单位代码：10359
学　　号：2009005J015

分 类 号：TP319
密　　级：

合肥工業大學
Hefei University of Technology

硕士学位论文

MASTER DEGREE THESIS

论文题目：网络小说分类与推荐研究

学位类别：高 校 教 师

学科专业：
(工程领域)　计算机应用技术

作者姓名：李 春 秋

导师姓名：樊玉琦　讲师

完成时间：2013 年 5 月

网络小说分类与推荐研究

Internet Novel Classification and Recommendation

作　者　姓　名　李春秋

学　位　类　型　高校教师

学　科、专　业　计算机应用技术

研　究　方　向　数据挖掘

导 师 及 职 称　樊玉琦　讲师

2013 年 4 月

合 肥 工 业 大 学

本论文经答辩委员会全体委员审查，确认符合合肥工业大学硕士学位论文质量要求。

答辩委员会签名：（工作单位、职称）

主　席：

安徽省经济信息中心　　高级工程师

委　员：

中国电科第三十八所　　高级工程师

合肥工业大学计算机与信息学院　副教授

合肥工业大学计算机与信息学院　副教授

合肥工业大学计算机与信息学院　副教授

导　师：

合肥工业大学计算机与信息学院　　讲师

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志和致谢的地方外，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__合肥工业大学__或其他教育机构的学位或证书而使用过的材料。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。

学位论文作者签字：李秀秋　　签字日期：2013 年 6 月 03 日

学位论文版权使用授权书

本学位论文作者完全了解__合肥工业大学__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，允许论文被查阅或借阅。本人授权__合肥工业大学__可以将学位论文的全部或部分论文内容编入有关数据库进行检索，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适用本授权书）

学位论文作者签名：李秀秋　　导师签名：樊飞跃

签字日期：2013 年 6 月 03 日　　签字日期：2013 年 06 月 03 日

学位论文作者毕业后去向：

工作单位：安徽商贸职业技术学院　　电话：0553-5971071

通讯地址：安徽省芜湖市弋江区文昌西路 24 号　　邮编：241002

网络小说分类与推荐研究

摘　要

随着互联网技术的飞速发展,互联网上的信息呈现指数级增长，人们通过传统的搜索引擎越来越难以获得自己感兴趣的信息，个性化推荐系统就是在这种背景下产生的一种帮助用户解决信息过载问题的技术。

本论文采用了文本分类方法来对网络小说进行分类和推荐。文本分类是自然语言处理领域一项重要技术，它将文本集合中每篇文本划分到一个预先定义的类别之中。文本分类的关键技术之一是如何提取特征向量，本论文通过对比TF-IDF、信息增益、开方校验法等三种不同的特征抽取技术，以选择最适合网络小说文本分类应用的特征抽取技术。进而，本论文采用支持向量机(SVM)对小说文本进行分类，并通过分类器的组合来提高分类速度。

通过基于内容的推荐技术、基于协同过滤的推荐技术、和以上这两种推荐技术的融合，本论文构建了网络小说推荐系统框架。实验结果表明本系统可以很好的为用户推荐其可能感兴趣的网络小说。

关键词: 文本分类；特征提取；文本向量；推荐系统；支持向量机

Internet Novel Classification and Recommendation

Abstract

With the rapid development of Internet technology, information on the Internet grows exponentially. It is increasingly difficult to obtain the information people are interested in through traditional search engines. Personalized recommendation system is a technology to help users to solve the problem of information overload. In this thesis we deal with the problem of information overload by using text classification and recommendation technologies.

This thesis uses text classification for classification and recommendation of novels on the Internet. Text classification is an important technology in the field of natural language processing, and it divides the text collection into pre-defined categories. One of the key technologies for text categorization is how to extract feature vectors. By comparing three different feature extraction techniques, TF-IDF, information gain, and χ^2 statistics, this thesis selects the most suitable method for network novel text classification applications. This thesis uses the support vector machine (SVM) to classify the text of the novel, and improves the speed of classification classifier combination.

Based on content recommendation technique, collaborative filtering recommendation technique, and the technique combining the above two techniques, this thesis constructs a recommendation system architecture for novels on the Internet.

Keywords: Text Classification; Feature Extraction; Text vector; Recommendation System; SVM

致　谢

三年的硕士学习阶段转瞬即逝，我即将告别艰苦而又充实的求学过程，回顾这三年的时光，我遇到过很多挫折，困惑过，焦虑过，也有克服困难后的喜悦，在此期间我所取得的所有成绩都离不开老师、同学、朋友的关心和帮助，在此，我由衷的向他们表示感谢。

首先，我要特别感谢的是我的论文指导恩师：樊玉琦老师。在樊玉琦老师的无私帮助以及悉心指导与鼓励下，我的论文才能如此顺利完成。樊老师严谨的治学态度和兢兢业业的工作作风，使我受益匪浅，给我以后学习工作树立了榜样，樊老师时刻教导我写论文要踏踏实实、认认真真。樊老师的这种工作作风对我以后的工作大有裨益，对我以后的学习生活也有很大的帮助。樊老师的耐心教诲和鼓励，让我有勇气和信心去面对困难，挑战困难，战胜困难。

再次，我要感谢与我共同工作以及共同合作的同事，与他们一起研究讨论，给我很多启发和帮助，有了他们的帮助我的论文才能如此成功和高效的完成。

最后，我要感谢的是我的家人，他们是我坚实后盾，无论是工作上、学习上还是生活上，他们都给予我很大的帮助与支持。虽然我不能一一列举所有帮助过我的人，但在此我要由衷的对所有关心过我、帮助我的老师、同事、同学和朋友说声：谢谢，谢谢您们给予的帮助与支持。

作者：李春秋

2013 年 4 月

目　录

插图清单

表格清单

第一章 绪论

1.1 课题背景与意义

自从 web 浏览器出现后，经过短短十多年的发展， Internet 已经发展成为一个全球信息化的存储空间，尤其网络技术的迅速普及，使得 web 信息飞速增长，网络资源越来越丰富，互联网上的各种应用已经深入到我们生活的点点滴滴之中，但是互联网在给我们带来方便的同时，也引发了一系列的问题产生，人们在面对如此浩瀚的信息时，无从下手寻找自己感兴趣的信息。因此，如何有效的方便人们查找、处理这些信息，成为当务之急。

百度、google 等搜索引擎工具的出现在一定程度上给人们寻找资源提供了便利，但是人们所熟知的这些搜索引擎工具，依旧是被动搜索，需要用户或者访问者有一个明确的要求，或者用户可以大致描述所要搜寻的信息。如果用户没有明确的需求，面对互联网的海量信息时，使用这些搜索引擎查询时依旧存在信息过载的问题。如何帮助用户有效的搜寻和分析海量的 web 信息，使用户快速查找、定位到自己感兴趣的信息，成为当前的研究热点。文本分类与推荐系统成为解决上述问题的最好办法。文本分类不用人工干预，具有分类速度快，精准度高的优点；推荐系统可以根据用户的浏览历史，根据用户的兴趣爱好以及用户购买行为，向用户推荐用户可能感兴趣的信息和商品。从而帮助人们有效的解决了信息过载的问题。

互联网的飞速发展，大量的应用随之产生。其中一个广泛应用是：越来越多的用户通过互联网来阅读书籍，并且其中绝大部分用户是用来阅读小说类书籍。然而，互联网上的小说数以千万计，信息过载信息十分严重，用户很难通过传统的搜索引擎获取自己感兴趣的小说，使得用户越来越难以找到自己感兴趣的小说，因此迫切的需要一种能够帮助用户选择其感兴趣小说的技术，通过用户的历史记录等用户信息挖掘出用户的兴趣特征，根据用户的兴趣将其最可能感兴趣的小说主动的推送给用户，使得用户可以不通过主动的搜索动作就可以获取自己感兴趣的小说，缩短用户获得自己感兴趣小说的距离、提高用户的体验。本轮文就是基于这样的出发点，通过将网络信息抽取、文本分类、个性化推荐技术等融合起来构建一个网络小说分类与推荐系统来满足小说用户的个性化需求，使得用户可以方便、快捷的获得自己感兴趣的小说。

1.2 文本分类研究现状

随着互联网技术的迅速发展和普及，大量的文字信息开始存储在计算机中，并且其数量越来越大，人们已经从信息匮乏时代转变到信息过载的时代，面对如此浩如烟海的文献、资料和文本数据，如何对其进行分类、组织、管理是当

前研究的重点课题。文本自动分类简称文本分类（text categorization）是目前解决此类问题的很好方法。

文本分类是自然语言处理领域一项重要技术，它是指将文本集合中每篇文本划分到一个预先定义的类别之中，在互联网信息爆炸发展的今天，文本分类在数据挖掘、信息检索、数字化图书馆等领域扮演着越来越重要的角色[1]。文本分类从分类方式上可以分为人工分类、自动分类两种。人工分类一般情况下可以具有更高的准确率、可靠性，但是由于个人的背景知识限制、人力资源限制等因素，想通过人工的方式将海量信息进行处理就变得不可行。自动分类是只通过机器学习等方法，利用计算机代替人进行文本分类工作，如果我们能够有效的应用机器学习方法，不仅可以提高分类速度，同样可以保证分类的可靠性。

国外对于文本分类技术的研究早于国内。上个世纪 50 年代末，H. P. Luhn 在文本分类领域进行了开创性的工作，提出了基于词频统计思想的自动化文本分类方法。不久以后，Maron 发表了关于自动化文本分类的第一篇论文，随后 K. Spark，G. Salton 以及 K. S. Jones 等众多学者都在这一领域进行了卓有成效的研究工作[2]。目前，文本分类技术在国外已经从实验性阶段进入到了实用化阶段，已经在邮件分类(垃圾邮件过滤)，电子会议、数字化图书馆等方面取得了广泛的应用和价值[3,4,5]。

概况地说，文本分类在国外经历了如下几个发展阶段：

第一阶段（1958-1964）：主要进行自动分类的可行性研究；

第二阶段（1965-1974）：进行自动分类的可行性研究；

第三阶段（1975-1989）：进入实用化阶段；

第四阶段（1990 年至今）：面向互联网的文本自动分类研究阶段。

在国内，通过借鉴国外技术以及与中文汉语特点相结合，中文文本分类技术也得到了快速的发展。国内文本分类技术的发展方向分为基于外延的分类方法和基于概念的分类方法两种方向，国内的高校和研究所也对中文文本分类进行了深入的研究，主要包括北京大学、清华大学、浙江大学、东北大学、哈尔滨工业大学、复旦大学、中科院计算所等单位[6,7,8,9]。由于中文与英文在语境、语法上存在较大的差异性，所以我们不能完全照搬国外的研究成果和工程技术，中文文本分类的研究过程中充分的考虑了中文文本自身的特点，最终形成中文文本分类研究体系。首先是侯汉清教授在 1981 年对计算机在文本分类工作中应用作了探讨和阐述[10]。此后，中文文本分类系统如雨后春笋般的发展了起来，其中有具有代表性的有上海交通大学研制的基于神经网络算法的中文自动分类系统，清华大学的自动分类系统等等，同时在不同的分类算法方面也展开了广泛的研究和实现，中科院计算所的李晓黎、史忠植等人，中国科技大学的范众等人，复旦大学和富士通研究中心的黄营著、吴立德等人，上海交通大学的刁倩、王永成等人，都在文本分类算法方面取得了突出的成就[11,12,13]。

汉语分词是中文处理的基础工作之一，它的性能直接决定后续工作的质量。同样，中文文本分类的效果也依赖于汉语分词技术。自上个世纪 80 年代以来，众多的专家和学者都为提高汉语分词质量不懈的努力着，在分词算法层面上可以分为 3 类：机械分词、基于理解的分词和基于统计的分词，目前来讲效果最好的是基于统计的分词、并融合一些规则处理的分词系统，在工程层面上也涌现了许多成功的汉语分词系统，比较著名的有北京航空航天大学研制的 CDWS 和 CWSS 分词系统，清华大学黄昌宁、马晏等开发的 SEG 系统，东北大学姚天顺建立的基于规则的汉语分词系统，南京大学王启祥等人实现的 WSBN 分词系统，中科院计算所研制出的汉语词法分析系统 ICTCLAS 等等[14,15,16,17,18]。文本分类另外一个重要方面是如何选择那些对分类有帮助的特征，即特征选择。目前用于特征选取的统计量有文档频度、特征频度、信息增益、特征熵、互信息、开方校验法、统计量、文本证据权、期望交叉熵、低损降维方法、Bayes 准则法等一些列特征选择方法，各种方法并无绝对的优劣，需要在特定的应用场景选择合适的特征选择方法[19,20,21,22,23]。

目前，常用的文本分类算法主要包括以下几种：基于 VSM 的向量距离法、基于贝叶斯分类算法、基于 KNN 分类算法、基于 SVM 分类算法、基于决策树分类算法和基于神经网络分类算法、以及利用这些算法构建多个分类器混合使用等等策略都取得了不错的效果[24,25,26,27,28]。

1.3 推荐系统研究现状

在信息化时代，互联网在满足用户对各种信息的不同需求的同时，也使用户经常会迷失在海量的互联网信息中，较为困难甚至无法获得自己感兴趣的知识资源，即著名的信息过载问题。个性化推荐系统就是在这种背景下产生的一种帮助用户解决信息过载问题的技术[29]。它是通过挖掘用户的兴趣爱好和购买行为，进而推荐能够满足用户兴趣爱好的信息资源。推荐系统与搜索引擎的主要区别是：推荐系统中用户不需要主动的操作就可以获得满足自身兴趣的资源，而搜索引擎是用户主动的获取信息资源，显然推荐系统更能够满足用户的需求并有着更好的应用体验。推荐系统已经在许多领域都得到了广泛的应用，例如电子商务、个性化图书馆等[30,31,32]。

1995 年 3 月，卡耐基.梅隆大学的 RobertArmstrong 等人在美国人工智能协会上提出了个性化导航系统 Web Watcher，斯坦福大学的 MarkoBalabanovic 等人在同一会议上首次推出了个性化推荐系统 LIRA，在该研究领域，我国的专家学者对个性化推荐系统方面的研究工作也越来越重视，如周涛等人提出了一个基于网络结构的个性化推荐系统，随着互联网技术的日益成熟和发展，出现了许多商业化的个性化推荐系统，如世界最大的网上购物市场 EBay、最大的网上书店 Amazon、电影推荐系统 Movielense 及百分点科技等等[33,34]。

推荐系统从算法层面主要分为基于内容相似度的推荐、基于协同过滤的推荐技术、基于用户统计信息的推荐、基于效用的推荐、基于关联规则的推荐以及基于组合的推荐等等。但是目前最常用的推荐技术是前两种推荐技术：基于内容相似度的推荐、基于协同过滤的推荐技术，这两种算法有着各自的优势和不足，很多学者和公司都根据应用场景选择合适的算法或者融合多种算法来进行推荐，以其获得较高的推荐效果，推荐系统是目前互联网上最有前景、最火的应用技术[35,36,37,38,39,40]。

1.4 网络小说分类与推荐研究现状

在前两小节当中本论文分别描述了文本分类和推荐技术的国内外研究现状，国内外众多学者都在这两方面做了很多的研究工作并且应用到工业界当中。但是针对网络小说的文本分类与推荐技术尚为空白，主要是受网络小说兴起的时间较短、没有现成的小说数据集、难以获取用户数据等方面客观原因的限制，网络小说的分类与推荐技术研究变得相对更加繁琐，不过随着互联网的普及，许多文学网站如雨后春笋般出现，从而带动网络小说的迅速普及，在线阅读人数越来越多，客户需求也越来越大，如何方便在线读者这方面的工作就显得越来越迫切。因此，针对网络小说的文本分类和推荐技术研究就迫在眉睫。

1.5 本文主要内容

利用在互联网阅读小说的需求随着互联网技术的发展越来越多日益增加的形势下，本轮文主要针对网络上的小说数据，试图寻求为用户提供更便捷、更高效获得感兴趣的小说的方法。由于互联网上小说数量巨大、网页结构也不一致，本论文首先构建一个精确信息采集器，使用网络爬虫抓取互联网上小说数据并转成结构化数据进行存储到 mysql 中。由于小说的种类繁多，本论文需要按照类别对小说进行归类，因此本论文研究了自动化小说分类算法将网络上的小说根据分类打上类别标签。本论文最终目的是为用户提供其感兴趣的小说，因此本论文构建了融合内容相似度与协同过滤算法的小说推荐系统为用-户推荐其可能感兴趣的小说。本文阐述了多个算法，并且对每个算法都做了实验并评价，获得最适合小说场景的算法。

1.6 本文组织结构

本轮文利用网路上的小说数据为基础，利用特征提取和支持向量机的相关理论和方法，构建了融合内容相似度与协同过滤算法的小说推荐系统，解决了用户想阅读感兴趣小说的问题。本轮文组织结构如下：

第一章：绪论。介绍了课题研究的背景和意义，文本分类研究现状，推荐系统研究现状以及网络小说研究现状，阐明了研究的意义以及本文主要内容和组织结构。

第二章：推荐系统系统架构设计。阐明了网络小说推荐系统的系统架构。

第三章：基于规则的精准信息采集器与文本特征提取技术。介绍了如何采集网络小说数据，详细介绍了文本特征提取技术相关概念、功能、方法等，为下一章文本分类奠定理论基础。

第四章：基于支持向量机网络小说分类技术。详细介绍了支持向量机的相关概念、功能、方法等。并利用特征提取技术提取文本向量，结合支持向量机分类技术研究了自动化小说分类算法并进行相应的实验验证。

第五章：融合用户兴趣与文本内容的小说推荐技术。根据分类后的网络小说数据并融合基于内容和基于协同过滤的推荐技术向用户推荐可能感兴趣的小说，并进行实验评价，最终得到了最适合网络小说的推荐系统。

第六章：总结与展望。对课题的总结与展望。

第二章 网络小说推荐系统架构设计

2.1 网络小说推荐系统

本轮文根据目前网络小说的文本分类与推荐技术的空白，而在线小说阅读人数越来越多，客户需求也越来越大的情况，设计了一个融合内容相似度与协同过滤算法的小说推荐系统为用-户推荐其可能感兴趣的小说。

本轮文主要针对网络上的小说数据，试图寻求为用户提供更便捷、更高效获得感兴趣的小说的方法。由于互联网上小说数量巨大、网页结构也不一致，本论文首先构建一个精确信息采集器，使用网络爬虫抓取互联网上小说数据并转成结构化数据进行存储到 mysql 中。由于小说的种类繁多，本论文需要按照类别对小说进行归类，因此本论文研究了自动化小说分类算法将网络上的小说根据分类打上类别标签。本论文最终目的是为用户提供其感兴趣的小说，因此本论文构建了融合内容相似度与协同过滤算法的小说推荐系统为用-户推荐其可能感兴趣的小说。本轮文阐述了多个算法，并且对每个算法都做了实验并评价，获得最适合小说场景的算法。

2.2 推荐系统的架构设计

2.2.1 推荐系统客户端设计

客户端采用 js 以及 jquery 来实现页面与服务器的通讯工作，采用 css 与 html 来实现网页样式设计工作利用。客户端最主要的工作如何与服务器进行通信以及通信包的解析工作。客户端与服务器的通讯采用 http 协议，http 协议是一种面向连接的、可靠的通信方式，客户端将通信数据打包以后发往服务器端，同时等待服务器端的数据返回，当服务端返回后将数据进行解析并展示给用户。这里一个关键问题是如何容错异常情况的发生，即当网络发生异常的情况下，如何友好的处理该次请求操作，客户端通过将超时时间设定为 3 秒来解决，即当客户端发出请求 3 秒以后仍然没有收到返回的数据信息，那么客户端主动关闭该链接，从而应对由于服务器端异常而导致的没有结果返回等情况。

客户端与服务器端通信数据的组织方式采用 json 格式，json 是一种轻量级的数据协议格式，特点是数据信息所占空间较小，解析方便快速，跨平台，因此特别适合作为客户端与服务器端信息通信的格式，本论文将用户请求打包成 json 格式发往服务器端，服务器(采用.NET 实现)利用 jsoncpp 开源工具将信息解析、计算后，将回包数据重新组织成 json 形式发给客户端，客户端对回包进行处理后展示给用户。

2.2.2 推荐系统后台设计

推荐系统的后台主要包括两方面的工作：接入模块、推荐逻辑模块。

（一）接入模块

接入层实现的功能是接收客户端的请求并将数据原封不动的发往服务器端，同时接收来自服务器端的返回结果并将该结果返回给客户端。在实际应用中，接入层服务分布在多台机器上，这样当某台机器发生当机时，也不会影响用户的使用。接入层服务在接受到请求时，将该请求随机发送到后端的某台服务器上，若后端的某台服务器发生当机，那么由于没有心跳包返回，接入层服务会在自己的转发列表中将该机器的 IP 删除掉，从而保证不会往发生当机的后端服务器上发送请求，进而不会影响用户的使用体验。接入层实现了客户端与服务器端数据交互的转发工作，同时规避了当机等风险。

（二）推荐逻辑模块

服务器逻辑是本论文的小说分类与推荐技术的算法所在模块，本论文后文中所描述的算法都是在这一模块中实现的。在实际工作中，本论文将服务器逻辑布置到多台机器上，从而减轻每台机器的负担同时应对宕机等意外所产生的风险。服务器逻辑采用多线程方式，每当一个请求过来时，主逻辑创建一个线程为该请求进行服务，本论文通过维护一个线程池来实现该工作，采用多线程的好处是，线程的创建和回收速度快，线程间可以共享数据，特别适合小说的分类与推荐这种使用静态数据资源较多、逻辑运算较重的应用场景。

推荐系统设计可能要用到以下技术：小说数据持久化以及缓存实现。本文在下一节中对小说数据持久化和缓存实现做了详细描述。

2.3 小说数据持久化以及缓存实现

2.3.1 小说数据持久化

通过下文对推荐系统的分析，推荐系统在推荐过程中要么采用基于内容的推荐技术、要么采用基于协同过滤的推荐技术，但是无论采用哪种推荐技术都需要大量的计算过程，在实际应用中，如果每个用户到来时都进行相应的计算，显然系统无法支撑如此大的开销，就会产生雪崩现象，因此我们尽量在线下为每个用户计算好推荐结果，当用户到来时直接给出推荐结果，这里包含了两个问题：一是如果一个新用户到来时，我们怎样快速的为其找到推荐结果；二是如果一个已经留下一些使用记录的用户到来时，我们怎样快速的为其找到推荐结果。

通过第五章的分析可知，当一个新用户到来时，我们可以利用基于内容相似度的推荐技术为其进行推荐，实质就是在小说候选集合中找到与当前用户浏览的小说最相似的 N 本小说，将这些小说推荐给用户。显然，我们是可以获得小说集合的，我们只需要在线下根据分类器分类结果，然后计算每本小说最相似的 N 本小说并存储起来，然后在推荐时直接查询存储表就可以快速的获得结

果。这个存储过程就称为持久化过程，下文是对持久化采用的策略进行详述。

当一个老用户到来时，我们需要利用基于协同过滤的推荐技术为其进行推荐，实质就是根据与当前用户最相似的 N 个用户的行为习惯为其找到最可能喜欢的 N 本小说，并将这些小说推荐给用户。显然，我们可以获得用户的集合，并且可以获得用户之间的喜好偏向，进而根据系统过滤算法为每个用户计算出其最可能喜欢的一些小说，本论文只需要将这些小说存储起来，然后在一个老用户到来时，通过查询存储表就可以获得结果。

不难发现，无论对于老用户还是新用户，关键问题转移为：一是什么时候计算出上文提到的“表”；二是这个“表”如何进行持久化。下面就将解决这两个问题。

对于新用户来讲，这个“表”是小说到与其最相近的 N 本小说的映射关系，由于小说文本不具有时效性，即对固定的小说集合来说，任何一本小说和与其最相近的 N 本小说的映射关系是一成不变的，但是当有新的小说到来时，这个映射关系就会发生改变，因此我们需要在新的小说加入时重新计算这个映射关系。那么什么时候会有新的小说加入呢？如第二章所叙述，当爬虫抽取到新的小说时，就说明我们此时应该重新计算这个映射关系了。因此我们在每天运行完爬虫时，就会根据是否有新的小说进入来触发重新根据内容计算相似度的程序模块，从而生成新的映射关系。对于老用户来讲，这个“表”是用户到与其最感兴趣的 N 本小说的映射关系，由于用户的兴趣具有漂移性，一旦用户的兴趣发生了漂移，我们就需要重新计算这个关系，那么我们如何来判断用户的兴趣是否发生了漂移呢？如果一旦用户浏览了新的小说，就说明用户的兴趣有可能发生了变化，因此我们在每次用户的浏览行为数据发生了变化时，重新计算这个新的映射关系，因此我们在每天运行完用户行为抽取程序时，就会对用户行为发生变化的用户重新计算这个映射关系。

接下来的主要问题就是如何存储这个映射关系，显然无论是小说到与其最相似的 N 本小说，还是用户到与其最感兴趣的 N 本小说都是典型的 Key-Value 格式，对于前者 Key 是小说 ID，Value 是与其最相似的 N 本小说的 ID；对于后者 Key 是用户的 ID，Value 是与其最感兴趣的 N 本小说的 ID。因此，本论文采用 Google 发布的开源数据存储形式 Protobuf 来存放这种数据结果，protobuf 最擅长的序列化的存储数据到磁盘，并通过反序列化加载到内存，并且用户只需要定义结果，protobuf 编译工具可以为本论文自动生成对应的代码，并且支持 c++、java、python 等三种语言环境，因此特别适合为本论文做这个存储工作。本论文就是基于 protobuf 实现小说数据的持久化工作。

2.3.2 缓存实现

上一节中本论文实现了小说数据的持久化，利用磁盘中存储的数据，本论

文就可以实现这个推荐系统了。但是在实践中，本论文一定会遇到这样一个问题：小说、用户都是海量数据，内存的空间无法承载所有的数据，然而，如果本论文每次用户请求都查询磁盘的话，显然是无法进行实际应用的。因此，本论文采用了缓存的思想来帮我们完成这样的工作，即我们把活跃用户的信息加载到内存，而活跃用户拥有更高的请求频率，当活跃用户的请求到来时，直接从内存将其感兴趣的数据返回；当一个陌生用户到来时，即内存中没有命中该用户，本论文从磁盘加载该用户的数据到内存并返回，根据局部性原理，这个用户下次操作一般会很快到来，当新的请求到来时，就会缓存命中。根据这种缓存的策略，可以极大的缓解海量数据下数据请求延迟的问题。

第三章　基于规则的精准信息采集器与文本特征提取技术

3.1 引言

本章将描述如何及时、精准的采集互联网中的小说数据，以及如何利用 mysql 数据库持久化这些数据。讲述了基于词频-反文档频率、信息增益法、开方校验法等三种特征选择方法，以便下一章中利用这些特征项来描述小说文本。本章的最后利用网络上的一些标签数据构建了原始训练数据集，并且通过对这部分原始训练数据集进行了人工标注从而得到本论文使用的训练数据集。

3.2 基于规则的精准信息采集器

3.2.1 精准信息采集器的实现

信息采集器又称网络蜘蛛(Web Spider)。网络蜘蛛是通过网页的链接地址来寻找网页，从互联网中某一个页面开始，读取该网页的内容并解析出其中的 url 链接，然后通过这些链接地址来抓取下一个网页，这样一直循环下去，直到把整个互联网上所有的网页都抓取下来。

信息采集器在抓取页面时一般有两种策略：广度优先和深度优先。广度优先是指信息采集器会首先抓取起始网页中链接的所有网页，然后再选择其中一个链接网页，继续抓取在此网页中链接的所有网页，通常利用队列来实现该策略。深度优先是指信息采集器会从起始页面开始，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接，通常利用栈来实现该策略。

信息采集器采集下来的只是页面的 html 源码，而本论文需要的是小说的名称、作者、评分、摘要、目录、正文等能结构化信息数据。因此本论文需要对采集下来的 html 源码做进一步的解析工作，找到本论文需要的有利用价值的字段。由于网页的结构是稳定的，因此本论文利用正则表达式根据页面标签解析出本论文需要的字段。至此，本论文通过在传统的网络爬虫基础上加载一个基于正则表达式的信息解析模块来实现精准信息采集器。

本论文采用 c++语言以及 boost 库的正则表达式库 regex 实现精准信息采集器。根据网页的 html 格式设计正则表达式，将所有正则表达式进行编号后存放在文件 regex.txt 中，当网页结构发生变化或者需要采集新的数据来源时，不需要重新编写代码，只需要更新或者增加相应的正则表达式到 regex.txt 中，由此实现了正则表达式与代码的逻辑分离，从而增加了程序的扩展性，并且使得程序维护起来更加容易。基于正则表达式的精准爬虫的缺点是当网站结构发生变化时会引起正则表达式失效，因此要求本论文需要在正则表达式发生失效时及时的发现问题，因此本论文会对每条正则表达式的性能进行监控，每当正则表

达式抽取的字段为空时就记载到错误日志当中并报警，从而克服了正则表达式的缺点，使得基于正则表达式的信息采集器可以稳定的运行。精准信息采集器的工作流程如下：

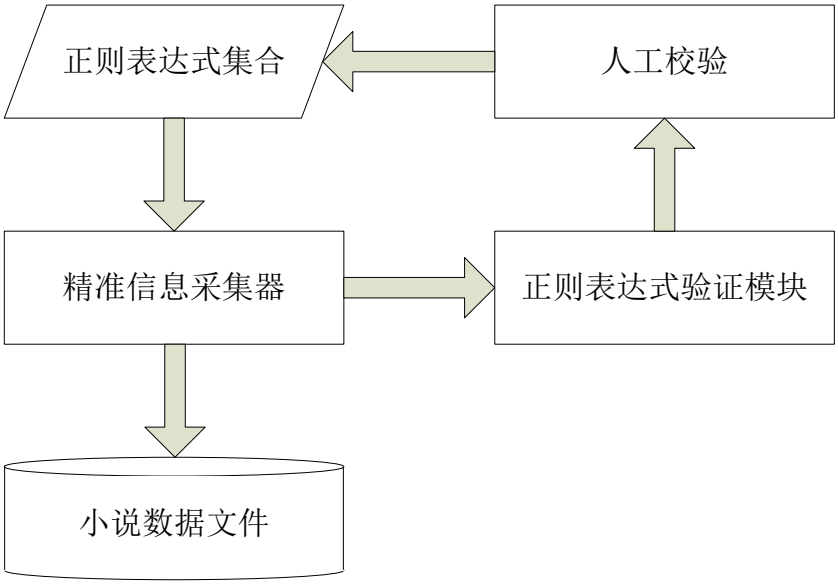

图 3.1 精准信息采集器

3.2.2 网络小说采集与存储

本论文使用 c++语言开发信息采集器，采集的小说网站包括起点中文网、小说阅读网、言情小说吧、起点女生网、红袖添香等一些小说网站，信息采集器在每天凌晨运行，根据页面结构利用正则表达式解析出本论文需要的字段。

以起点中文网为例，页面如下：

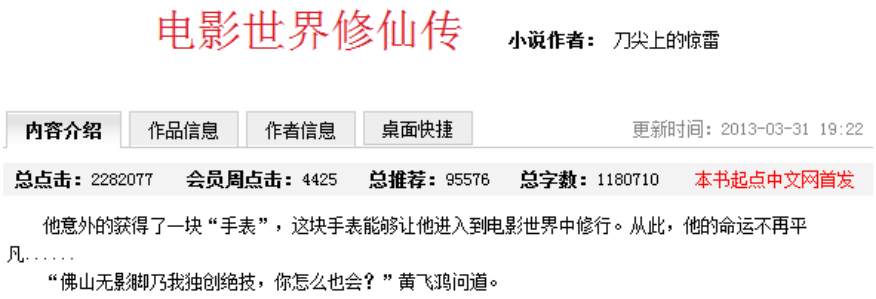

图 3.2 网络小说页面

该页面包含的本论文需要的字段有小说名，作者名，更新时间等信息。同理，本论文会通过其它页面得到我们需要的其它信息。正则表达式就是基于该页面的 html 源码，该页面的源码片段如下：

图 3.3 网络小说页面 html 源码

本论文根据 content 字段就可以得到小说名称“电影世界修仙传”、作者名称“刀尖上的惊雷”，同理，通过其他的 html 源码片段本论文就可以解析出本论文需要的字段信息。

本论文采集的所有字段详细描述如表 2.1。

表 3.1 网络小说页面字段提取

字段名称	字段描述
id	小说的唯一标识
name	小说的名称
user	小说的作者
public_time	小说发布的时间
summary	小说简介
catalogue	小说目录
content	小说正文内容
score	网友评分

本论文抽取出这些结构化字段并将其存储到 mysql 数据库当中。在 mysql 中本论文用 gb2312 编码来存储小说文本，因此在存入数据库之前对小说文本做相应的转换工作。至此本论文完成了小说页面的抓取与存储工作。

3.3 网络小说文本特征提取

3.3.1 基于词频-反文档频率的特征提取技术

词频-反文档频率的英文缩写是 TF-IDF。TF-IDF 是一种广泛应用于信息检索与资讯探勘的加权方法，已经得到工业界的实践验证，TF-IDF 加权的各种形式常被搜寻引擎应用，作为网页与用户查询之间相关性的度量或评价。通过 TF-IDF 可以评估一个字或一个词在语料库或语料库的某一份文件中的重要程度。TF-IDF 是基于这样的假设：字词的重要性与它在文件中出现的次数成正比增加，但却与它在语料库中出现的频率成反比下降。TF-IDF 实际上是 TF * IDF 的一种特定写法，其中 TF 称为词频(Term Frequency)、IDF 称为反文档频率(Inverse Document Frequency)。词频 (TF) 指的是某个特定的词在该页面中出现的次数，是一个局部的度量，在实际应用中，通常需要对 TF 做归一化，以期在

较长或较短页面间获得平衡。反文件频率 (IDF) 反应的是一个词在整个语料库中的重要性，是一个全局的度量。

TF-IDF 的计算方法，假如本论文有一个包含 10000 篇页面的语料库，其中某篇页面的总词数为 1000，并且词“十八大”在该页面中出现了 10 次，那么“十八大”一词在该页面中的词频就是 10/1000=0.01，即 TF=0.01。本论文再次假设在该语料库中有 100 个页面出现了词“十八大”，则反文档频率为 log(10000/100)。至此，本论文可以得到 TF-IDF=TF*IDF=0.01*log(100)。

对于网络小说资料库，本论文通过分词、停用词过滤等预处理以后，计算每篇小说中字词对应的 TF-IDF，对于当前小说页面本论文选择 TF-IDF 得分较高的 k 个词作为其特征词，假设本论文的小说语料库共有 10000 篇小说，那么可以得到 10000*k 个特征词，接下来本论文对这些词进行去重处理就可以得到最后特征词集合，假设这个集合的大小为 m。

至此，对于一篇小说，本论文可以将其表示成如下的数学向量：

$$p=\{TF-IDF_1, TF-IDF_2, \cdots\cdots, TF-IDF_m\}$$

一旦本论文可以把一篇小说表征为数学形式，后续就可以利用机器对其进行各种度量和处理。

3.3.2 基于信息增益的特征提取技术

信息增益又称 Information Gain。在概率论和信息论中，信息增益是非对称的，可以用来度量两种概率分布 P 和 Q 之间的差异。信息增益描述了当使用 Q 进行编码时，再使用 P 进行编码的差异。通常情况下，P 代表样本或者观察值的分布，Q 代表理论模型或者对 P 的近似。在机器学习领域，本论文可以使用信息增益来衡量一个特征项对于训练数据的分类的能力。利用特征取值的情况划分训练样本空间，根据所获得信息量的多少进行特征选择。对于文本分类来讲，信息增益表征对知道一个词在一篇文章中出现或不出现的信息所需要的比特的数量。单词 w 的信息增益值计算公式如下：

$$G(w)=-\sum_{i=1}^{m}P(c_i)\log P(c_i)+P(w)\sum_{i=1}^{m}P(c_i|w)\log P(c_i|w)+P(\bar{w})\sum_{i=1}^{m}P(c_i|w)\log P(c_i|w)$$

其中 $P(c_i)$表示类别 c_i 对应的分布概率值，在小说文本处理过程中，该概率值等于某个类中的所有出现过的词的词频总数与训练数据中所有词的词频总熟的比值。$P(c_i|w)$是类别 c_i 对词条 w 的条件概率，$P(c_i|\bar{w})$是类别 c_i 对除词条 w 的其它词的条件概率。

根据上述信息增益的概念可知，在信息增益中，衡量一个词条的价值就是根据一个特征它所包含的能够帮助预测分类类别的信息量，带来的信息量越多，该特征越重要。从理论上讲，信息增益可以说是最好的特征选取方法，但是，现实生活中，很多很多信息增益较低的特征出现频率往往很高，许多信息增益

较高的特征出现的频率反而很低，这就造成了当信息增益选择的特征数目较少时，容易出现数据稀疏的问题，这个时候如果采用信息增益分类，往往效果很差。针对此类情况，本论文首先对训练文本中出现的每个词计算信息增益，然后指定阈值，去除信息增益低于阈值的词条，选择增益值较高的组成特征向量。

因此，在文本分类过程中，信息增益最能反应一个词条对所有类别的价值，在进行特征选择过程中，本论文选择信息增益值较大的词条作为特征。

3.3.3 基于开方校验的特征提取技术

开方检验又称为 x^2 统计。开方校验的核心思想就是通过观察实际值与理论值的偏差来确定理论是否正确。在实践过程中，本论文先假设两个变量确实是独立的，然后观察实际值与理论值的偏差程度，如果偏差非常小，本论文就认为误差是自然的样本误差，是由于测量手段不够精确导致或者偶然发生的，而两者确确实实是独立的，此时就接受原假设；如果偏差大于一定的阈值，本论文就认为两者实际上是相关的，即否定原假设，而接受备假设。

本论文可以利用如下公式来计算开方校验值：

$$\chi^2(w,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中， A 代表词 w 和类 c 一起出现的次数，B 表示 w 出现但 c 不出现的次数，C 表示 c 出现但 w 不出现的次数，D 代表两者都不出现的次数。

该值越大，表示词 w 越能代表类别 c。

一个词对所有类别的平均代表能力可以利用如下公式计算：

$$\chi^2_{avg}(w)=\sum_{i=1}^{m}P(c_i)\chi^2(w,c_i)$$

Yang 的研究表明，x^2 统计量法是目前特征选择方法中效果最好的一个，x^2 统计量法是很多中文分类系统的首选方法之一。本论文选择对所有类别平均代表能力较大的词作为特征值。

除本论文列举的几种评估函数外，相关系数、证据权值法、几率比法等等也是一些常见的评估函数，每一种评估函数的算法都有自己的优点与不足，本论文根据具体的文档分类算法选择不同的特征选取算法，因为同一种特征选取算法在不同的文档分类算法上效果明显不同。Yang 和 Pedersen J.O.对不同的特征选取算法和分类算法做了很多实验，得出一些结论显示：x^2 统计量法表现较好，分类准确度最高；信息量增益法表现次之，但是其准确性比较接近于 x^2 统计量法；互信息法效果较差，相对于 x^2 统计量法以及信息量增益法，其在准确度上有明显的差距，其中词频-反文档频率是最简单的方法，成本也比较低，其效果比互信息法相比要好很多。

3.4 训练数据准备

本论文将小说分为玄幻、奇幻、武侠、仙侠、都市、青春、历史、军事、游戏、竞技、科幻、灵异等 12 个类别，通过自动与人工相结合，即半自动方式获得本论文的训练数据集合。首先，本论文根据起点中文网上的分类标签获得相应类别下的小说共 2400 篇，每个类别下 200 篇小说作为原始数据，即自动化获取训练数据的过程；然后人工对这些数据进行标注，本论文将每篇小说分给 3 个不同人进行标注，如果自动化标注的类别标签正确那么得分为 1，如果自动化标注的类别标签错误那么得分为 0，若一篇小说文本得分大于等于 2，那么认为自动化标注的类别标签正确，根据这种方式，本论文最终获得训练数据集共 2289 篇小说，各个类别数量分布如表 2.2。

表 3.2 网络小说分类训练集

类别	**数量**
玄幻	192
奇幻	196
武侠	196
仙侠	188
都市	181
青春	198
历史	185
军事	195
游戏	186
竞技	187
科幻	193
灵异	192

3.5 本章小结

本章主要是实现了一个精准小说信息采集器，进而可以及时、准确的采集网络上的小说数据，从而为后文中进行小说文本分类和为用户推荐感兴趣的小说奠定了数据基础，是本论文的基础工具之一。本论文讨论了将采集到的小说文本数据表示为特征向量过程中的核心部分——特征抽取的策略和方法，主要描述了基于词频-反文档频率、基于信息增益、基于开方校验法等三种特征抽取技术。最后通过半自动与人工相结合的方式标注了一份数据集合，本论文后文中的实验工作都是在该数据集合上开展的，它是本论文衡量各种策略和算法性能的依据。

第四章　基于支持向量机网络小说分类技术

4.1 引言

本章首先将描述什么是支持向量机(SVM)，支持向量机是 Cortes 和 Vapnik 于 1995 年提出的一种机器学习算法，支持向量机在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中，它是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。其次讨论了支持向量机分类器的训练方法和在实际应用时的组织方式，以其在训练和分类过程中都能有较高的性能和速度。接下来本论文还对上一章中提出的三种特征抽取技术的效果进行了比较，从而确定最优的特征抽取方法。最后本论文对采集到的小说文本利用支持向量机进行分类，并对分类的结果进行评价。

4.2 基于支持向量机的分类技术

4.2.1 最优分类面的定义

支持向量机 SVM(Support Vector Machine）是一种可训练的机器学习方法，其核心内容是在 1992 年到 1995 年间提出的。SVM 的主要思想包括以下两点：一是它针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能，二是它基于结构风险最小化理论在特征空间中建构最优分割超平面，使得学习机器具有全局最优化，并且在整个样本空间的期望风险以某个概率满足一定的上界值。

支持向量机 SVM 是从线性可分情况下的最优分类面一步步发展而来的, 其发展过程可用如图 3.1 所示，实心点和空心点分别代表不同类型的样本, H 为分类线,H1、 H2 分别为各类中离分类线最近的样本且平行于分类线的直线, H1 与 H2 之间的距离叫做分类间隔，最优分类线就是要求该分类线不仅能够将两类样本完全正确的分开,且使得分类间隔大。

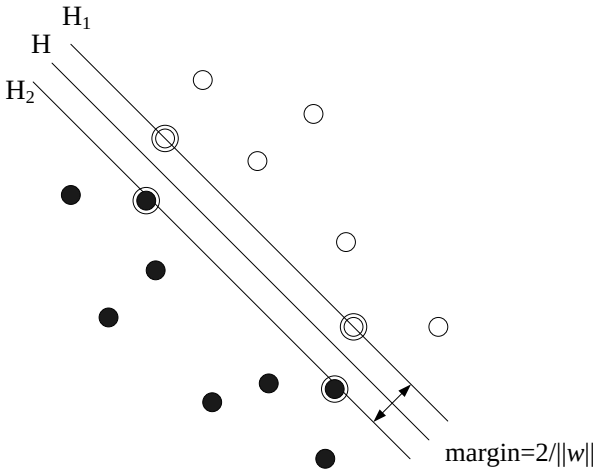

图 4.1 线性支持向量机

分类线对应的方程可以写为：

$$(x_i, y_i), i = 1, \cdots, n, x \in R^d, y \in \{-1, +1\}$$

满足 $y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, \cdots, n$

此时，分类间隔即为 2/||w||,不难看出，要想使分类间隔最大，其实是使得$||w||^2$最小。本论文可以给出如下的定义：满足上述公式且使得$\frac{1}{2}\| w \|^2$值最小的分类面称为最优分类面，落在 H1、H2 上的训练样本点称为支持向量。使分类间隔最大在实践上就是对推广能力的控制。本论文利用 Lagrange 优化方法可以把上述的最优分类面问题转化为其对偶问题来解决，即在约束条件：

$$\sum_{i=1}^{n} y_i \alpha_i = 0 (\alpha_i \geq 0, i = 1, \cdots, n)$$

下对 α_i 求解如下函数的最大值，函数形式为：

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

α_i 为与每个样本一一对应的 Lagrange 乘子，求解对应的样本就是支持向量，得到最优分类函数。

4. 2. 2 支持向量机

对于 N 维空间中的线性函数, 其 VC 维为 N + 1,即使在十分高维的空间中也可以得到较小 VC 维的函数集, 这个特性保证了支持向量机有较好的推广性. 根据上文，本论文不难发现，通过把原问题转化为对偶问题,使得计算的复杂度不再取决于空间维数, 而是取决于样本数, 确切的说是是样本中的支持向量数，这些特点使得支持向量机可以高效的处理高维问题成为可能。对于一个非线性问

题，本论文可以通过非线性变换将其转化为某个高维空间中的线性问题，从而在变换到的更高维空间求解最优分类面。这种变换通常来讲是非常复杂的，因此这种思路在实践中不容易实现。但是幸运的是，在上面的对偶问题中，都只涉及到了训练样本之间的内积运算而已，即在高维空间实际上只需进行内积运算，从而本论文根本没有必要知道变换的具体形式，这种内积运算可以用原空间中的函数实现。因此，在不增加计算复杂度的前提下，在最优分类面中采用适当的内积函数 K 就可以实现某一非线性变换后的线性分类，此时目标函数变为：

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i\cdot x_j)$$

对应的，分类函数变为：

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*\right)$$

支持向量机是一种开创性影响比较大的学习方法，支持向量机有如下几个显著特点：

（1）利用大间隔的思想降低分类器的 VC 维，实现结构风险最小化原则，控制分类器的推广能力；

（2）利用 Mercer 核实现线性算法的非线性化；

（3）稀疏性，即少量样本（支持向量）的系数不为零，就推广性而言，较少的支持向量数在统计意义上对应好的推广能力，从计算角度看，支持向量减少了核形式判别式的计算量；

（4）算法设计成凸二次规划问题，避免了多解性。

总体来说，支持向量机就是通过使用内积函数定义的非线性变换将输入的低维空间变换到某个更高维的空间，在这个高维空间求解最优分类面。

4.2.3 回归方法

支持向量机中不同的内积核函数对应着不同的算法，目前来说，应用较为广泛的核函数主要包括两类：多项式核函数与向基函数。

支持向量机方法可以很好的应用于函数拟合问题。本论文用线性回归函数 $f(x)=w\cdot x+b$ 拟合数据 $\{x_i,y_i\},i=1,\cdots,n,x_i\in R^d,y_i\in R$ 的问题，首先假设所有的训练数据都可以在精度 ε 下无误差地用线性函数拟合，即有如下不等式组：

$$\begin{aligned}&y_i-w\cdot x_i-b\le\varepsilon\\&w\cdot x_i+b-y_i\le\varepsilon\quad i=1,\cdots,n\end{aligned}$$

控制函数集复杂性的方法是使回归函数最平坦，它等价于最小化 $\frac{1}{2}\|w\|^2$，

最终得到回归函数为：

$$f(x)=(w\cdot x)+b=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)(x_i\cdot x)+b^*$$

其中，在实际计算过程中α_i,α_i^*也只有很小的一部分不为 0，这些不为 0 的样本就是支持向量，本论文只要用核函数$K(x_i,x_j)$来代替上式中的内积运算就可以实现非线性函数的拟合。

4. 2. 4 基于支持向量机的分类器构造

一般来讲，支持向量机是一种典型的两类分类器，即它只回答属于正类还是负类的问题。而现实要解决的问题中往往是多类的问题，例如小说文本分类问题，本论文需要支持向量机来判断小说是武侠、言情、军事或是其他等类别。如何由两类分类器得到一个多类分类器是一个在实践过程中经常遇到的问题。

以传统的文本分类为例，有一种一劳永逸的做法，即一次性考虑所有样本并且求解一个多目标函数的优化问题，进而一次性的得到多个分类面，如下图 3.2 所示 。

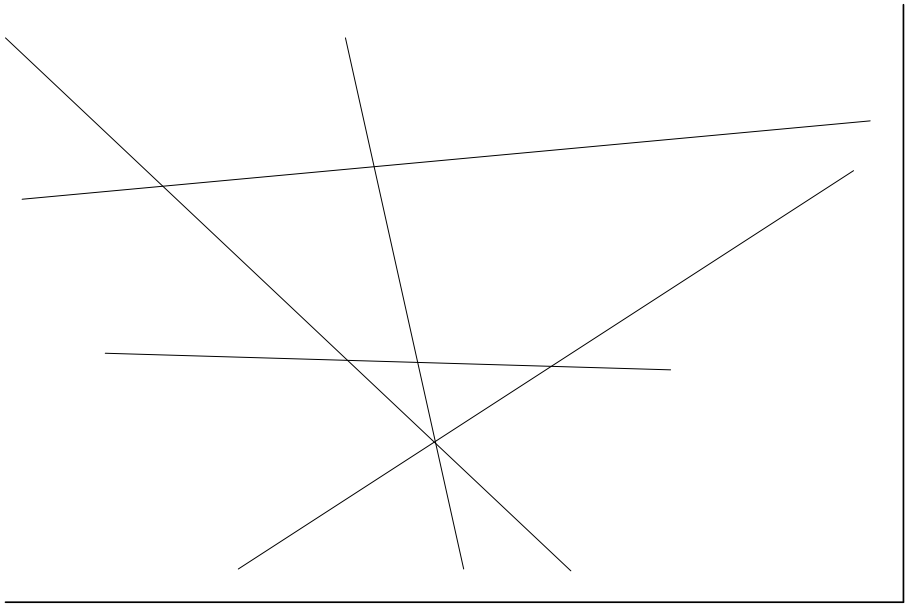

图 4.2 多类支持向量机

可见多个超平面把空间划分为多个区域，每个区域对应于一个类别，给定一篇文本，根据其落在哪个区域来确定其对应的分类。不幸的是该方法从理论上看起来很完美，但是一次性求解的方法的复杂度已经超过了计算机的处理能力，因此无法将该方法应用到实践当中。折衷地，本论文想到了一种“一类对其余”的方法，就是每次仍然求解一个两类分类的问题。例如现有五个类别分

别记为 A、B、C、D、E，第一次就把类别 A 的样本设定为正样本，其余 B、C、D、E 四个类别的样本设定为负样本，这样就可以得到一个把类别 A 与其它类别区分开来的两类分类器，即它能够指出一篇文本的类别是否是 A；类似地，第二次把类别 B 的样本设定为正样本，把 A、C、D、E 的样本设定为负样本，从而得到一个能够判断一篇文本是否是 B 类的分类器；如此下去，最终可以得到五个这样的两类分类器。当有一篇文本到来时，分别询问这五个分类器来确定该文本的类别，这种方法的优点是每个优化问题的规模较小、分类速度快，但是可能会出现一下这些问题，例如一篇文本到来时，分类器 A、B 都说该篇文本属于自己的类别，该现象称为分类重叠，一般地我们利用该篇文本到各个超平面的举例为其确定最可能的类别；另外一种现象是所有分类器都无法为该篇文本确定类别，该现象称为不可分类，一般的只能认为该文本确实无法确定类别了；另外一个问题是分类器只拿该类的数据作为正样本，而其他所有类别的数据作为负样本，这就是数据集偏斜问题，该问题会引起分类器性能的下降。为了避免数据集偏斜问题，是否可能每次选一个类的样本作正样本，另外一个类的样本作为负类样本，这样就很容易的避免了偏数据集斜问题。我们仍然假设有 A、B、C、D、E 等五个类别，因此过程就是算出这样一些分类器，第一个分类器只回答“是类 A 还是类 B”，第二个分类器只回答“是类 A 还是类 C”，第三个分类器只回答“是类 A 还是类 D”，如此下去，我们可以最终得到 10 个分类器来对 A、B、C、D、E 这五个类别进行划分，一般地，如果有 k 个类别，则需要 k(k-1)/2 个分类器来实现类别的划分。虽然分类器的数目较多，但是由于复杂度的降低，在训练阶段所用的总时间却有所减少，在分类的过程中，当一篇文本到来时，第一个分类器会判定其是类 A 还是类 B、第二个分类器会判定其是类 A 还是类 C，最后让每一个分类器都给出自己的分类结果，最后统计所有的分类结果，如果类别 A 得票最多，就判这篇文本属于类别 A。这种方法显然也会出现分类重叠的现象，但不会再出现不可分类现象，因此该分类策略解决了上述的所有分类缺陷。但是又引起了新的问题，假如我们的类别数量是 1000，那么要调用的分类器数目会上升至约 500,000 个，显然在实践过程中很难容忍这样巨大的开销。因此，我们必须想办法在保持上面一对一方法那样来训练分类器，同时优化分类器之间的组织方式来提高分类的效率，使其能够在实践中得到应用。本论文首先按照图 3.3 的方式来组织分类器（这是一个有向无环图，因此这种方法也叫做 DAG SVM）。

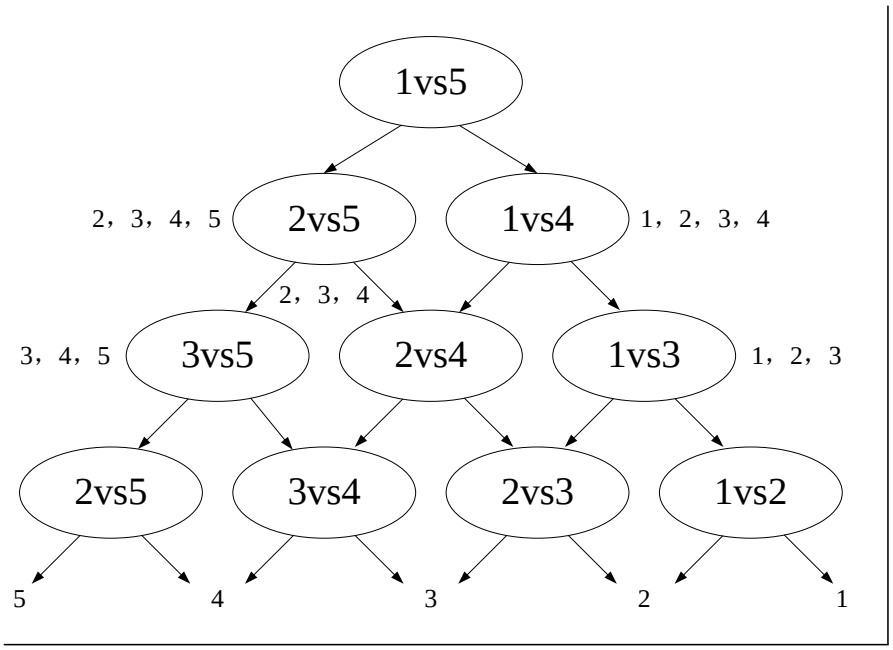

图 4.3　DAG 支持向量机分类器

这样在分类时,我们就可以先询问分类器“1vs5”，如果它的类别结果是 5，那么本论文就往左下走，再询问分类器“2 对 5”，如果它还说是“5”，本论文就继续往左下走，这样一直问下去，就可以得到分类结果，我们其实只调用了 4 个分类器就是实现了类别的划分，该方法既解决了重叠划分和不可划分的现象，又提高了分类速度，但是引出了一个新的问题，假设在最一开始的分类器回答错误，那么后面的分类器再也无法纠正该错误。因此在实践当中，当类别较少时，本论文一般采用一对一方式训练和组织分类器，当类别数量较多时，本论文采用该段描述的方法来组织分类器。使用支持向量机进行分类的时候，可以分为训练和分类两个完全不同的过程。训练阶段的复杂度就是上文所说的求解最优分类界面的时空开销。分类阶段的复杂度就是上文所说的分类策略所产生的开销。在本论文中，在训练阶段采用一对一的方式训练分类器，由于本论文为小说共划分了 12 个类别，根据上文分析共需要 66 个分类器，同时由于单个类别下的样本数量不大，因此可以快速的训练分类器，本论文采用一对一的方式来组织这 66 个分类器，每次分类时，这 66 个分类器都投上自己的一票，将投票最多的类别设定为该文本的最终类别。

4.3 网络小说文本预处理及向量表示

当前，在处理文本信息的过程中，一般情况下，选取词作为文本的特征项表现要优于选择词组或者字。因此，本文选择使用词作为文本的特征项。本论文首先将小说文本分成词，由这些词组成向量元素来表示文本。但是，本论文

使用中文分词器切分词条时，经常含有大量的单个独立字，这些独立字不仅携带信息量较少，而且对文本分类的准确性和处理效率产生极大影响。因此，本论文在进行文本分词时，首先要过滤单个独立字。此外，一些数学符号以及英文字符对文本分类贡献极小，可以忽略。最后，需要过滤掉纯英文词条。经过上述预处理过程，可以有效地降低文本词条向量的维数。提高文本特征向量的中文纯度。

网络小说文本预处理实现过程如图 3.4。

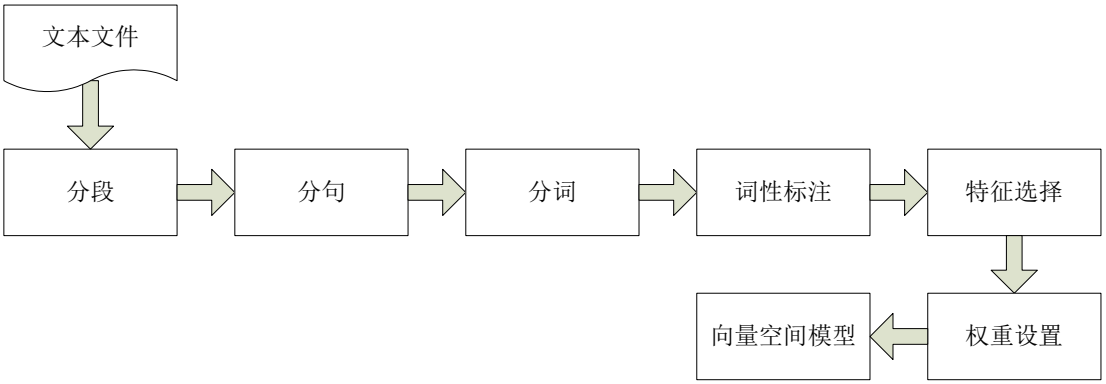

图 4.4 网络小说文本预处理过程

（1）首先，删除文本中的空格；

（2）对文本进行分段，并标记段落标识；

（3）识别断句符号，并标记断句符号；

（4）使用算法对句子进行分词，并标记词性，去除独立字以及英文符号，数学符号；

（5）特征词识别；

（6）特征词权重设置；

（7）形成文本向量空间模型。

在第二章中，本论文讨论了 3 种经典的特征抽取技术以及上节的文本预处理过程。通过将小说文本预处理本论文就可以把小说文本表示为相应的特征向量，从而展开相应的计算工作。假设本论文通过特征抽取技术抽取到的特征词个数为 M，第 i 个特征词表示为 T_{i}，那么一篇网络小说文本表示为：$\{W_{T_1},W_{T_2},\cdots\cdots,W_{T_M}\}$，其中 W_{T_i} 表示特征词的权重。权重计算通常采用 TF-IDF 方法。

为了实现对小说文本的分类，本论文需要将小说文本表示成特征向量，再将其送入支持向量机进行类别预测。将小说文本表示成特征向量的核心问题是如何选取特征向量。

根据上文的分析，本论文共有 3 种方法选择特征，这三种方法分别是基于词频-反文档频率、基于信息增益、基于开方校验法的特征选择。

本论文首先对训练数据集中的 2289 篇文本小说进行分词、去停用词处理，然后分别利用基于词频-反文档频率、基于信息增益、基于开方校验法进行特征选择，选择出来的特征空间分别记为 T1、T2、T3。

一篇文本的三种特征向量空间分别表示如下：

$$A1=\{w_{1-tf-idf},\cdots,w_{|T1|-tf-idf}\}$$
$$A2=\{w_{1-tf-idf},\cdots,w_{|T2|-tf-idf}\}$$
$$A3=\{w_{1-tf-idf},\cdots,w_{|T3|-tf-idf}\}$$

$w_{i-tf-idf}$ 表示特征 w_i 的词频-反文档权重。$|T1|$ 表示基于词频-反文档频率方法选择出的特征向量个数、$|T2|$ 表示基于信息增益方法选择出的特征向量个数、$|T3|$ 表示基于开方校验法选择出来的特征向量个数。对于这三种方法，本论文需要选择出最优一种特征选择方法来解决本论文的问题。

在上一章中本论文描述了三种特征抽取的方法将小说文本表示为相应的空间向量，为了衡量基于词频-反文档频率、基于信息增益、基于开方校验法等三种特征抽取算法的性能，本论文分别利用以上三种方法做特征抽取工作，然后用 SVM 分类器进行类别预测，预测结果的优劣就反应不同特征抽取算法的效果。

本论文利用人工标注的 2289 条数据作为训练集合，根据不同的特征抽取算法得到三个不同的分类器，然后再将这 2289 条数据依次送入三个分类器进行封闭条件下的类别预测，实验结果表明基于信息增益的特征抽取算法具有最好的分类效果，基于开方校验法的特征抽取算法的分类效果位于第二，并且与基于信息增益的特征抽取效果差距不大，而基于词频-反文档频率的特征抽取算法的分类效果最差。因此本文采用基于信息增益的特征抽取算法生成特征空间。图 4.5 为基于支持向量机的分类模型图，下一节本论文将对此进行实验。

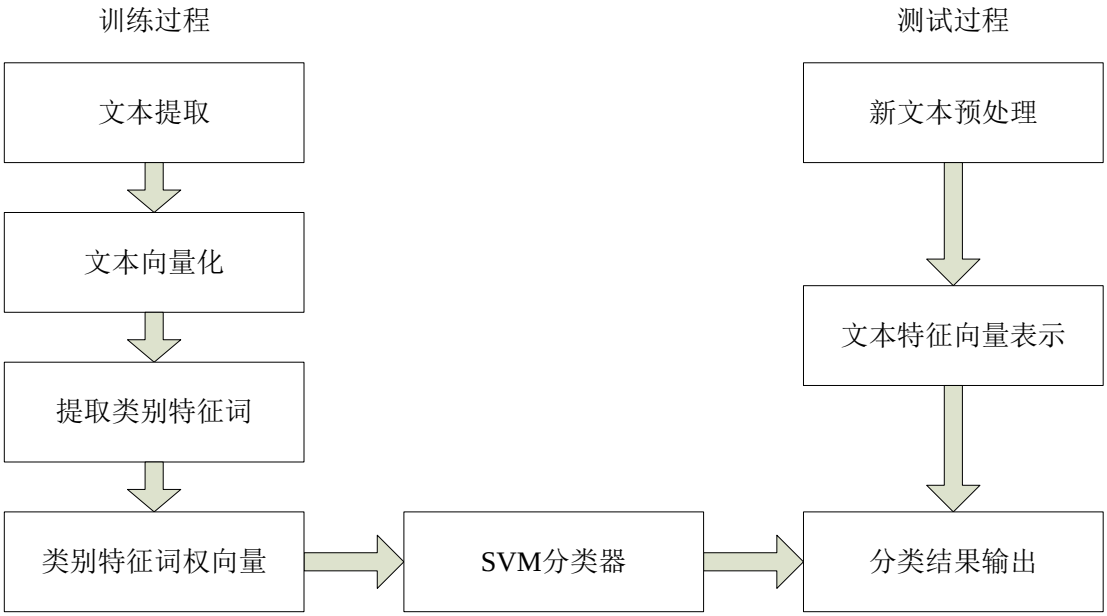

图 4.5 基于支持向量机的分类模型图

4.4 实验

分类实验主要包括两个部分：分类器的训练过程与类别的预测过程。下面将分别描述上面两个部分的具体操作。

分类器的训练过程主要包括以下几个步骤：文本的分词以及去停用词预处理、基于信息增益法的特征选择、将小说文本表示成相应的特征向量、分类器训练等四个主要模块。文本的分词以及去停用词预处理工作模块中，本论文采用中科院的开源分词工具对网络小说文本进行汉语词的切分，并去掉例如“的”、“是么”等没有实际意义的停用词，将小说文本表示成一些词的集合。在基于信息增益法的特征选择模块中，本论文利用信息增益公式计算得出得分较高的词作为本论文的特征。在将小说文本表示成相应的特征向量中，本论文将文本表示为第二步的特征集合，并为每个特征赋予权值。在分类器训练模块中，本论文利用开源的 SVM 工具包进行分类器的训练，共计得到 66 个分类器。

类别的预测过程主要包括以下几个步骤：文本的分词以及去停用词预处理、基于信息增益法的特征选择、将小说文本表示成相应的特征向量、分类预测等四个主要模块。其中前三个模块的具体工作与分类器训练过程相应的模块一致。在分类预测模块中，本论文采用开源的 SVM 工具包的类别预测接口进行类别判定，具体做法是，当一篇小说文本到来时，本论文将其分别送入得到的 66 个分类器中，每个分类器都会给出自己的类别预测，本论文综合所有分类器的类别预测结果，选择出现次数最多的类别作为该篇小说最终的类别结果。

在具体的实验过程中，本论文取 2289 条人工标注的数据集的前 2250 条参与到实验过程中，进一步地将这 2250 条数据随机拆分成 5 等份，每份包括小说数量为 450 篇，本论文将这五份数据分别记为 part_1、part_2、part_3、part_4、part_5。接下来本论文在这五份数据上进行交叉实验，即随机选择其中的四份数据作为训练集合来训练本论文的分类器，剩下的一份数据作为测试集来衡量分类器的效果。从而本论文共得到了 10 组实验结果，分别记作 Test1 到 Test10，具体数值如表 3.1。

交叉实验的结果表明，开放测试条件下分类器的性能介于 91.5%到 92.3%之间，而通常封闭条件下的测试结果分类器的正确率达到了 97.4%，可见封闭测试下的性能比开放测试高出 5 个百分点以上。封闭测试下性能远高于开放测试下的性能说明本论文的样本空间过小，分类器对于“没有见过”的特征预测效果不好，因此本论文设想通过扩大训练集规模来提升分类器的效果。

目前，本论文拥有的训练集规模人工标注的 2289 篇小说文本，要扩大训练集的规模的途径有两种：一种方法是采用人力标注更多的训练集，该方法的优点是标注结果准确度高，缺点是人力开销过大；另一种方法是利用机器自动标注一部分训练集，该方法的优点是标注速度快、成本小，缺点是标注结果的准

确度没有人工标注高。

根据实际情况，本论文只能采用第二种方法扩大训练集的规模。采用机器标注面临的首要困难就是如何提升标注的准确率，因为一旦标注的准确率过低，再把这部分数据作为可信数据加入到训练集当中，不仅不会提升分类器性能，反而会引入更多的噪音。幸运的是，根据本论文的类别预测策略，本论文的 66 个分类器都要给出自己的类别预测结果，因此可以认为得票率大于 M 的类别是值得信赖的，这个问题从策略上就得到了解决，接下来就是 M 如何取值。本论文综合上一小节中的 10 次实验结果，将类别预测正确的小说分为一组，记为 A，将类别预测错误的小说分为另一组，记为 B，本论文分别统计 A、B 两组中每篇小说最终类别的得票数量，发现 A 组中最终类别的得票数量均值远高于 B 组中最终类别的得票数量，这恰好符合本论文的预期，因此本论文取 M 为 A 组中最终类别的的票数量均值最为阈值，若最终预测的类别得票数大于 M，那么认为该类别是可信的，则将其加入到本论文的训练集合当中。至此，本论文可以利用机器标注任意大规模的数据集。

表 4.1 分类器类别预测结果

实验组别	**正确率**
Test1	91.7%
Test2	91.5%
Test3	92.2%
Test4	92.3%
Test5	91.7%
Test6	91.6%
Test7	91.8%
Test8	91.8%
Test9	92.2%
Test10	92.3%

本论文通过上面描述的方法，利用机器标注 100000 篇小说为本论文的训练集，并将该训练集平均分成 5 份，每份包括 20000 篇小说，利用每份训练集分别进行分类器的训练，从而本论文可以得到 5 组实验，分别记为 Test1 到 Test5。在每组实验上利用人工标注的 2289 篇小说作为测试集来衡量分类器的效果。具体数据如表 3.2。

增大训练数据集规模后的实验表明，正确率介于 96.7%到 97.3%之间，已经比较接近封闭测试下正确率，比小规模数据集上的分类器的性能提升了 5 个百

分点左右，可见增大训练集的规模可以提升分类器的效果。

表 4.2 扩大训练规模后分类器类别预测结果

实验组别	**正确率**
Test1	96.7%
Test2	97.3%
Test3	97.1%
Test4	96.9%
Test5	97.0%

4.5 本章小结

在本章中首先详细描述了支持向量机的原理以及其在机器学习方面的优势和应用。接下来详细叙述了基于支持向量机的分类器的训练以及预测过程的应用背景、常见问题和解决方法。本论文构建了基于支持向量机的小说分类器，并对比了三种不同特征抽取算法的优劣，最终选定基于信息增益法的特征抽取算法作为本论文特征选择的依据，利用实验评测了本论文分类器的效果，并通过扩大训练集规模提升了的分类器效果，最终本论文分类器的正确率稳定在 97%以上，可以应用到实践过程中。

第五章　融合用户兴趣与文本内容的小说推荐技术

5.1 引言

本章主要包括以下几个部分，首先是如何从互联网上获取小说用户的行为数据，这是本论文推荐系统的物质基础；其次描述了基于内容相似度的推荐技术并对该方法的推荐效果利用用户的真实数据进行评价；最后描述了基于用户协同过滤的推荐技术并对该方法进行评价，同时对比了两种不同推荐技术的技术指标和各自的优缺点。

5.2 获取用户数据

对于推荐系统来说，用户行为数据是最重要的，因为用户行为数据是衡量推荐系统的基础，只有通过用户行为数据才能衡量本论文推荐系统的质量。对于本论文的小说推荐系统来说，最重要的是需要获取一批实际互联网用户的有关小说的行为数据。因为本论文没有实际的小说阅读系统，因此本论文只能通过抓取第三方网站来获得用户小说行为数据。

豆瓣是中国最富盛名的评论和社交网站，上面有大量用户的读书行为数据。对于一个用户来讲，其有关读书的行为数据包括三种：在读、读过、想读等类型，一旦某篇小说出现在在读、读过、想读三者之一的列表中，本论文就认为用户对这本书感兴趣，本论文的目的就是获取用户这样的行为数据，并利用这样的数据来衡量推荐系统的质量。本论文利用网络爬虫从豆瓣上抓取用户以及该用户的读书记录，并将结果保存到数据库当中。网页页面形式如图 5.1

图 5.1 用户登录后小说网页页面显示

本论文可以通过该页面获得用户 waits 感兴趣的小说，包括在读的小说 6 本、读过的小说 387 本、想读的小说 234 本，利用这种方式，本论文就可以建立一个用户感兴趣小说的知识库文件。

5.3 基于小说内容的推荐技术

5.3.1 文本相似度计算

基于内容的推荐技术是基于这样的假设：用户的兴趣在一段时间内是稳定的，对某一类或某几类相似的事物感兴趣。例如张三喜欢《天龙八部》，那么本论文猜测他也可能对《射雕英雄传》感兴趣，根据我们的实际经验，对《天龙八部》感兴趣的人往往是武侠迷、金庸迷，对《射雕英雄传》感兴趣的可能性非常大，因此本论文的任务就落到了当我们知道张三喜欢《天龙八部》时，如何将《射雕英雄传》推荐给他，我们都知道《射雕英雄传》与《天龙八部》在作者、背景、武功、人物上有很多的相似之处，如果我们能够通过某种方法得到《射雕英雄传》与《天龙八部》的相似距离，假如两者的距离非常小，我们就可以认为两者比较相似，进一步的认为当张三对《天龙八部》感兴趣的同时非常可能对《射雕英雄传》也感兴趣，这个计算小说间相似性并根据相似距离向用户推荐小说的过程就称为基于内容相似性的文本推荐技术。

本论文首先讨论如何衡量小说直接的相似性。本文采用向量空间模型来计算小说之间的相似性。向量空间模型(VSM)是由 Salton 等人于上个世纪 60 年代提出的，目前在信息检索、知识挖掘等领域都有着广泛的应用。其基本思想是将文档看成是由彼此之间相互独立的一个个词组成的，即上文所说的特征向量，同时根据每个词在文档中的重要程度为其赋予一定的权重，即上文所说的词频-反文档频率，这样就可以将一篇小说表示为$(w_1, w_2, \cdots, w_n)$这样的向量形式。进而本论文可以用如下的公式来计算两篇小说之间的相似性。

$$Sim(T_1, T_2) = \sum_{i=1}^{n}(T_{1,i} \times T_{2,i}) / \left(\sqrt{\sum_{i=1}^{n} T_{1,i}^2} \times \sqrt{\sum_{i=1}^{n} T_{2,i}^2} \right)$$

其中 T_i 表示文本 i 的特征向量，$T_{1,i}$ 表示文本 1 的第 i 个特征词对应的权重。

5.3.2 实验

本论文在人工标注的 2289 条小说数据集上进行基于余弦空间向量夹角的相似度计算，并对该方法的准确性进行评价。

实验流程如下，首先需要对小说文本进行分词、停用词过滤、特征提取、TF-IDF 权重计算等预处理工作，然后再依据余弦空间向量夹角公式计算小说直接的相似度。假如本论文需要处理小说集合的同类别下样本数量为M，则本论文需要计算$(M-1)\times(M-2)\times\cdots\times 1$组小说直接的相似度，可见当 M 较大时，这个计算的时空开销是非常巨大的，不过在实践过程中，本论文可以离线的将这

些数据计算好，即本论文可以通过计算得到每篇小说与其它小说直接的相似度，然后本论文选择相似度较高的 10 部小说存放到数据库当中，当用户请求到来时，就可以将与用户当前浏览小说相似度最近的 10 部小说推荐给他，至此本论文就完成了基于内容相似度的小说推荐。流程图如图 4.2。

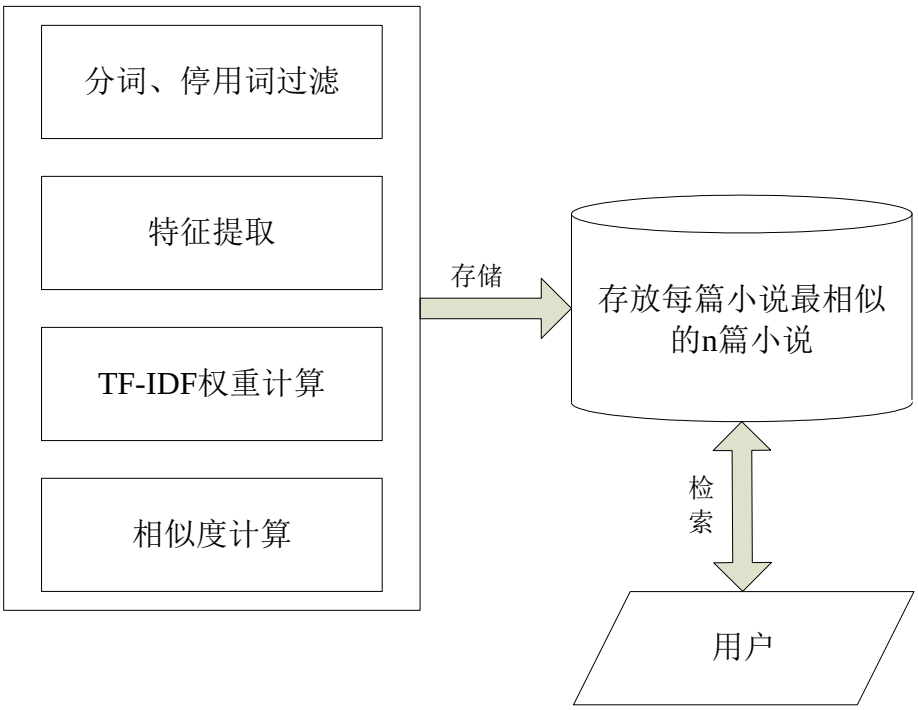

图 5.2 基于内容相似度的小说推荐流程图

根据上面的流程图，本论文就可以实现基于内容相似度的小说推荐系统，本论文接下来的主要任务是评价推荐系统的效果，从而确定一些实践中的参数。推荐系统的评价是从两个角度进行的，一是相似度算法本身的效果、二是用户对推荐小说的满意程度，下面本论文就分别描述以上两种评价方式。

从 2289 篇小说中每个类别下随机抽取 50 篇小说，然后分别对与其最相似的 10 篇小说进行人工相似性的评价。具体做法如下，本论文将这些随机抽取的小说以及与其最相似的 10 篇小说数据下发给 3 名同学，如果相似则打 1 分，反之打 0 分，若有两个以上同学打 1 分，则认为两篇小说是相似的。本论文最终利用 P@N 指标来衡量相似性效果，P@N 常用于信息检索效果的评价中，这里是指前 N 篇小说中正例的比例，即准确性。

从指标上，本论文可以得到如下总结，对于某篇小说来讲，与其最相似的前 10 篇小说的准确率可以维持在 90%以上，与其最相似的前 5 篇小说的准确率可以维持在 95%以上，而相似性的效果是整个推荐系统效果的基础，在一般工程实践中，正确率在 95%以上的技术才能应用的实际产品当中，因此本论文选择前 5 篇小说推荐给用户，同时由于实际推荐过程小说的候选集合较大，因此

计算开销十分巨大，不过我们知道，类别不同的小说之间的相似度一定不高，因此我们在实际工程中，只对同类别的小说之间进行相似度计算，而第三章中的讨论可以帮助本论文很好的为一篇小说文本判断类别，从而使得本论文整个流程顺利的进行。然而，并不是将与当前用户浏览小说最相似的前 5 篇小说推荐给用户，就可以获得 95%的准确率。因为用户并不一定喜欢相似的小说，本论文需要根据用户对所推荐小说的反馈情况来确定用户是否对所推荐的小说感兴趣，只有用户感兴趣才算有效的推荐。本论文利用前面抓取的豆瓣用户数据来评估本论文的推荐系统，首先本论文统计出浏览小说数量大于 100 篇的用户，并从中选择出 100 个用户对本论文的推荐系统效果进行评价，评价方法如下：如果本论文推荐出的小说曾经被用户阅读过，则记为有效推荐，反之记为无效推荐。

表 5.1 用户对基于内容的推荐结果的满意度评价

N	P@N
1	99.0%
2	97.9%
3	96.7%
4	96.1%
5	95.4%
6	93.3%
7	92.1%
8	92.0%
9	91.0%
10	90.3%

从实验可以得出，推荐与用户当前浏览小说最相似的 5 篇小说可以获得 6%以上的准确率。在实际应用中，如果本论文给用户推荐出了有效的推荐，可以提高用户的兴趣和粘性，反之，则可能会伤害用户的体验，因此，本论文取 N=5，即在推荐系统中，选择与用户当前浏览小说最相似的前 5 篇小说推荐给用户。至此，本论文从基础算法选择、算法效果、推荐系统参数选择等方面完成了基于小说内容的推荐系统的描述。

5.4 基于协同过滤的推荐技术

5.4.1 基于用户的协同过滤算法

基于用户的协同过滤算法是推荐系统的最经典算法之一，是大多数推荐系统的核心。该算法的核心思想用中国的一句古话来讲就是“人以群分”，即假设甲乙两人兴趣相似，若甲喜欢物品 C，那么乙也可能喜欢物品 C，因此，基于用户的协同过滤算法可以分解为以下两个子任务：一是如何找到兴趣相似的用户

集合，二是如何找到这个用户集合中用户喜欢的物品推荐给目标用户。下面本论文就这两个子任务的算法分别进行描述。

子任务一的关键点是如何计算用户之间的兴趣相似度，本文利用用户的小说浏览记录行为来衡量用户相似度。假设两个小说用户 a 和 b，令 $N(a)$表示用户 a 曾经浏览过的小说集合，令 $N(b)$表示用户 b 曾经浏览过的小说集合，那么可以通过余弦相似度来衡量两个用户之间的兴趣是否相似，公式如下：

$$w_{ab}=\frac{|N(a)\cap N(b)|}{\sqrt{N(a)\| N(b)}}$$

不过，我们都有这样的实际经验，如果两个用户都有过浏览《新华字典》的行为，我们能说明这两个用户相似么？显然不能，因为几乎每个中国人可能都浏览过《新华字典》，但是若两个用户都浏览过《射雕英雄传》，我们能说明这两个用户相似么？显然可以，因为看过《射雕英雄传》的用户一定是对武侠比较感兴趣的。然而，在上述公式中，我们对《新华字典》和《射雕英雄传》是同等对待的，这显然是不合理的，因此我们希望能够对《新华字典》这类几乎每个人都会浏览的热门物品进行降权处理，使得我们的算法更合理。因此本论文引入了如下的公式：

$$w_{ab}=\frac{\sum_{i\in N(a)\cap N(b)}\frac{1}{\log^{1}+|N(i)|}}{\sqrt{|N(a)\|N(b)|}}$$

可见，通过引入权重因子 $\frac{1}{\log^{1}+|N(i)|}$ 惩罚了出现在两个用户公共物品列表中的热门物品对两者相似度的影响，通过该公式本论文可以很好的计算用户直接的相似度。

在实际应用过程中，假设共有 M 个用户，那么计算用户之间相似度的时间开销为 $O(M^2)$，当用户数量较大时，使得计算难以推广。通过实际操作发现，绝大多数用户之间没有重合的行为数据，即用户 a 与用户 b 的行为记录不相交，因此本论文只需要计算出用公共行为的用户对$(a，b)$，然后在对这些用户进行相似度计算，从而提高相似度计算过程的速度。

子任务二是找到用户兴趣集合中目标用户可能感兴趣的小说。首先需要确定用户兴趣集合，当给定目标用户 a 时，利用子任务一中提到的公式计算出与其兴趣最相似的 K 个用户，这些用户的集合记为 $S(a，K)$，进而可以利用如下的公式来衡量用户 a 对物品 i 的感兴趣程度：

$$p(a,i)=\sum_{b\in S(a,K)\cap N(i)}w_{ab}\gamma_{bi}$$

其中，$N(i)$表示对物品 i 感兴趣的用户集合，w_{ab} 表示用户 a 与 b 之间的相似度权重，γ_{bi} 表示用户 b 对物品 i 的感兴趣程度。利用如上的公式，本论文就可

以计算出用户 *a* 感兴趣的物品集合，从而按照兴趣程度将物品推荐给用户。

至此，本论文根据如上的算法就可以实现基于用户协同过滤的推荐系统，接下来会描述该推荐系统的流程并对系统的效果进行评价。

5.4.2 实验

基于用户协同过滤的推荐系统主要包括如下几个步骤：首先要在抓取的用户行为数据集之上根据子任务一中描述的公式计算两两用户之间的相似度；其次利用子任务二中的公式计算出每个用户最感兴趣的小说集合，并选择前 20 个小说存储到数据库当中；最后当用户到来时，从数据库取出其最感兴趣的 20 篇小说并过滤掉其已经浏览过的小说，最后推荐出 10 本小说给用户。流程图如图 5.3。

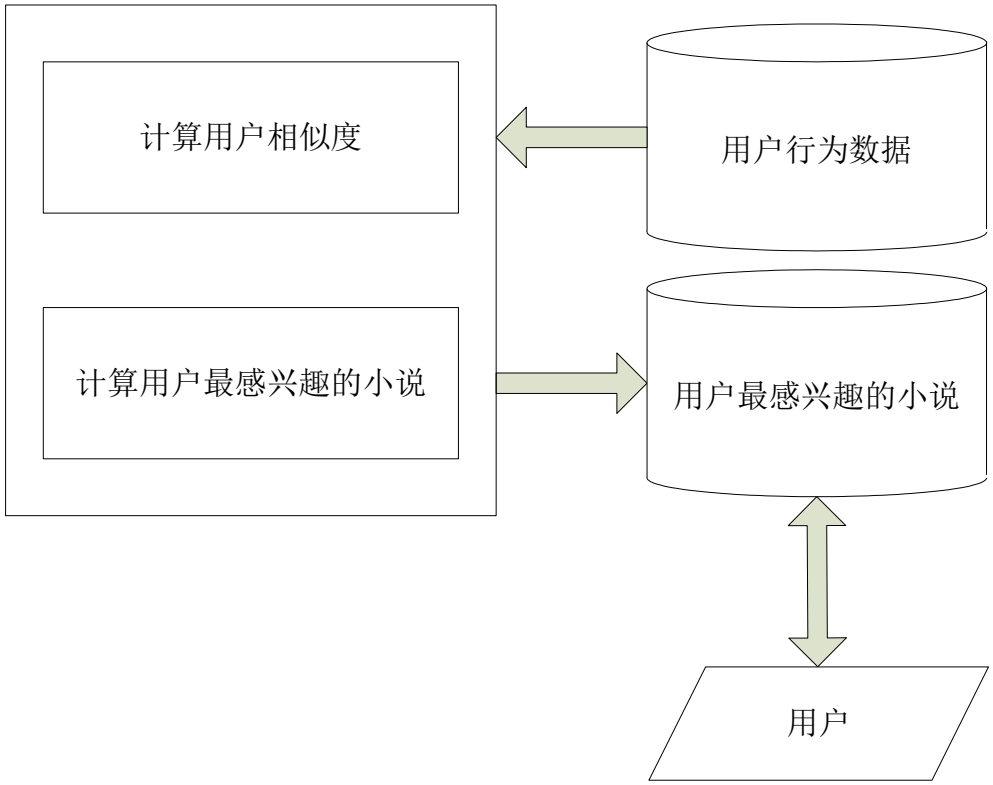

图 5.3 基于协同过滤的小说推荐流程图

至此，本论文就实现了基于用户协同过滤的推荐系统。接下来就需要对本论文的推荐效果进行衡量。与上一节中一样，本论文利用 P@N 指标来衡量本论文的推荐系统的效果，本论文这里首先取 N=10。

从以上的指标可以看出，基于用户协同过滤的推荐算法要优于基于内容相似度的推荐算法，因此本论文选择基于用户协同过滤的推荐算法作为最终小说推荐系统的核心算法。同时，本论文发现当 N=5 时，推荐的准确率掉到 15%一下，因此在实际应用中，本论文为每位用户推荐 5 本小说。至此，本论文完成

了基于协同过滤技术的网络小说推荐系统。

表 5.2 用户对基于协同过滤的推荐结果的满意度评价

N	P@N
1	20.1%
2	19.9%
3	18.7%
4	18.1%
5	17.4%
6	14.3%
7	13.3%
8	10.0%
9	9.9%
10	9.6%

5.5 基于内容的推荐技术与基于协同过滤的推荐技术相融合

在上面的两个小节中，本论文分别实现了基于内容的推荐技术与基于协同过滤的推荐技术。两种推荐技术有着各自的优劣：基于内容的推荐技术只能为用户推荐与用户当前浏览小说内容相似的小说，基于协同过滤的推荐技术可以通过兴趣相似的其它用户的行为为用户推荐小说，但是如果当一个新用户到来或者无法获取当前用户的历史行为记录时就会无能为力。由于以上的原因本论文在实际的应用中会融合基于内容的推荐技术与基于协同过滤的推荐技术，当我们可以获取用户的历史行为记录时就使用基于协同过滤的推荐技术来为用户完成推荐，当无法获取用户的历史行为记录时就使用基于内容的推荐技术为用户完成推荐，从而克服了推荐系统中冷启动的难题。

推荐系统流程图如图 5.4。

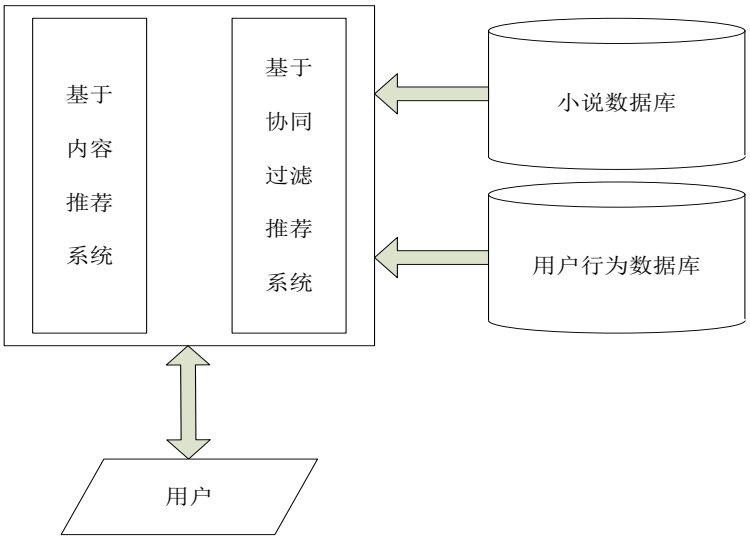

图 5.4 基于内容和协同过滤的小说推荐流程图

5.6 本章小结

本章中主要描述了如何从网页中获取小说用户行为数据，同时描述了基于内容相似度的小说推荐流程和基于用户协同过滤的小说推荐流程，并分别对两种核心算法的推荐效果进行评价，最终在实际应用中将两种算法相结合获得了最优的小说推荐效果。

第六章 总结与展望

6.1 本文所做的工作

本论文采用了文本分类方法来对网络小说进行分类和推荐。在本论文中，网络小说的分类和推荐分为三步：网络小说文本特征提取；基于支持向量机的网络小说分类；融合融合用户兴趣与文本内容的小说推荐。本论文通过对比TF-IDF、信息增益、开方校验法等三种不同的特征抽取技术，选择最适合网络小说文本分类应用的特征抽取技术。采用支持向量机(SVM)对小说文本进行分类，并通过分类器的组合来提高分类速度。采用基于内容相似度的文本推荐技术和基于用户协同过滤的推荐技术。实验研究表明本设计方案能够很好的为用户推荐用户感兴趣的网络小说，能够满足大多数用户的需求。

6.2 对未来的展望

推荐系统研究是一个新兴的领域，虽然目前在各种推荐算法已经取得不错的研究成果，尤其在基于内容的推荐和基于协同过滤的推荐算法，但是还存在很多问题，如数据的获取主要依赖显示评价，而获得用户隐式信息方面做得还远远不够，对推荐系统的开发和应用，尤其与企业其它系统的集成的研究还远远不够，这个将是未来的一个研究重点。

参考文献

[1] 周水庚.中文文本数据库若干关键技术研究[D].复旦大学硕士学位论文，2000

[2] 汪传建.基于混合模型的文本分类的研究[D].东北大学硕士学位论文，2005

[3] 王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J].计算机研究与发展，2000.37(5)：513-520

[4] 王斌,潘文锋.基于内容的垃圾邮件过滤技术综述[J].中文信息学报，2005.19(5):1-10

[5] 王本年,高阳,陈世福等. Web 智能研究现状与发展趋势[J].计算机研究与发展,2005.42(5)：721-727

[6] 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[J].计算机研究与发展,2000.37(9)：1033-1038

[7] 刘永丹.文档数据库若干关键技术研究[D].复旦大学硕士论文,2004

[8] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报，2006.17(9):1848-1859

[9] 黄萱菁.独立于语种的文本分类方法[J].中文信息学报,2000.14(6):1-7

[10] 侯汉清,分类法的发展趋势简论[M].中国人民大学出版社,1981

[11] 张滨.中文文档分类技术研究[D].武汉大学硕士学位论文,2004

[12] 张治平.Web 信息精确获取技术研究[D].国防科学技术大学硕士学位论文，2004

[13] 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004.18(1):26-32

[14] 黄昌宁.中文信息处理中的分词问题[J].语言文字应用,1997,1：72-78

[15] 孙茂松,左正平,黄昌宁.汉语自动分词词典机制的实验研究[J]，1995,14(1):1-6

[16] 吴栋,滕育平.中文信息检索引擎中的分词与检索技术[J].计算机应用，2004.24(7):128-131

[17] 陈桂林,王永成.一种改进的快速分词算法[J],计算机研究与发展，2000,37(4):418-424,452

[18] 步丰林.一个中文新词识别特征的研究[J].计算机工程.2004,30(B12):369-370

[19] 周茜,赵明生,扈曼.中文文本分类中的特征选择研究[J].中文信息学报，2004,18(3):17-23

[20] 朱颢东,钟勇.一种新的基于多启发式的特征选择算法[J].计算机应用，

2009,29(3):849-851
[21] 张海龙,王莲芝.自动文本分类特征选择方法研究[J].计算机工程与设计,2006,27(20):3838-3841
[22] 尚文倩,黄厚宽,刘玉玲,林永民,瞿有利,董红斌.文本分类中基于基尼指数的特征选择算法研究[J].计算机研究与发展,2006,43(10):1688-1694
[23] 柴玉梅,王宇.基于 TF-IDF 的文本特征选择方法[J].微计算机信息,2006,22(08-3):24-26
[24] 巩知乐,张德贤,胡明明.一种改进的支持向量机的文本分类算法[J].计算机仿真,2009,26(7):164-167
[25] 张东礼,汪东升,郑纬民.基于 VSM 的中文文本分类系统的设计与实现[J].清华大学学报: 自然科学版,2003,43(9):1288-1291
[26] 庞剑锋,卜东波,白硕.基于向量空间模型的文本自动分类系统的研究与实现[J].计算机应用研究,2001,9:23-26
[27] 张宁,贾自艳,史忠植.使用 KNN 算法的文本分类[J].Computer Engineering,2005,31(8):171-172,185
[28] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学,2006,17(9):1848-1859
[29] 顾犇.信息过载问题及其研究[J].中国图书馆学报,2000,26 (129):42-45,76
[30] 黎星星,黄小琴,朱庆生.电子商务推荐系统研究[J].计算机工程与科学,2004,26(5):7-10
[31] 马文峰,高凤荣,王珊.论数字图书馆个性化信息推荐系统[J].现代图书情报技术,2003(2):16-18
[32] 高凤荣,马文峰,王珊.数字图书馆个性化信息推荐系统研究[J].情报理论与实践,2003,26(4):360-362
[33] ZHOU T,REN J,MEDO M,et al,Bitpartite network projection and personal recommendation[J]. Phys Rev E,2007,76(4):046-115
[34] LINDEN G,SMITH B,YORK J.Amazon.com recommendations: Item-to-item collaborative filtering[J].IEEE Internet Computing,2003:76-80
[35] 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展[J].自然科学进展,2009,19(1):1-15
[36] 吴丽花,刘鲁.个性化推荐系统用户建模技术综述[J].情报学报,2006,25(1):55-61
[37] 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法[J] .软件学报, 2003,14(9):1621-1628
[38] 周军锋,汤显,郭景峰.一种优化的协同过滤推荐算法[J].计算机研究与发

展,2004,41(10):1842-1847
[39] 王辉,高利军,王听忠.个性化服务中基干用户聚类的协同过滤推荐[J].计算机应用,2007,27(5):1225-1227
[40] 李聪,梁昌勇,马丽.基于领域最近邻的协同过滤推荐算法[J].计算机研究与发展,2008,45(9):1532-1538
[41] 郑勇涛,刘玉树.支持向量机解决多分类问题研究[J].计算机工程与应用,2005,8(23):190-192
[42] 姜鹤.SVM 文本分类中基于法向量的特征选择算法研究[D].上海交通大学硕士学位论文,2010
[43] 张宁.基于语义的中文文本预处理研究[D].西安电子科技大学硕士学位论文,2011
[44] 吴国进.基于支持向量机的文本分类研究[D].安徽大学硕士学位论文,2011

攻读硕士学位期间发表的论文

1. 李春秋，何军，基于数据挖掘技术的高校学生成绩管理研究[J]. 宿州学院学报,2013，28（2）:79-82
2. 李春秋，电子商务推荐系统研究. 内江科技[J]，2013，（4）:166,202
3. 李春秋，个性化图书推荐系统研究. 河南科技[J]，2013，12

特别声明

本学位论文是在我的导师指导下独立完成的。在研究生学习期间，我的导师要求我坚决抵制学术不端行为。在此，我郑重声明，本论文无任何学术不端行为，如果被发现有任何学术不端行为，一切责任完全由本人承担。

学位论文作者签名：李春秋

签字日期：　2013 年 4 月 22 日