGMOD Conference on Management of

data:207-216,1993

[49] David Gilbert, Michael Schroeder. FURY: Fuzzy unification and resolution based on edit distance[A]. International Conference on Bioinformatics and Biomedical Egineering, IEEE. 2000

[50] Larker L.S., Croft W.B. Combining classifiers in text categorization[A]. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval,289-297,1996

[51] Yu Hwan Kim, Shang Yoon, Byoung Tak Xhang. Text filtering by boosting naïve Bayes classifiers[A]. Proceedings of the 23rd annual international ACM SIGIR conerence on Research and development in information retrieval, 168-175,2000

[52] Kok-wee Ganetal. A statistically emergent approach for language processing application tomodelling context effects in ambiguous Chinese word boundary perception Computational Linguistics 1996 22 (4)

[53] Ronen Feldman and Ido Dagan.KDT-Knowledge Discovery in Textual Databases. In Proceedings of the 1st Annual Comference on Knowledge Discovery and Data Mining, 1995. 112-117

[54] Joel Larocca Neto, Alexandre D. Santos, Celso A.A Kaestner and Alex A. Freitas. Document Clustering and Text Summarization. In Proc. Of the 4th Int.Conference on Practical Applications of Konwledge Discovery and Data Ming(PADD-2000), 2000. 41-55

致　谢

衷心感谢傅彦导师对我的精心栽培和悉心教导。在本文的撰写过程中，从选题、定题、修改、定稿，都离不开导师的悉心指导和耐心帮助，在此致以最诚挚的谢意。感谢导师对我无微不至的关怀和帮助。

感谢计算机科学与工程学院的各位老师对我多年的培养和学业上的指导，以及给予过我帮助的各位领导和老师。

我要特别感谢我的父母，他们对我的养育之恩和所倾注的无数心血让我终生难以回报；特别感谢帮助过我的亲人和朋友们，他们一直给予我极大的关心、支持和鼓励。

最后，衷心感谢各位评审专家以及所有关心和帮助过我的人们！谢谢你们！

个人简历、在学期间的研究成果及发表的学术论文

一、个人简历

出生年月：1981 年 9 月 24 日

2000 年 9 月－2004 年 7 月　四川大学数学学院，获得了理学学士学位

2004 年 9 月－2007 年 6 月　电子科技大学计算机科学与工程学院，计算机应用技术专业，攻读工学硕士学位

二、科研项目

1. 国家自然科学基金项目《基于神经网络的大规模科学数据分析》（基金号：10476006）

2. 四川省应用研究基金《各类大规模高维数据的智能数据挖掘理论和技术基础研究》（基金号：05JY029-067-2）

三、发表论文

牛海根，傅彦．一种有效的短信聚类算法，“电子科技大学研究生学报”录用并发表

四、获奖

获 2006-2007 年度电子科技大学优秀学生奖学金。

获 2006-2007 年度电子科技大学计算机学院优秀党员称号。

获 2006-2007 年度电子科技大学计算机学院优秀学生干部称号。