浙江大学计算机科学与技术学院

硕士学位论文

资源自适应个性化新闻推荐系统的研究与实现

姓名：唐朝

申请学位级别：硕士

专业：计算机应用技术

指导教师：卜佳俊;王灿

20100301

摘要

随着互联网的迅猛发展，网络数据量不断增长，在给网络用户获取信息带来便利的同时也造成了信息过载问题。我国上网用户数量达到 3.84 亿，占人口总数的 29%以上，其中超过 80%网络用户使用网络新闻资讯服务。个性化新闻推荐系统可以为用户推荐个性化新闻资讯，帮助用户发现感兴趣的内容，具有广阔的应用前景。

目前市场上的新闻推荐系统通过让用户回答问题或者主动定制新闻的方式实现个性化，其个性化特性不够完善。本文提出了一种基于隐式反馈的用户建模方法，该方法通过用户终端收集用户使用记录并进行分析，建立基于加权关键词表示的多模型用户模型。该模型可以同时反映出用户的长期兴趣和短期兴趣，并可以实时自动更新。

学术界对推荐系统的研究主要集中在推荐效果上，但是对于实际应用，推荐系统的性能也十分重要。本文提出一种资源自适应的推荐算法，使系统在推荐效果和系统性能之间取得了动态平衡。该方法监测系统资源使用情况，通过调整新闻时间窗口、文档向量维度和用户模型关键词维度的方式来动态调整相似度计算的运算量。当访问压力增大时，系统降低运算量来提高系统性能。

本文实现了上述算法并设计了个性化新闻推荐系统 EagleNews。该系统从国内知名新闻网站采集新闻，使用基于内容的推荐方法，以网络服务的方式运行于互联网上，可以使用通用的 HTTP 通信协议和标准的数据交换格式 XML 进行访问。实验表明系统取得了良好的效果。

关键词：个性化推荐，资源自适应，用户建模，长期兴趣，短期兴趣

Abstract

With the rapid development of Internet, the amount of the Web data is keeping increase. It is difficult for users to find the valuable information which attracts them in large-scale data. Additionally, there are 0.38 billion Web users in China, amounting for 29%, and more than 80% of Web users read Web news. Personalized recommendation systems are designed to provide personalized news to users. However, these systems are in primary stage in China.

Firstly, available news recommendation systems are not good enough currently, since they can merely recommend news to users by asking users some questions. This paper proposes a recommendation method based on implicit feedback from user-generated profile. It gathers user behavior through client software and creates a multi-model user profile by analyzing these records. It considers both the short-term and long-term user interests and updates these interests automatically.

Secondly, traditional researches of recommendation system mainly focus on recommending quality. However, the efficiency problem is very crucial in practical applications. This paper proposes a resource-adaptive algorithm to address this problem, which tries to balance the precision and the efficiency of the system. It monitors the free resource of the system and adjusts a sliding time window on news stream accordingly. This way, the dimensions of document vector and user profile vector change dynamically. When system load increases, the system can still reduce the computing time.

Finally, this paper implements the algorithms and designs the news recommending system EagleNews. The system gathers news from famous news websites in China, and recommends news using the content-based method. Furthermore, it is a standard web service by using HTTP protocol and XML as its data exchange way. The results of experiments show that the performance of the systems is quite good.

Keywords: recommendation system, resource-adaptive, user modeling, long-term interest, short-term interest

图目录

表目录

浙江大学研究生学位论文独创性声明

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。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，也不包含为获得浙江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。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解浙江大学有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江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检索和传播，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：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

第1章 绪论

我国互联网历经二十多年的发展，普及率已经超过四分之一，而且网民的数量还在高速稳定增长中。互联网已经越来越多地深入到了人们的工作、学习、娱乐等日常生活当中。由于具有低成本、高覆盖面和实时传播的特性，互联网已经成为当今信息传播的主要途径之一，也因此，通过访问互联网来获取资讯信息也成为人们获取资讯信息的主要手段之一。

由于互联网的优越特性，在其上发布信息极为便捷，这就使得互联网上的信息数量以近乎爆炸的速度增长。我国的新闻资讯类网站数量数以万计，如果再加上非专业的个人维护的资讯类网站，更是多不胜数。由于用户各自的兴趣爱好不同，应用和目的也有差异，所以他们对信息资讯的需求千差万别，但是，面对如此海量的信息，用户往往不能够有效地获取高质量的信息。个性化服务技术可以为每位用户单独定制以提供不同的服务，将该项技术应用于新闻资讯信息的推荐，可以使用户更有效地获得新闻资讯。然而，目前对推荐系统的研究往往偏重推荐效果方面，而商业应用系统，则对推荐系统的性能也有着较高要求。

鉴于此，有较高性能的网络新闻资讯推荐系统的研究势在必行，以期能够为用户提供实时、准确的新闻资讯推荐服务。

1.1 研究背景

1.1.1 个性化新闻推荐技术的产生

中国互联网络信息中心（CNNIC）于2010年1月发布的《中国互联网络发展状况统计报告》显示，截止到2009年12月底，中国网络用户数量达到3.84亿人，互联网普及率达到29.6%。互联网用户数量还在持续增加，普及率也在平稳上升。图 1-1 显示了最近几年来我国网络用户规模和互联网普及率都在不断增加的变化趋势。

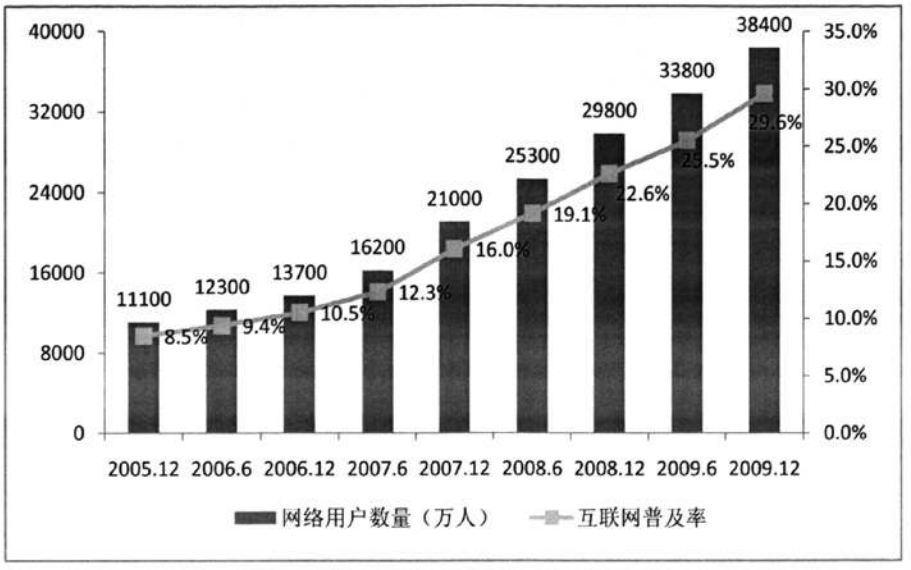

图 1-1 中国网络用户规模和互联网普及率

用户使用互联网主要是为了满足信息获取、交流沟通、学习娱乐和商务交易的需要，可以说，互联网已经深入到了生活的方方面面。用户经常使用的网络服务多种多样，数据显示，网络新闻的使用率稳定在一个很高的水平，占 80.1%，较上一次统计又有所增长，是用户使用最多的网络服务之一，仅次于网络音乐，位于第二位，而绝对使用人数也在高速增长，具体数据参见图 1-2。

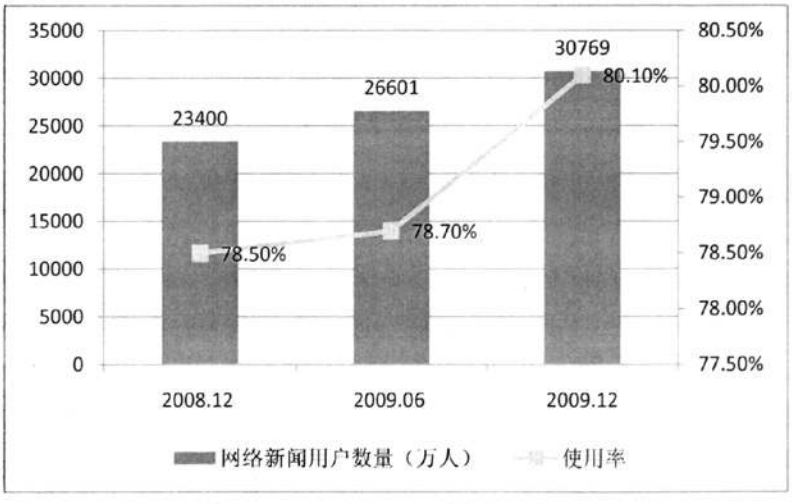

图 1-2 2008.12-2009.12 网络新闻用户对比

由于网络新闻资讯具有庞大的用户群体，所以成为各类门户网站和内容型网站必备的板块之一。网络新闻资讯类网站在整个互联网中占据了举足轻重的地位，

是其重要组成部分之一。典型的综合型新闻门户类网站有新浪、网易、新华、腾讯、搜狐、东方网等等，除去这些知名的大型综合型门户网站，还有地方性门户网站，音乐、汽车、军事、财经、IT 等行业领域新闻网站。在互联网上，网络新闻资讯网站无处不在。

大量的新闻网站为人们提供了充足的信息资讯，然而与此同时，也为用户带来了新的问题和挑战。网络新闻类网站一般由商业公司、组织机构或者个人经营，由于各个经营主体有着不同的社会资源、媒体资源、不同的专业和行业背景，导致各个网站开办的栏目良莠不齐，有些网站以体育新闻见长，而另一些则以国际政治见长，而专业性较强的网站则干脆专注于一个细分类别的新闻，如 cnBeta[1]，只发布 IT 业界新闻。而用户往往不会只喜好或需要一个单一分类的新闻资讯，而是一般都同时关注数个类别的新闻，这就需要用户同时关注数个网站的更新情况，而如此做法，无论是时间上，还是精力上，都会给用户带来负担。

另一个方面，新闻类网站因为成本、版面等原因的限制，往往只能把新闻资讯划分为数个或十数个分类，而这样的分类粒度往往不能满足用户的需求，比如一位球迷可能只关注体育类新闻中的足球类，而在足球中又对四大联赛感兴趣较多，对中国足球兴趣索然。这就造成用户需要在大量的新闻中寻找自己感兴趣的新闻，浪费了时间。

个性化服务技术可以有效地解决上述问题。所谓个性化服务就是为不同的客户提供有针对性的服务，以满足其独特的需求。个性化推荐系统是个性化服务的典型应用，这类系统一般会为每个用户建立单独的用户模型或者用户概貌，并以此为依据预测用户可能感兴趣的信息或者商品推荐给用户。

个性化新闻推荐系统，就是将个性化推荐技术，应用于新闻资讯的推荐，它可以帮助用户轻松获得自己的感兴趣的新闻资讯信息，并从互联网上海量信息中，发掘可能感兴趣的内容，并且不需要用户花费时间去寻找，可以为用户节约大量的时间。

个性化新闻推荐系统在国外已经有了成功的案例，但是在我国，其研究才起步不久，成熟的应用系统尚不多见，已有的新闻资讯推荐系统，大都是通过与用户进行简单交互，为用户定制新闻内容板块，并不能自动学习用户的兴趣，其推

[1] http://www.cnbeta.com cnBeta 网的首页

荐效果并不十分理想。

除此之外，无论是互联网上的信息还是网络新闻的用户数量，都十分巨大，且增长速度快，这种特性对个性化新闻推荐系统的性能提出了较高要求，一个实用的个性化新闻推荐系统必须是能够适应较大的新闻数据量和用户访问量的产品。

1.1.2 个性化推荐技术的应用现状

由于个性化新闻推荐系统有着很高的应用价值，国内外许多机构和公司都进行过相关的技术研究和应用开发。但是其中很多停留在理论层面和原型系统的研发，实际投入运营的为数不多。

Digg[2]是创始于 2004 年，是美国一家非常著名地用于推荐网络新闻、图片、视频的网站。该网站并自身不发布内容，而是发动用户将他们发现的内容链接推荐到该网站，并允许用户继续推荐其他用户推荐过的内容，从而将更有趣的内容推荐给用户。该网站在性质上属于协同过滤的推荐服务，但是更强调用户的作用，个性化方面并不突出。

图 1-3 Findory 的首页

Findory[3]是一个个性化的电子报纸网站，可以算得上是真正意义的个性化新闻推荐网站，用户在该网站阅读新闻，系统在用户的使用过程中学习出用户的兴趣，并以此为依据为用户创建一个新闻头条页面。Findory 使用的是混合的协同过滤推荐算法，它将与口味相似的用户阅读过的文章相似的文章也推荐给用户。

豆瓣[4]是国内最大也是最成功的综合型个性化推荐网络服务之一，该网站记

[2] http://digg.com Digg.com 的首页

[3] http://glinden.blogspot.com/2008/01/brief-history-of-findory.html Findory 的特性介绍文章

[4] http://www.douban.com 豆瓣网的首页

录每个用户读过的书籍、看过的电影、听过的音乐以及用户为这些项目的评分和评论，然后为每个用户推荐相似的内容，或者其他好友喜爱的内容。除此之外，豆瓣还开通了博客文章的推荐功能，也是以协同过滤为主的推荐算法。

图 1-4 豆瓣网首页为用户提供电影和书籍推荐

百度新闻搜索[5]是百度公司推出的新闻推荐服务，该服务上线于2003年7月，每天发布新闻数量在150,000到160,000条。该服务本身并不发布或转载新闻内容，仅发布新闻标题和链接，其全部新闻收集于其他的新闻门户网站。百度新闻搜索的首页按照新闻内容分类将页面分成若干板块，每个板块显示若干焦点新闻。

该项服务还提供个性化新闻服务，允许用户以两种方式定制新闻：1）按照关键词定制；2）按照地区定制。用户定制后，会在页面上显示一个板块，该板块内容为用户定制的内容。百度新闻搜索采用的是静态用户建模方法，其个性化特性不能自动识别用户的兴趣，也不能发现用户兴趣的变化，都需要用户手动调整。

Google资讯[6]是谷歌公司推出的新闻推荐服务，其性质与百度新闻搜索相似，从1,000多个新闻资讯来源中获得新闻，并允许用户个性化定制。Google资讯允许用户按照三种方式定制板块：1）按照新闻分类；2）按照地区；3）按照关键

[5] http://news.baidu.com/ 百度新闻搜索主页

[6] http://news.google.cn/ 谷歌资讯首页

字。与百度新闻搜索相似，该服务业仅支持静态的用户定制。

1.2 研究目的及意义

本文研究的课题为，资源自适应的个性化网络新闻推荐系统的研究与设计，该系统主要能够实现自动学习用户兴趣，并且其性能能够满足实际应用需要。在理论研究和实际运用方面都有着一定的意义：

1.2.1 研究目的

本文主要的研究目的可以归纳成以下几点：

1) 为每个用户建立用户模型，在该推荐系统中，使用比较灵活高效的用户建模方法建立模型。该用户建模方法可以准确挖掘用户阅读兴趣，既可以快速反应出用户短期兴趣的变化，又可以反应用户稳定的长期兴趣；

2) 资源自适应特性，该系统要有较高的性能，能够容纳较大的新闻数量，如几十个甚至上百个新闻源的数据量，又能允许较多的用户同时访问，与此同时在新闻数量和用户访问量发生突变时，能够自适应地调整系统参数，保持系统性能稳定，同时又不过分牺牲推荐效果。

3) 设计与实现一个完整的个性化新闻推荐系统，该系统可以提供通用的个性化服务，该通用体现在推荐结果的格式通用，能够推荐的内容通用（可以推荐新闻、博客、资讯、电子书等）；

1.2.2 研究意义

个性化网络新闻推荐系统的研究与实现，无论是在理论研究还是在实际应用领域都有着积极的意义：

1. 理论意义

个性化推荐领域的学术研究，大多集中在对推荐效果的提升，如推荐的准确率、全面性、新颖性等方面，而较少有研究工作涉及到推荐系统的性能方面。究其原因，多数研究都是较为偏重理论，不以实际运行的商业系统为研究目标，故而也没有地专注提高系统效率的需求。

本课题的目标是设计与实现一个高性能的个性化推荐系统。该系统因为课题

的需要，必须是能够承受一定访问压力的系统。所以，本课题的研究，在为个性化推荐系统在提升性能方面的研究提出了有意义的探索。

2. 应用前景

目前，全球信息化程度越来越高，互联网的普及率也是越来越高。我国在互联网建设上投入逐渐增多，互联网逐渐成为我国又一大产业。各行各业对互联网的依赖也因此逐渐提高。个性化推荐系统已经遍布了国内的网络，如推荐书籍、电影、CD等的豆瓣网；推荐电影的Mtime时光网；推荐视频的土豆、酷六、优酷等视频分享网站；如淘宝、易趣、拍拍、有啊等网络购物网站更是使得个性化推荐系统大放异彩。

然而，面向文本类内容的新闻资讯个性化推荐系统，却并不多见。目前国内最大的个性化新闻推荐系统是如百度新闻搜索，谷歌新闻资讯等网站。这些新闻推荐系统，大多只能为用户提供一定程度的内容定制，如按照地区、频道进行定制，却不能随着用户的使用来动态调整推荐的新闻，其推荐效果还可以在此基础上有所提高。

除此之外，由于残疾人中有大量的视障人群，无法像普通人一样轻松地获取信息资讯，享受互联网带来的便利，个性化网络新闻推荐系统的研制，配合为视障人群专门设计的终端设备，还可以为视障人群获取互联网上的新闻资讯信息提供极大的便利。

本课题设计与实现的个性化新闻推荐系统，一方面填补了国内没有新闻资讯推荐系统的空白；另一方面，本课题研究的系统，有着较好的扩展性。能够根据市场的需求，灵活的调整系统的功能。并可以简单地与其他系统集成在一起，构建丰富的应用，如可以将新闻、博客、电子书等都纳入到推荐系统中，成为被推荐的对象。

由于本课题所研究的系统，为每个用户都建立独特的用户模型，所以也具备了向每个用户有差别的推送特定信息的能力，可以在日后拓展成为广告、商务咨询等内容的推荐系统，有着广泛而深远的应用前景和现实意义。

1.3 本文组织

第1章是绪论，介绍了个性化新闻推荐技术的研究背景和本课题提出的动因，还介绍了本文的研究目的和意义，最后介绍了本论文的组织。

第2章介绍了个性化推荐技术的概念和原理，并将个性化推荐系统按照推荐方法的不同分成三类，逐一展开介绍，并在此基础上讨论了目前世界上主流的推荐系统研究项目，以及学术和工业界在提高推荐系统性能方面做出的努力。

第3章介绍了用户建模的概念和本文中提出的系统使用的用户建模方法，该方法通过隐式信息收集方式，为用户建立多模型用户模型。使用该方法建立的用户模型能同时兼顾用户的短期兴趣和长期兴趣，使得系统的推荐结果能够及时反应户兴趣的变化。此后，通过实验确定了用户建模方法的参数设定，并通过用户模型实例分析验证了该方法的有效性。

第4章首先分析了个性化新闻推荐系统使用的推荐算法以及选择此算法的原因，提出了优化推荐算法性能的几个方法，最后，提出了资源自适应的推荐系统模型，给出了该系统模型的原理及其实现方法，使用该模型的推荐系统，可以兼顾到推荐性能和推荐效果，并动态地保持平衡。

第5章介绍了以前两章技术为基础构建的实例系统EagleNews。分析该系统的特点和体系结构，并逐模块说明了设计的细节和关键问题的解决办法。

最后，第6章总结了本文的工作及对此问题研究的贡献，并对今后的研究方向提出了展望。

第2章 个性化推荐系统技术综述

在这个信息爆炸时代，信息过载[1]已经成了用户在使用互联网过程中所面临的一个主要的问题，虽然搜索引擎可以帮助人们在众多信息中检索到自己想要的信息，解决了主动检索网络信息的难题，但是自动从网络上获得自己感兴趣的信息的问题，却没有得到解决。个性化推荐系统的研究始于上个世纪九十年代中期[2]，最初的推荐系统用于帮助人们从新闻讨论组中过滤出自己感兴趣的话题。由于可以协助用户发现网络上的内容，向用户推荐他们感兴趣的事物，个性化推荐系统在书籍、音乐、电影乃至交友等越来越多的领域大展宏图，此外，该技术也被广泛应用于电子商务领域，不但可以推荐商品、广告等信息，还可以用于将买卖双方互相推荐，有着巨大的商业价值。因此，至今十五年过去了，个性化推荐系统研究领域依旧是当今的研究热点，工业界与学术圈在该领域的研究都十分活跃。

虽然，推荐系统得到了广泛的应用，在互联网上几乎无处不在，但个性化推荐技术之中还有许多有待改进之处。一方面是要提升推荐系统的效果，而另一方面就是提升推荐系统的性能。对于前者，近几年来，涌现出了很多卓有成效的研究[3, 4, 5]。而后者，学术界则涉足不多，但是并不是说推荐系统的性能没有研究价值，而是多数研究机构和学者，因为资源的限制而无法开展有效的研究。而要让推荐系统能够投入运营，这方面的研究必不可少，所以工业界对此领域进行了许多有益的探索[6]。

本章内容首先会对个性化推荐技术展开综述，介绍个性化推荐系统研究的问题，系统的分类方法，以及各个类型推荐系统的特点与局限性。然后，将展开对个性化推荐技术性能问题的探讨。

2.1 个性化推荐技术

个性化推荐系统作为一个独立的研究领域，开始于上世纪九十年代中期，是一个非常年轻的研究领域。推荐系统研究的核心问题，是预测一个用户对其未看到的事物的喜好程度，一般用户对其查看过的被推荐项进行打分来表明对其的喜

好程度，所以推荐问题就可以抽象成预测用户对其未见过的事物评分的问题。预测出用户对其未见过的事物的评分高低，就可以以此为依据，将得分高的事物推荐给用户。

为了便于研究，将这个问题进一步地形式化：在一个推荐系统中，用C表示所有用户的集合，用S表示所有事物的集合。一个用户对一个事物的喜好程度，可以表示为二元关系$C\times S$到一个全序集合R上的映射：

$$u:C\times S\rightarrow R \qquad \text{公式 (2.1)}$$

将该映射定义为一个函数u，集合C中的用户数量一般非常庞大，每个用户，一般使用一个数据结构来表示，称为用户概貌或者用户模型，用户模型标识了用户的特性，如性别、年龄、爱好，又或者用户感兴趣的关键词等。集合S中的事物数量也非常庞大，每个事物，也用类似的数据结构来表示，标识了事物的特性，如一张CD的专辑名称、歌手姓名，或者一部电影的导演、主演以及主要情节内容等。

一般在推荐系统中，用数值来衡量用户对一个事物的喜好程度，所以集合R的元素一般为一个正整数区间或者是实数区间。推荐问题，就是对于用户$c\in C$，找到使得$u(c,s)$取值尽可能高的事物s'_c：

$$\forall c\in C, s'_c=\arg\max_{s\in S}u(c,s) \qquad \text{公式 (2.2)}$$

问题的难点在于，用户仅能对自己看到过的事物进行评分，没法对未见过的事物进行评分，所以，要实现推荐，首先就要预测用户对未看到的事物的评分，而仅有的参照就是用户对已经看见的事物的评分情况。

表 2-1 一个电影推荐系统中的评分矩阵

用户	《刺陵》	《孔子》	《阿凡达》	《十月围城》
张三	4	3	4	4
李四	?	4	5	5
王五	2	2	4	?
陈六	3	?	5	2

如表 2-1 中数据所示，是一个电影推荐系统的评分矩阵，用户只能对已经观看过的电影进行评分，而系统则需要根据已经存在的评分，来预测出用户对未观看过的电影的评分情况，如表中“?”所示的部分。

预测用户对事物评分的方法有很多种，推荐系统一般依据此来进行分类，通

常来说，可以分为三类：

- **基于内容的推荐系统**——将与用户曾经喜好的事物相似的事物推荐给用户；
- **协同过滤的推荐系统**——将与用户相似的用户过去表示喜好的事物推荐给用户；
- **混合的推荐方法**——将上述两类方法按照某种方式结合的推荐方法；

2.1.1 基于内容的推荐方法

基于内容的推荐方法基本原理是，利用用户c过去对事物s'的评分$u(c,s')$来预测用户对于其未见过的事物s的评分，其中s'是与s相似的事物。例如，用户过去喜爱某部电影，则现在如果一部新电影与过去的电影有着相同的导演、主要演员或者题材，用户可能也会喜欢。该推荐方法起源于信息检索[7]和信息过滤[8]的研究。较常见应用于文本类信息的推荐系统，如文档、网页和新闻推荐系统等。

形式化地描述基于内容的推荐过程，就是首先将待推荐的对象s，表示成一种数据结构$Content(s)$，如上文中所说，常见的文本信息推荐系统，将待推荐文档d_j表述成一个关键词k_i的集合。这个集合中的关键词依据其重要性的不同，带有不同的权值w_{ij}。一个计算权值的经典方法，就是词频逆文档频率（TF/IDF）方法[9]。N 为所有待推荐文档集合的总数，关键词k_i在其中n_i个文档中出现过，假设$f_{i,j}$为关键词k_i在文档d_j中出现的次数，则关键词k_i在文档d_j中的词频$TF_{i,j}$定义为：

$$TF_{i,j}=\frac{f_{i,j}}{\max_z f_{z,j}} \qquad \text{公式 (2.3)}$$

其中，分式中分母为任意关键词k_z在文档d_j中出现的次数的最大值。一般来说词频能够代表一个关键词的重要性，但是如果一个词在所有文档中都出现，则该次的意义就相对平凡，所以经常使用逆文档频率与词频结合，来评估一个关键词的重要性，关键词k_i的逆文档频率被定义为：

$$IDF_i = \log \frac{N}{n_i} \qquad \text{公式 (2.4)}$$

关键词 k_i 的 TF/IDF 权重 w_{ij} 就被定义为：

$$w_{i,j} = TF_{i,j} \times IDF_i \qquad \text{公式 (2.5)}$$

然后 $Content(d_j)$ 就可以用向量表示为：

$$Content(d_j) = (w_{1j}, w_{2j}, ..., w_{kj}) \qquad \text{公式 (2.6)}$$

然后，用户 c 过去已经评分的事物集合可以表示成 $ContentBasedProfile(c)$，基于内容的推荐方法，就是要预测用户 c 对 s 的评分：

$$u(c,s) = score(ContentBasedProfile(c), Content(s)) \qquad \text{公式 (2.7)}$$

继续使用上面的例子，假设要对网页文档进行推荐，则 $ContentBasedProfile(c)$ 和 $Content(s)$ 都可以表示成加权关键词的向量 $\vec{w}_c$ 和 $\vec{w}_s$，则 $u(c,s)$ 经常使用某些启发式评分算法，如余弦相似度算法[10]：

$$u(c,s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\|\vec{w}_c\|_2 \times \|\vec{w}_s\|_2} \qquad \text{公式 (2.8)}$$

除了传统的信息检索的方法，还有其他的启发式算法被用来预测评分，如贝叶斯分类器[11]，各种机器学习方法如聚类、决策树和人工神经网络等[12]。

基于内容的推荐方法有几个优点：推荐不依赖其它用户的数据，没有协同过滤推荐的冷启动和稀疏性问题；能立刻将新项目或旧项目加入推荐，没有新项目问题；由于文本处理技术和信息检索技术的成熟，对于文本类内容的推荐有相当的优势。

该推荐方法也存在一定的局限性，如对于无法通过机器理解，或难于对其提取特征的内容，无法进行有效推荐，如音视频流等；过度特性化，由于仅通过分析用户的使用习惯来进行推荐，则用户未曾访问过的内容，就不会推荐给用户，使得用户失去的了拓展视野发掘新兴趣的机会。

2.1.2 协同过滤的推荐方法

协同过滤推荐方法的本质是，利用与用户 c 相似的用户 c' 对用户 c 未见过的

事物s的评分$u(c',s)$来预测用户c对s的评分$u(c,s)$。例如，一个用户的朋友都喜欢电影《阿凡达》，那么这个用户虽然没有看过这部电影，但是也会对这部电影产生浓厚的兴趣。协同过滤算法的发展时间不长，这个算法的根本思想来自于人们之间自然而然地交流想法的行为，它的基本假设就是兴趣相近的人们对相同的事物评分相近。

最早的应用协同过滤技术的系统，如 Grundy[13]，Tapestry[14]等，将用户对事物的评价记录下来，并允许用户检索这些记录来获取自己感兴趣的内容，属于被动的协同过滤，也即用户必须主动的利用其他用户的评价。此后，GroupLens 系统[15][16]，视频推荐[17]，和 Ringo[18]系统才真正实现了自动化的协同过滤推荐，也即系统自动为每个用户预测其可能感兴趣的被推荐项。

根据[19]，协同过滤的推荐系统可以分为两种类型，基于存储的（Memory-based）和基于模型的（Model-based）两种。基于存储的协同过滤系统使用所有已经被用户评分过的项目的集合来进行新项目评分的预测。而基于模型的系统与此不同，是利用用户已经进行评分的项目来学习出一个模型，最后根据这个模型来进行推荐。

由于在计算过程中，要使用全部的历史数据进行计算，纯粹的基于存储的方法并不适合在实际系统中使用，因为在实际商业应用时，存储资源，计算资源都是十分有限，无法随意调用。所以，在实际的协同过滤推荐系统中，都或多或少的会进行提前计算，生成一定的模型。目前，主流的协同过滤系统，不是纯粹的基于模型的系统，就是基于存储与基于模型的混合系统。因此，在本节，我们把协同过滤方法，分为概率方法和非概率方法来进行讨论。

2.1.2.1 非概率方法

最近邻算法是协同过滤算法中最为经典的算法之一。该算法主要有两种：面向用户的最近邻算法和面向被推荐项的最近邻算法。面向用户的最近邻算法的思路非常简单，就是利用了协同过滤算法的最基本的假设，使用与用户c相似的用户对项目s的评分来预测用户c对s的评分：

$$r_{c,s} = \mathrm{aggr}_{c' \in \hat{C}} r_{c',s} \qquad \text{公式 (2.9)}$$

在上面的公式中，$\hat{C}$表示用户c的N个邻居的集合。aggr代表一个聚合的函数，用于将$\hat{C}$中用户对s的评分聚合在一起。具体的聚合往往采用启发式的算法，这里列举几个简单的例子：

$$(a)\ r_{c,s}=\frac{1}{N}\sum_{c'\in\hat{C}}r_{c',s}$$
$$(b)\ r_{c,s}=k\sum_{c'\in\hat{C}}\mathrm{sim}(c,c')\times r_{c',s} \qquad \text{公式 (2.10)}$$
$$(c)\ r_{c,s}=\bar{r}_c+k\sum_{c'\in\hat{C}}\mathrm{sim}(c,c')\times(r_{c',s}-\bar{r}_{c'})$$

（a）是最简单的聚合方法，将N个最近邻居对s的评分求平均数，作为对$u(c,s)$的预测值。由于用户与其邻居相似程度不可能都一致，所以，加权平均的方式，相比简单平均来说，可能更为合理，所以在（b）中，使用了加权平均的方式，其中k为归一化参数$k=1/\sum_{c'\in\hat{C}}|\mathrm{sim}(c,c')|$。（b）的方法基本上已经比较完善，但是还有一些不足之处，用户对一个事物进行打分时，可能使用了不同的分值区间，这样的话用（b）的公式，可能就无法得到准确的值。于是改进成了（c）的样子，这个公式将每个用户的评分与平均值的偏差聚合，然后将结果加到平均值之上，公式中$\bar{r}_c$可以这样定义：

$$\bar{r}_c=\frac{1}{|S_c|}\sum_{s\in S_c}r_{c,s},\text{其中}S_c=\{s\in S_c\mid r_{c,s}\neq\varnothing\} \qquad \text{公式 (2.11)}$$

在上述的公式中可以看出，如果要得到更为准确的预测值，则判断两个用户之间的相似度也是一个关键。绝大多数的方法，判断两个用户相似程度，都依赖于两个用户共同评分的项目，两个较为常见的方法是相关系数法和余弦法。使用户x和y都进行过评分的项目的集合为S_{xy}，相关系数法计算两个用户相似度的方法为：

$$\mathrm{sim}(x,y)=\frac{\sum_{s\in S_{xy}}(r_{x,s}-\bar{r}_x)(r_{y,s}-\bar{r}_y)}{\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-\bar{r}_x)^2\sum_{s\in S_{xy}}(r_{y,s}-\bar{r}_y)^2}} \qquad \text{公式 (2.12)}$$

而在余弦法中，则将用户x和用户y共同评分的项目的分值表示为两个m维向量，其中$m=|S_{xy}|$，则用户x和y的相似度可以这样计算：

$$\mathrm{sim}(x,y)=\cos(\vec{x},\vec{y})=\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|_2\|\vec{y}\|_2}=\frac{\sum_{s\in S_{xy}} r_{x,s}r_{y,s}}{\sqrt{\sum_{s\in S_{xy}} r_{x,s}^2}\sqrt{\sum_{s\in S_{xy}} r_{y,s}^2}} \qquad \text{公式 (2.13)}$$

在不同的系统中，用来判断用户相似度的具体算法略有不同，以上介绍的只是最基础的算法。实际在应用过程中，系统一般事先计算好所有用户的相似度保存在数据库中，当用户有了新的评分操作后，再更新用户的相似度值，这样才能保证在推荐时，可以高效地计算。

2.1.2.2 概率方法

协同过滤的概率方法直接通过概率预测用户 c 对 s 的评分。大多数概率的协同方法计算用户 c 对项目 i 评分为 r 的概率 $p(r|c,s)$，然后用概率最高的评分值或者 r 的期望值作为 $u(c,s)$ 的预测值：

$$r_{c,s}=\mathrm{E}(r_{c,s})=\sum_{i=0}^{n} i\times\Pr(r_{c,s}=i\,|\,r_{c,s'},s'\in S_c) \qquad \text{公式 (2.14)}$$

在上述公式中，假设用户评分的范围是从 0 分到 n 分，后面的概率表达式，表示在已知用户先前对其他事物 s' 的评分的情况下，会对 s 评定某个特定分值的概率，要估算这个概率值，有两个常见的模型：聚类模型和贝叶斯网络[19]。

在第一种方法中，首先将兴趣相似的用户聚类到相同的分类下，假设用户的评分行为都是独立的，用户类别的数量和参数都是从数据中学习得来；第二种方法将每个待推荐对象都表示成一个贝叶斯网络的节点，节点的每个状态，对应着评分范围内的一个分值。网络结构和条件概率都可以从以前的用户数据中学习得到。

这类方法有一个局限性，那就是用户只能被分到一个类别中，而实际上，用户如果可以同时属于多个类别，还可以提高用户的使用体验，简单举个例子，在一个图书推荐系统中，一个用户可能即对自己工作相关的内容感兴趣，也同时对休闲内容如旅游等感兴趣。

与基于内容的推荐方法相比，协同过滤的推荐方法有着不少优势：

1. 更好的兼容性，因为主要是依靠用户评分进行推荐，所以对被推荐事物本身的特性没有要求，可以用于特征难于提取的音乐、电影等媒体流或

艺术品等。

2. 在用户之间分享知识，避免了基于内容推荐系统容易产生的过度特化的问题。对于用户曾经没有涉足的领域，也有被推荐给用户的机会。

同样，协同过滤系统，也存在着其局限性：

1. 新用户问题，由于要依赖口味相似的用户的使用数据，对于一个系统的全新用户来说，难以产生有效的推荐结果；
2. 新待推荐项的问题，该问题与新用户问题对称，新的待推荐事物加入到系统后，由于没有任何用户对齐评分，则该事务难以被有效推荐给用户；
3. 稀疏问题，对于待推荐事物集合数量庞大的情况下，协同过滤的推荐系统难以有效运转，由于数量太过庞大，导致同一事物被口味相似用户评分的概率降低，在系统进行推荐时，失去了参照。

2.1.3 混合推荐方法

由于基于内容的推荐方法和协同过滤推荐方法都有各自的局限性，所以，很多个性化推荐系统都采用二者的混合推荐方法，来弥补互相之间的缺陷。不同的混合推荐系统，使用的混合方法也各不相同，概括起来，常见的混合方法有如下几种：

1. 加权混合

加权混合是一种极为自然的混合方式，该方法将两种推荐方法为被推荐项测的评分，按照一定权重组合，作为被推荐项新的预测分值，并重新排序，以此提高推荐效果。P-Tango 系统[20]正是一种以线性组合方式加权混合的推荐系统。而 Pazzani 提出的推荐系统[21]，则采用类似投票的机制，来加权混合两种方法预测出的分值。

2. 切换选择

切换选择是在两种推荐方法中，选择结果较好的一方作为最终的推荐结果。例如 DailyLearner 系统[22]，首先选用基于内容的推荐方法，如果不理想，再尝试协同过滤的推荐方法；而 Tran 和 Cohen 提出的方法[23]则按照推荐结果与用户先前评分情况的吻合程度高低来选择较好的一方作为结果。

3. 结果混合

在需要大量推荐结果的系统中，有时候也会采用结果混合的方法，如 PTV 系统[24]，首先利用电视节目附带的文本信息进行基于内容的推荐，然后利用用户偏好信息进行协同推荐，然后将两个推荐结果放在一起呈现给用户。

4. 特征结合

特征结合法是将协同过滤方法的结果作为额外的数据附加到每个待推荐项上，然后再使用基于内容的推荐方法，[25]将该方法应用于电影推进，得到了准确率明显好于纯协同过滤方法的结果，不过，这仅是在对每部电影的特征信息进行了手工过滤后的结果，如果应用每部电影全部的信息，则只能提高结果的查全率，不会提高结果的准确率。

5. 级联

级联的混合方法，是一个多阶段的方法，首先使用一种方法得到粗略的结果，然后使用另一种方法，从第一种方法预处理后的结果中，继续进行推荐。EntreeC[26]就是一个级联的推荐系统。

6. 统一模型

该种混合方法合并基于内容方法和协同过滤方法两者的特征，构建一个综合的统一的模型。例如基于规则的分类器[27]，统一概率方法[28]、[29]，贝叶斯 mixed-effects 回归模型[30]、[31]。

2.2 推荐系统的性能挑战

个性化推荐系统已经越来越多地深入到了人们的生活中，随着其应用范围越来越广，以及互联网上的信息数量日益增长。推荐系统所面临的挑战，日益显现。性能问题，就是推荐系统必须面临的一大问题。没有高效的推荐算法，个性化推荐系统，在未来将无法承受越来越大的计算压力。如何使推荐系统能够更好地适应时代的发展，是本文的一个研究课题。

2.2.1 推荐系统性能压力原因

造成推荐系统性能压力最直接的原因，就是推荐系统用户数量的激增和待推荐数据量的激增。拿个性化新闻推荐系统来说，在第 1 章已经分析过，我国网民数量快速增多，互联网普及率以超过 5%的速度高速增长，其中网络新闻用户数量在半年来更是激增了 13%以上。而美国 Google 公司的产品 Google News，几

天之中就有数百万的独立用户访问[6]。

而另一方面，网络新闻的数量也在爆炸性增长，谷歌资讯和百度资讯每天公布的新闻数量超过15万条，这还仅仅是中文新闻的数量，Google News[32]（英文）有超过4500个新闻源，Yahoo! News[33]有超过5000个新闻源，其新闻数量都在百万量级。

对于一个新闻推荐系统来说，这样庞大的数据量是对其性能的第一个挑战。

另一方面，各个个性化推荐系统的运营商之间竞争激烈，为了增加用户体验，他们需要不断提升推荐系统的效果，以得到更好的推荐准确率和全面性，这势必就造成了推荐算法复杂性地不断增加。

2.2.2 研究现状

虽然个性化推荐系统在性能方面面临着极大的挑战，但是先前关注此问题的研究并不多见，多数研究者仍旧是以提高个性化推荐系统的全面性和准确性作为研究目标。

出现这种状况有几个原因，一是因为学术研究用的系统，多数不需要投入商业运营，所以提升性能的需求并非十分迫切；另一方面，多数研究采用数据集进行研究，没有动态变化的数据，故也没有提升性能的需求；还有一个原因是学术研究领域用于研究性能问题的客观条件也远远不够，综上，个性化推荐系统研究领域，关注性能的研究并不多见。

个性化推荐系统的研究与信息检索和信息过滤的研究有着很大的相关性，在此相关领域，却有着一些卓有成效的研究，值得借鉴。

ONED[34]是一个用于在新闻流数据中检测新的新闻事件发生的系统，该系统采取了一系列的措施，用以提高系统的性能，使得系统能够应付更大的数据量。该系统提出了一个系统框架，在该框架中，系统通过调整计算参数，预处理，建立索引，新闻源排序，结果优先级等数个有效手段，使得系统能够在数据量增加的情况下，保持相对稳定的性能，并且其事件监测的结果也不会让用户感觉到数量上的压力。

除此之外，工业界也有一些研究，如Google News的个性化系统[6]，该系统中，采用了MinHash聚类，PLSI索引，以及使用MapReduce，将个性化推荐任

务进行分割，然后分布式运行来实现对大规模计算的优化。虽然此方法卓有成效，但是并不是适合于一般的公司和用户使用，这也存在一定程度上的客观原因。

2.3 本章小结

本章主要介绍了个性化推荐系统的研究起源，以及研究现状。对现有的个性化系统进行了分类，并分别介绍了每个类型的个性化推荐技术的原理和特点，以及现有的系统案例。个性化推荐系统按照推荐方法的不同，可以分成三类，基于内容的推荐系统、协同过滤的推荐系统和混合的推荐系统。此外，还分析了每个类型的个性化推荐系统的优点和其局限性。然后，本章分析了个性化推荐系统面临的挑战，重点介绍了推荐系统所要面临的性能的压力。目前，学术界在提高推荐系统性能方面的研究并不多见，本章最后介绍了与推荐系统研究领域使用方法类似的研究领域关于性能的研究。

第3章 用户建模与兴趣发现

个性化推荐系统可以分别为每个独立用户筛选其感兴趣的信息，为了做到这一点，系统必须了解每个用户的兴趣偏好。个性化系统一般都要通过某种途径，来收集用户的信息，通过分析整理这些信息，从中识别出用户的兴趣偏好，这个过程，称之为用户建模。

目前国内已经出现了很多的网络新闻门户网站，其中不乏能够提供个性化服务者，然而，这些新闻个性化服务，普遍是在用户首次使用时，通过询问用户一些简单问题，获取用户的回答，为用户建立永久的用户模型，对呈现给用户的新闻资讯做一定程度的定制，然而，这种静态化的用户建模方式所建立的用户模型，并不能准确反映出用户的兴趣，而且也不够灵活，无法兼顾到用户兴趣的变化。此外，这种与用户交互的建模方式还需要用户额外的操作，给用户带来额外的负担。

本章将介绍 EagleNews 系统中所使用的用户建模方式。该方法不需要强制要求用户参与交互，就可以从用户的使用记录中识别出用户的兴趣，此外，该方法使用一种折中的方法，在一个用户模型中，同时反映出用户的长期兴趣和短期兴趣。

3.1 用户建模概述

个性化推荐系统可以筛选互联网上的海量数据信息，将有可能符合用户兴趣喜好的内容，推荐给用户，过滤掉与用户兴趣毫不相干的信息，此外，系统还会将推荐给用户的内容，按照预测出来的用户对其感兴趣的程度，进行排序。

为了能够做到这一点，需要个性化推荐系统对用户有足够的了解。所以，任何个性化推荐系统都需要收集、整理、理解和利用用户信息，以实现为用户推荐的目的，这个过程统称为用户建模。

用户建模的研究与个性化推荐系统的研究密切相关，绝大多数的个性化推荐系统，都需要某种形式的用户模型（User Model）参与才能正常工作。用户模型其实就是一系列与用户有关的数据的集合，并以某种方法表示出来，如：用户的

账户名或标识符、性别、年龄、地理位置、国籍、教育程度等等信息。用户模型即是通过用户建模得到的产物，通常也称为是用户概况（User Profile），本文在后文中将不区分二者区别。

用户建模的方法多种多样，但是用户建模的过程，可以简单归纳成三个步骤，如图 3-1 所示。第一个步骤是收集用户信息，第二个步骤是建立用户模型，最后一个步骤是应用模型。

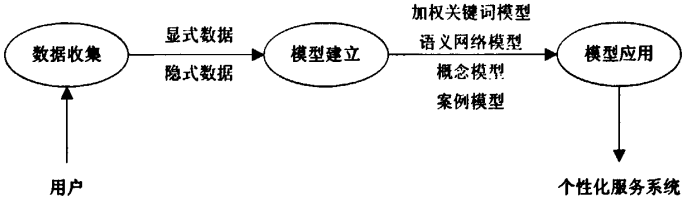

图 3-1 用户建模过程概述

收集用户信息的方法一般有两种，显式收集（Explicit）或者隐式收集（Implicit）。“显式”是指在收集用户信息的过程中，需要用户参与交互的收集方式，如让用户填写表单、回答问题或者其他的可视化交互方式；而“隐式”则是指在收集用户信息过程中，不明确要求用户参与，用户只需使用系统一段时间，其数据信息就会被系统通过某种方式得到，常见的如浏览器的缓存、Cookie 信息或者专用客户端的反馈信息。这两种方式各有利弊，采集到的信息也多有不同。现代个性化系统中，多采用隐式的信息收集方式，或二者结合的方式[35]。

如果用户模型建立后，只要不强制修改就一成不变的，称之为静态模型，如：现在国内投入商业运行的如百度新闻等，多为静态模型；而在模型建立后，会在用户使用过程中，通过自动学习，不断更新的，称之为动态模型，动态模型根据其能够反映的用户兴趣时间期限上的差别，又可分为短期兴趣模型和长期兴趣模型[36]。

事实上，用户建模并非个性化推荐领域所独有，其他领域也会涉及。而本章内容则只关心用于个性化推荐领域的用户建模。个性化推荐领域所使用的用户模型有多种表示方法，常见的有：加权关键词模型[37]、语义网络模型[38]、概念模型[39]等。

在本文设计与实现的 EagleNews 新闻系统中，用户建模使用的是隐式用户数据收集方式，用户模型表示方法采用的是较为常用的加权关键词模型，选择这种

表示方式，也是因为本系统采用的是基于内容的推荐方法，会在第4章中介绍。为了使用户模型更加准确，还引入了多模型的方式，这些内容会在3.2中详细阐述；另外，系统中使用的用户模型还兼顾了用户的短期兴趣与长期兴趣，这部分内容将在3.3用户模型生成和更新算法中介绍。

3.2 基于加权关键词的隐式多模型用户建模

为了使用户得到更好的体验，在EagleNews系统中，我们选用了隐式的用户数据收集方式，由于系统采用基于内容的推荐方法，所以，与之相对应的，用户模型也就选用了加权关键字的表示方式。为了能够更加准确地表达用户兴趣，我们还引入了多模型用户建模方法。

3.2.1 数据隐私

个性化系统，顾名思义，就是要使系统理解用户的兴趣以及偏好，这就意味着，用户必须要向系统透露一部分个人信息，这也就有了用户隐私被泄露的风险。隐私安全问题历来与个性化推荐系统形影不离。

在EagleNews新闻推荐系统中，使用的推荐方法是基于内容的推荐方法改进而来，所以在计算推荐列表时，系统更加关心的是用户兴趣的表示与新闻资讯文档的吻合程度，而对用户的本身属性信息（如年龄、性别、地理位置等隐私信息）并不关心。由此，系统在建立用户模型时就需要回避用户隐私信息。

在第5章中会介绍，EagleNews系统以网络服务的形式存在，则访问系统获取服务，需要客户端软件。因此，我们把识别用户的方法与客户端软件结合起来。由客户端软件来产生用户的标识符，一般这类标识符由运行客户端的系统所在硬件的处理器或者其它硬件的串号生成，这些硬件的串号都是绝对唯一的存在，所以可以良好的胜任识别用户的任务。

如此一来，用户的兴趣偏好便与该用户使用的客户端软件唯一挂钩了。而用户模型中其他信息都不需用户的交互即可得到。由于用户在使用系统的时候没有透露任何个人的隐私，而在系统中，保存的信息仅是用户选用的终端软件产生的用户标识符。所以，EagleNews的用户建模方式从最大程度上避免了用户隐私被泄露的问题。

3.2.2 用户信息收集

前文已经提到，收集用户信息的方法有两类，一类是显式的用户信息收集，需要用户填写表单或者与相应的用户接口进行交互，才能完成用户信息的收集工作，通过这种方式进行收集的用户信息，往往是静态信息，诸如用户的性别、年龄、学历等等与用户本身的属性、特点相关的资料。

另一类则是隐式的用户信息收集，这种信息收集方式，不需要明确提示用户，提交自己的信息，只是在用户使用系统的过程中，系统从用户与系统的交互行为产生的记录中分析得到用户的信息，这种方式又叫做隐式反馈。

一般来说，隐式信息收集有较多的优势，使用这种方式一般不会给用户带来负担，用户不需要学习系统的使用方法，或者研究表单的填写方法；另一方面，如果要创建动态的用户模型，系统需要始终保持从用户处得到反馈，对于明确要求用户操作的显式信息收集方式来说，就需要用户在使用过程中，不断地向系统提交自己的意见信息，造成用户的疲劳感和厌烦感。

在 EagleNews 系统中，由于系统采用的是基于内容的推荐方法，不需要知道与用户本身属性相关的信息，就没有必要通过显式信息收集来获得这部分信息，所以在该系统中，采用以隐式信息收集为主的信息收集方式。

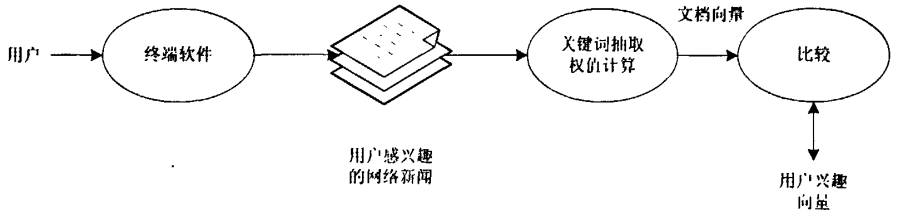

图 3-2 EagleNews 系统中使用的用户建模方式

EagleNews 系统是一个网络服务，用户需要通过终端软件来与之进行交互。用户使用的各类终端软件，就扮演了负责用户信息采集器的角色。在用户的每一次操作中，都会有信息被传送到服务器，在 3.3 节中，会详细介绍系统采集用户信息的方式以及具体采集的用户信息。

3.2.3 用户模型表示

用户模型有多种表示方法，其中最为常见的就是加权关键字表示法。这种表示法中所用到的关键字可以从Web文档的文字中提取出来，也可以由用户指定。每个关键词代表一个用户的兴趣或者一个主题。对于EagleNews系统来说，待推荐对象是网络新闻资讯信息，从文本中提取关键字信息相对来说较为容易也代价较低，所以使用这种方式来表示用户模型，无疑是比较合适的，在最后进行推荐时，系统使用的是基于内容的推荐方法，这样的表示方法也便于相似度的计算。而权重的值，比较常见的方式是采用其TF/IDF值作为权重。在EagleNews系统中，使用的权值不是单纯的TF/IDF值，这在3.3节中会详细介绍其生成方法。

在使用加权关键词方法表示用户模型时，经常将关键词按照类别进行划分，这样能够得到更精细的粒度。如图 3-3，就是一个按照类别将关键词分类后的用户模型。像上图中，如果不将用户模型进行分类，像关键词“周杰伦”和“拜仁”都有比较高的权值，则在最后进行检索时，可能只能得到“周杰伦观看拜仁慕尼黑比赛”这样两者兼顾的结果。

体育

拜仁	足球	德甲	联赛	……	……
0.75	0.35	0.55	0.42	……	……

音乐

周杰伦	R&B	Rap	排行榜	……	……
0.87	0.63	0.36	0.33	……	……

财经

版块	上市	跌停	大盘指数	……	……
0.92	0.42	0.35	0.88	……	……

图 3-3 基于加权关键字的用户模型示例

将用户模型按照分类进行划分，还可以减轻稀疏（Sparsity）问题[2]的影响。不同类型的新闻资讯信息，其产生频率和产生数量的差别很大。由于采用的是加权关键词表示法，所以，用户模型中的关键词都是从新闻文档中提取而来，如果采用单一兴趣模型，那么其中，与某兴趣相关的新闻的数量相比与其他兴趣相关的新闻数量少的话，那么代表该兴趣的关键词也会相对要少。这样一来，在进行推荐时，该兴趣主题的新闻资讯信息就无法得到有效推荐。建立多兴趣模型后，用户访问系统时，按照分类来访问新闻文档，则每个兴趣相关的内容，都会得到有效的推荐。

此外，在 EagleNews 系统中，使用按照分类分开的加权关键字方法表示用户模型，还有出于系统效率的考虑。在 4.3 节中，会详细介绍本系统中使用的资源自适应方法，使用这样的用户模型表示法，为在系统运行中进行参数的调整提供了很大的方便。

除了加权关键字模型外，由于 EagleNews 系统中的新闻按照分类进行分别推荐，所以，还会为每个用户储存一个分类列表。该分类列表会根据用户阅读习惯而进行重新排序，使经常阅读的分类被排在前面。

形式化地描述上述信息，对于一个用户 u，其用户模型分为两部分：

$$\begin{cases} \text{UserProfile}(u) = \{p_1, p_2, ..., p_n\} \\ \text{CategoryList}(u) = \{c_1, c_2, ..., c_c\} \end{cases} \quad \text{公式 (3.1)}$$

第一部分为代表用户兴趣的模型，其中 p_i 为分类 i 的子模型。其中，

$$p_i = \{t_1, t_2, ..., t_m\} \quad \text{公式 (3.2)}$$

每个子模型 p_i 都是一个元组 t_i 的集合。一个元组定义为：

$$tuple = (keyword, weight, originweight, timestamp) \quad \text{公式 (3.3)}$$

这个元组中保存的数据有关键词，此关键词的当前权值，此关键词的原始权值，以及此关键词添加入到用户模型的时间戳。下文关键词的更新方法中，会介绍其他两个值的作用。第二部分为用户感兴趣的分类列表，这个列表中是一个有序列表，基于用户对不同分类的兴趣程度来进行排序。

3.3 用户模型的生成与更新

在 EagleNews 系统中，用户信息数据由终端软件所采集。因为在整个过程中，没有用户的参与，所以采集的数据都是用户使用的记录。要创建用户模型，必须从用户的使用记录中分析得到。

用的通过终端访问系统的服务，其访问动作可以归结为以下种类：

- 请求分类列表
- 更换分类
- 阅读特定的新闻
- 更换阅读的新闻

在以上几种动作中，只有更换阅读的新闻这个动作会产生用户使用记录。因

为用户使用新闻推荐系统，其根本目的是为了要阅读新闻。其阅读的具体新闻，都带有分类信息，这样，只要记录了这个动作，那用户阅读的分类、新闻信息就都齐全了。

系统实际运行中，记录下来的信息有：用户的 ID，由终端软件生成，一般根据系统运行的硬件平台的芯片串号或者其他硬件如网卡Mac地址等信息生成，确保其唯一性；阅读的新闻 ID，该 ID 是用户请求特定分类新闻后，系统发送给用户的推荐新闻列表中所包含的信息。

阅读新闻的时间比例：在实际系统中，这个数值是一个百分比。起计算方法，按照终端的不同类型而不同。如，在我们为盲人设计的语音合成终端中，用户使用收听的方式来阅读新闻，则其阅读新闻时间比例是收听时间比上新闻总播放时间；而在为普通人设计的手机终端中，该时间比例为用户阅读的页数和新闻总长度的比例。

上述三种信息会在用户操作终端软件的时候，批量地反馈给服务器，服务器将此信息保存在数据库中。图 3-4 所示的即为用户模型的生成过程。

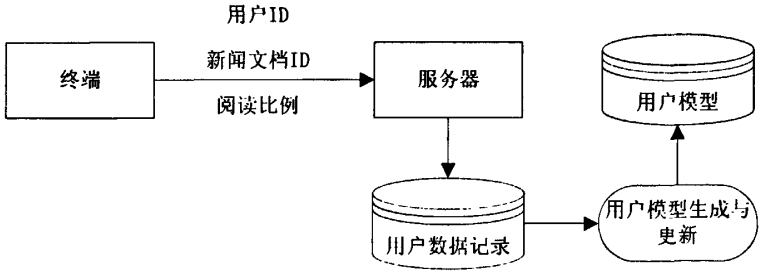

图 3-4 用户模型生成过程

EagleNews 系统中，有一个专门的进程用于维护和更新用户模型，该进程首先读取一条用户使用记录，提取该记录中包含的新闻条目，然后将其表示成文档向量，然后提取该用户对应新闻分类下的用户模型，将提取出的文档向量按照一定的算法归并到其中，该过程可以使用伪代码描述如下：

表 3-1 生成和更新用户模型的算法伪代码

```
WHILE 存在未处理的用户记录
    提取一条用户记录；
    找到该记录对应的新闻文本；
```

将该文档表示为加权关键词的向量；
将该向量归并到用户模型；
将该条用户记录标记为“已处理”；

在上述算法当中，最为关键的一个步骤，就是将表示文档的加权向量归并到表示用户模型的加权向量中的过程。

3.3.1 短期兴趣模型与长期兴趣模型

一般来说，每个用户都会对某些特定内容的新闻有着比较稳定的兴趣。比如，某用户喜欢体育新闻，兼且喜欢军事新闻，也有些用户会喜欢更多的内容类别。这种比较稳定的兴趣，我们称之为是长期兴趣。除了长期兴趣之外，用户可能会在某个特定的时期对某些特定的事件或某个特定的领域产生兴趣，但是随着时过境迁，又会渐渐淡忘。比如，某用户本身对体育新闻不感兴趣，但是08年时奥运会在北京召开，这件事情对中国意义非凡，该用户在8月份开始对体育新闻中的奥运新闻产生了浓厚的兴趣，密切关注，而随着奥运落幕，终于热情渐渐降低，不再关注体育新闻。

长期的兴趣与用户本身的特点有关，比如性格、成长经历等，一旦养成就较为稳定。这样的兴趣识别起来比较容易，一般来说，用户时常关注的内容，就是其长期兴趣所在。用户长期兴趣的获取，可以通过直接与用户交互得到，如百度新闻搜索，谷歌资讯等都是采用这样的方式；除此之外，也可通过用户的使用记录由系统自动学习出来。

短期兴趣和长期兴趣不同，一般只是用户临时关注的事件，这类事件往往是突发的，重大的或者有特殊意义的事件。这类事件往往会持续一段时间，但又不会持续很久，也有可能是周期性的，比如每年到了国庆长假，很多人会对旅游产生兴趣。识别用户的短期兴趣，由于其时效性的问题，通过与用户交互这种方式，并不理想，只能通过统计用户使用记录的手段，但是短期兴趣持续时间短，与用户偶发性的浏览不易区分。一个用户模型如果能够同时反映出用户的长期兴趣和短期兴趣，则能使用户在使用过程中获得较好的用户体验。

3.3.2 长短期兴趣复合模型

要使用户模型同时反映用户的长期兴趣和短期兴趣，有多种方法。较为简单的一种思路，就是为长期兴趣和短期兴趣分别建立模型。这种方法思路简单清晰，但是使用这样的方法，要为用户产生推荐列表，就要分别计算，而维护模型是，也需要分别去维护，在用户数量较大或者待推荐集合较大时，对系统的压力较大。

在 EagleNews 系统中，我们设计了一种方法，只使用一个模型，就同时兼顾到用户的短期兴趣和长期兴趣：

1） 如前文所述，在本系统中，用户模型使用加权关键字的形式来表示，其权值为该关键词的 TF/IDF 值。每次更新用户模型时，在将新的关键词及其权重归并到原有模型时，系统并不是直接将 TF/IDF 值累加，而是首先将原有模型中的关键词权值乘上一个衰减因子，然后再将新的关键词归并进去，归并的时候，如果关键词已经存在，则将其权值累加到衰减后的权值上，如果本不存在，则将关键词直接加入到用户模型中；

$$w_i = \tau w_i' + w_i'' \qquad \text{公式 (3.4)}$$

其中 w_i 表示关键词 k_i 的新的权值，w_i' 表示其原有权值，w_i'' 表示该关键词在日前所分析的文档中的 TF/IDF 权值，τ 为时间衰减函数。

这样设计的原因是因为用户的短期兴趣表现出随时间流逝而消失的特性，使用时间衰减因子来使关键词权值发生衰减，符合用户行为本身的规律。τ 是一个与时间相关的函数，系统中选用的是自然对数的指数函数：

$$\tau = e^{-\lambda(t-t_0)} \qquad \text{公式 (3.5)}$$

其中，λ 是一个衰减程度的因子，其值的设定，与待推荐的新闻的时间窗口的大小设定有关，如果系统推荐的新闻时间窗口根本就小于一个关键词权值衰减到 0 的时间，那么在用户模型中，将无法区分出短期兴趣和长期兴趣。根据实验分析，系统中默认采用的时间窗口大小是 14 天，而用户短期兴趣在 7 天内衰减到低于 10%，即能获得较好的区分度，所以可以求出 λ 的值为 0.3。

2） 用户模型维护的加权关键词表，分为有效词和候选词，所有的关键词按照其权值大小排序，设定一个权值阈值 w_t，如果一个关键词 k_i 的权值 $w_i > w_t$，则该关键词为有效词，否则为候选词。在最后进行推荐时，系统仅参考有效词进

行推荐。阈值w_t的值，则需要根据实验数据观察，得到一个较易区分长短期兴趣的值。

使用上述方法更新用户模型，可以识别用户的长期兴趣，由于用户经常性地浏览符合自己兴趣的内容，代表其长期兴趣的关键词的权重会不断提高，始终保持在有效词的行列中。

而对于短期兴趣来说，代表的关键词首先会停留在候选词的行列中，当短期反复浏览相关内容，其权值就会提高，进入到有效词列表中，而随着事件逐渐过去，这类词的权值也会不断衰减，最终重新成为候选词。

而对于偶发性的浏览行为，代表其的关键词进入候选词列表后，权值会快速衰减，最终被淘汰出用户模型。可以看出，代表短期兴趣的关键词可以通过不断的浏览行为转化成长期兴趣，也会因为不断衰减最终淘汰出用户模型。该方法的关键就在于w_t这个阈值和衰减因子的选定。在下一小节中，会通过实验来确定一组比较合理的取值。

3.4 实验分析

为了观察用户长短期兴趣的差异区分阈值，以及验证系统所使用的方法的有效性，需要实验数据进行观察。该实验采用从新浪、网易、腾讯、新华等主流国内新闻门户网站抓取的新闻资讯文档数据，以及若干测试用户使用数据，来进行测试。数据的时间限制是从2009年11月到12月。

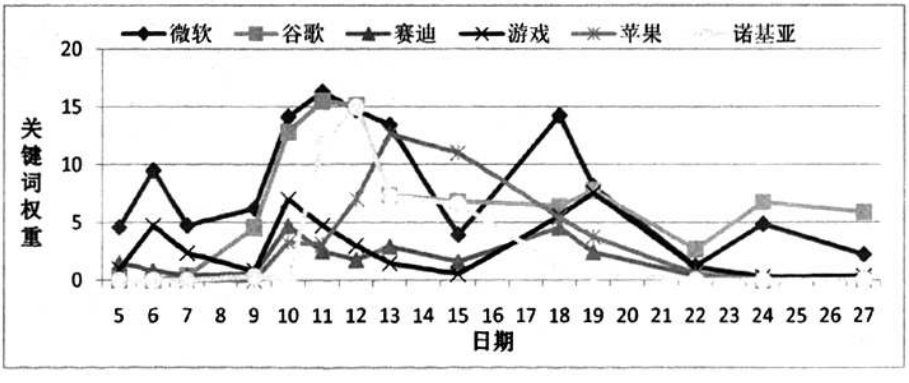

图 3-5 某用户11月浏览科技类新闻的用户模型变化图

在上图中，横轴表示的11月的日期，纵轴表示的是每天更新过用户模型后，各关键词的权值，如果一个关键词的权值衰减后，被淘汰出用户模型，则其权值

在图中用0表示。本图只绘出了用户模型中部分关键词的权值在一个月中的变化趋势，从中，我们可以看到，该用户对微软公司，谷歌公司等有关的新闻较为感兴趣，始终都保持着关注；而对诺基亚和苹果相关的信息，则表现出偶尔感兴趣的现象；从图中，我们可以看到将w_t设定在 4~6 之间，可以较好地区分出用户的兴趣差异来。

表 3-2 某用户科技类用户兴趣模型关键词权值变化表（部分）

日期	5	6	7	9	10	11
微软	4.555134	9.464528	4.716067	6.18424	14.11934	16.2937
用户	3.956537	6.077238	3.028219	2.876638	6.615592	12.62962
市场	1.613386	2.264624	1.128437	2.262942	7.551063	8.861453
移动	1.856283	0.985737	0.491181	0.455553	6.925885	5.397053
中国	3.094583	2.842066	1.416169	0.969393	1.960683	2.870906
科技	1.526944	1.278237	0.636931	1.993689	2.503396	2.672007
游戏	0.916266	4.694324	2.339129	0.789083	7.005834	4.734194
表示	0.604481	1.464375	0.955435	1.204668	3.383263	4.000017
技术	1.244284	1.046976	0.521696	0.840755	2.230879	5.839348
竞争	0.40434	2.07169	1.0323	1.733634	3.075968	3.711496
软件	1.781571	1.806848	1.134776	3.9074	8.496882	13.18072
美元	2.593087	1.377	0.686143	0.387585	13.16316	8.472416
网络	0	2.27388	1.133049	4.053889	4.430231	6.345918
赛迪	1.5137	0.803816	0.400532	0.597541	4.72496	2.521122
份额	0	0.529571	0.263879	0.318767	0.793533	2.44389
谷歌	0	0	0.377805	4.528632	12.79682	15.47939
广告	1.19194	0.632952	1.407217	0	4.364785	8.173197
window	2.381021	13.47601	6.714942	1.389092	0.871594	2.073297
网站	1.5193	1.932816	1.195427	0	1.990871	6.05449
产品	1.933725	2.424326	1.208014	0.89126	2.542605	6.809945
互联网	0.962712	0.813473	0.405344	0.949241	1.736659	3.48535
企业	0.37829	1.098053	0.547147	0	0	3.127258
厂商	0	1.033633	0.515047	0.379094	2.57228	5.798586
ceo	1.01877	1.106951	0.551581	2.808477	2.328128	1.921058
系统	1.855164	3.772629	2.518758	1.989821	2.022804	1.964131

3.5 本章小结

个性化推荐系统要实现个性化，必须要了解用户的兴趣和偏好。一般个性化系统都会为每个用户建立各自的用户模型，用于学习和记录用户的兴趣和偏好，以便在推荐时使用。建立用户模型有隐式和显示两种方法，所建立出来的模型要能反映用户的长期兴趣和短期兴趣。

本章中重点介绍了 EagleNews 个性化新闻推荐系统中，所采用的隐式用户建模方法，还介绍了系统为每个用户建立的多模型用户模型，此外，在该系统中，采用了一种将用户长期兴趣和短期兴趣融合在一起表示的模型表示方法，使得以此方法建立的用户模型能够兼顾到用户的长期兴趣和短期兴趣。

最后，使用实际系统运行数据，分析了该用户建模方法建立的模型案例，以及测定了用户模型建立方法所需采用的最佳参数。

第4章 资源自适应的新闻推荐算法

在个性化推荐系统中，推荐算法的好坏决定了系统的推荐效果。正如第 2 章所述，推荐算法可以分为三大类，每种算法并不完全通用，一般来说，个性化推荐系统要根据其目标用户的需求和被推荐对象的特性来选择合适的推荐算法，以达到最优的推荐效果。EagleNews 系统选用的是基于内容的推荐算法。选择这一算法，主要是考虑到网络新闻的特性以及系统性能的要求，在本章后续小节中，会介绍选择这一推荐算法的根据。

学术界对于推荐算法的研究，普遍关注的问题是推荐的效果，如推荐的准确性、全面性，较少关注推荐系统的性能问题，而性能对于一个应用系统来说，同样重要。但是推荐系统的性能和推荐效果，恰如跷跷板的两端，不可兼得。因为更精确的推荐势必需要更大的计算量，而计算资源本身有限，就只能以延长计算时间来弥补，如此一来，就降低了系统的性能。

在本章中，将要介绍在 EagleNews 系统中，使用的推荐算法，以及为了在推荐效果和系统性能之间取得平衡，而做出的改进。通过这种改进，系统在推荐效果和性能之间，取得了一种平衡，并且使这种平衡达到一种动态稳定，兼顾到了系统的性能和推荐效果。

4.1 个性化新闻推荐算法

推荐算法是一个个性化推荐系统的核心，决定了这个系统的推荐效果的好坏。由于同一种推荐算法，应用于不同的场景，其推荐效果是有差别的。所以，一个推荐系统为了取得较好的推荐效果，应该根据应用本身的特性来选择合适的推荐算法。

协同过滤的推荐算法是目前使用最为广泛的推荐算法，同时也是最为成功的推荐算法，如国内比较著名的豆瓣网，MTimes 网等，都采用了协同过滤的方法。但是协同过滤算法却并不适合在 EagleNews 系统中使用，其原因有以下两点：

首先，在基于协同过滤算法的推荐系统中，要求用户对查看过的项目进行评价，这也是系统判断用户之间相似度的依据，而 EagleNews 系统，为了提高用户

体验和减轻用户压力，采用的是隐式用户建模的方式，见3.2，通过这种方式取得的数据，仅能知道用户对某篇新闻的关注度，而这种关注度与偏好还存在一定的区别，虽然可以通过连续统计关注度来分析得到用户的兴趣偏好，但是单一的对某篇新闻的关注度，却不足以认定为用户对其感兴趣，这样一来，以此为依据的推荐算法效果就要打折。

通过公式来说明问题：协同过滤算法将与某用户相似的用户感兴趣的内容推荐给该用户，但是在EagleNews系统中，并没有办法得知，用户对某篇新闻的兴趣如何；关注某个内容，可能与用户本身特征有关，比如年龄，行业，地区等；但是关注了，并不等于就有兴趣。在用户明确打分的系统中，可以区分这两者，但是在EagleNews系统中，却没法区分。

其次，EagleNews系统，是一个个性化新闻推荐系统，其推荐对象为新闻报道。新闻的特性是失效性强，更新迅速，数量大。数量庞大，就使得两个用户看到同一篇新闻的概率降低，另一方面，协同过滤的推荐算法奏效有一定的延迟时间，而新闻则时效性很强，一旦时间过了，那么新闻的价值就会降低，把过期的新闻推荐给用户，并不能增加用户体验。

基于内容的推荐方法，起源于信息检索和信息过滤的研究。该方法比较适合EagleNews推荐系统。系统的推荐对象为新闻，新闻的内容为文本，而现今文本处理技术已经相当成熟，从文本中提取特征较为容易。其次，在基于内容的推荐系统中，不需要判断用户之间的相似度，仅依赖用户个体的历史行为来分析用户的偏好，也不存在上述协同过滤算法的问题。

4.2 推荐系统性能优化策略

提升个性化推荐系统性能的方法有很多种，对于EagleNews系统来说，由于其推荐的对象是新闻文档，有着一定的特殊性，可以根据新闻的特性来采取一系列的措施，降低系统在推荐时的运算量，以此达到提高推荐算法运行速度的目的。

4.2.1 基于时间窗口的新闻筛选

一般来说，新闻的价值与其时效性是密切相关的，滞后时间过长的“新”闻，其价值就大打折扣了。在个性化推荐系统中，推荐给用户的新闻，也要充分考虑到时效性的问题，尽可能将最新的新闻推荐给用户。

EagleNews 系统使用时间滑动窗口来控制被推荐新闻的集合。在使用推荐算法来挑选推荐给用户的新闻时，仅从发表时间处于时间滑动窗口中的新闻中挑选。时间滑动窗口，是指从现在向过去延伸的一个时间段，其长度是 W 天。在这里，W 是一个可以被系统资源调度调整的一个参数。如果一篇新闻其发表时间距离现在已经过了 W 天，则该新闻会被移出被推荐新闻的集合。从实验数据中，可以看出，W 的最佳取值在 10 天到 17 天之间。默认取值为 14 天。

4.2.2 TOP-K 文档关键词选取

在 EagleNews 系统中，判断一个文档符合用户的兴趣偏好，使用的方法是计算该新闻文档和用户模型的相似度。系统使用向量法来表示文档，一篇网络新闻文档 D 可以用关键词向量来表示 $d_i(k_1,k_2,...,k_n)$ 在文档向量中，关键词按照其 TF/IDF 权值来降序排列。因为高权值的关键词，能够更多的反映出该新闻的内容，而权值较低的词，其内容与文档本身关系较弱，所以，在计算用户模型和文档的相似度时，只需要使用最能够代表新闻特征的 K 个关键词的权值去计算，因为这些词对最后计算出的相似度数值贡献最大。

在计算相似度时，选用的关键词数量的多少，决定了计算速度的快慢。选用的关键词越多，计算的精确性就越高，但是计算速度就越慢。如果从所有关键词中，只挑选那些对反应文档内容贡献最大的 K 个关键词，则对于最终计算结果的影响，相对较小，而减少了那些大量的与文档主题关系不大的关键词，则可以大幅度提升系统的计算速度。通过实验分析，在 EagleNews 系统中 K 的取值在 150 到 250 之间较为合适，默认设定的值为 200。

4.2.3 TOP-K 用户关键词选取

在 EagleNews 系统中，推荐用户感兴趣的新闻，其本质就是找到与用户模型相似的新闻文档。用户模型的表示与新闻文档是一致的也是用关键词向量来表示，每个关键词的权重，也是该词的 TF/IDF 值。与上一节中提到的原理相同，在计算相似度时，权值最高的 T 个关键词，对最后相似度数值的共享最为巨大，所以，系统在计算时，仅采用最前面 T 个关键词进行计算。根据实验数据的观察，T 的取值被设定在 25 到 75 之间较为合适，实际系统默认选用的值是 50。

4.3 资源自适应的推荐系统模型

4.3.1 新闻推荐系统的突发特性

前面几节中，介绍了一系列提高系统性能的方法。这些方法采用的原理相同，主要是通过降低计算规模来实现提高计算速度的。但是，这样做，计算的速度固然提升了，准确度却被一定程度的牺牲。

互联网上的新闻事件发生具有一定的随机性，并不均匀。在发生社会普遍关注的突发性大事件发生的时候，如08年5月的汶川大地震时期和08年8月北京奥运会时期，由于用户普遍比较关心，各大网络门户媒体都加大了报道的频率和密度。那一时期，新闻量大增，而与此相对应的，用户对新闻提供媒体的访问频率也大大增加了。而随着时间推移，事件被淡忘，新闻数量和用户对系统的访问又会恢复正常。

4.3.2 资源自适应推荐模型

鉴于此，在EagleNews中，我们将采用一种资源自适应的算法，来平衡系统的性能的准确度。

当服务器端的CPU资源和内存空余容量不能适应用户访问的需要，造成系统响应速度降低时，我们设法降低系统的计算量。可以采用这么几种策略：

1）减少系统中新闻的数量。假设在平均状态下，系统中的新闻容量为N_{normal}，当系统中新闻量明显超过N_{normal}时，每隔时间t，就将新闻的时间窗口减小α，直到系统响应速度恢复平均水平，或系统中待推荐新闻数量降低到N_{normal}的一半停止，当系统资源空闲时，则每隔时间t，就将新闻的时间窗口增加α，直到待推荐新闻数量恢复N_{normal}为止（根据实验中测定，在我们的系统中N_{normal}取值为14天产生的新闻数量，α取值为5%，t取值为1小时）。

2）调整K的值，在平均状态下，进行相似度计算时，使用TF/IDF值最高K_{normal}个进行相似度计算，当系统资源不足时，每个时间t，将K的值减少β，降低到$K_{normal}/2$为止，当系统资源空闲时，则每隔时间t，将K的值增加β，直

到恢复K_{normal}为止（根据实验数据测定，K_{normal}取 200，t为 1 小时，β取值为 50）。

3）调整M的值，同样使用上述 2）中的方法，调整用于代表用户兴趣的关键词的数量。

4）对上述三个参数的联合调整。当新闻量过大，系统响应降低时，首先调整新闻的时间窗口，然后是K，然后是M，三个参数轮番调整，直到系统响应能够达到一般条件下的系统响应。而当情况逐渐好转时，则做相反的操作。

4.4 实验及数据分析

为了确立 4.3 节举措中提到的系统参数，在 EagleNews 系统之上，进行了实验。实验用的数据，为系统 2009 年 7 月 1 日到 7 月 30 日从互联网采集的新闻资讯文档（目前来源有：搜狐、新浪、新华网、网易等）32,532 篇，包含 24 个分类的新闻。用于测试的服务器为低端桌面服务器，使用 Windows XP 操作系统，CPU 为 1.86GHz 双核，内存 3G，硬盘 160G。

在所有新闻中，选用数量最多的体育类新闻，作为测试推荐算法准确度的分类，用娱乐新闻，财经新闻等测试系统在并发访问情况下的响应速度。

对于 EagleNews 新闻推荐系统，推荐准确度$R_{accuracy}$被定义为符合用户兴趣的新闻和推荐出来总新闻数量的比，系统响应速度$R_{response}$，则用于评价系统优劣的指标，被定义为

$$score = p \times R_{accuracy} + (1-p) \times R_{response} \qquad \text{公式 (4.1)}$$

其中$R_{accuracy}$为推荐准确度的归一化评价分值，其计算方法为推荐符合用户兴趣的新闻与总推荐新闻数量的比值。$R_{response}$为系统响应速度的归一化评价分值：

$$R_{response} = r_{\max} - \frac{r_{\max} - r_{\min}}{r_{\max}} \cdot t \qquad \text{公式 (4.2)}$$

其中，$r_{\max}$为最高评分，这里设为 100，$r_{\min}$为最低评分这里设为 60，$t_{\max}$为最低评分时，可忍受的系统响应时间，这里$t_{\max} = 8s$，s为系统实际响应时间，p为加权参数，取值为 0.8。

首先评估的是时间窗口参数设定对系统整体性能的影响。在这里我们将其余参数设定为计算相似度使用关键词数量为 $K=200$，标识用户兴趣的关键词数量为 $K_{profile}=50$，在限定了这两个参数后，从下图中，可以看到时间窗口从 7 天提升到 14 天，可以使得系统综合评分有较大幅度提升，而从 14 天到 21 天，则变化不大。

接下来是对相似度计算使用关键词数量和用户档案文件中选用关键词数量相互作用下，对系统整体性能的影响的评价。将其余参数设定为时间窗口 $W=14$ 天。

图 4-1 和图 4-2 中的曲线，反应了参数的不同取值对系统综合评分的影响。时间窗口 W 的取值在 14 天时，系统整体表现和在 21 天时，表现相差无几，却远远好于 7 天时的表现，由于在 7 天时，虽然由于整体新闻数量较少，提高了响应速度，但是较少的新闻数量也明显降低了推荐的效果，导致整体表现下降。在这里可以将 $W=14$ 作为系统时间窗口参数的一个初始值，以获得一个系统综合最优的初始状态。

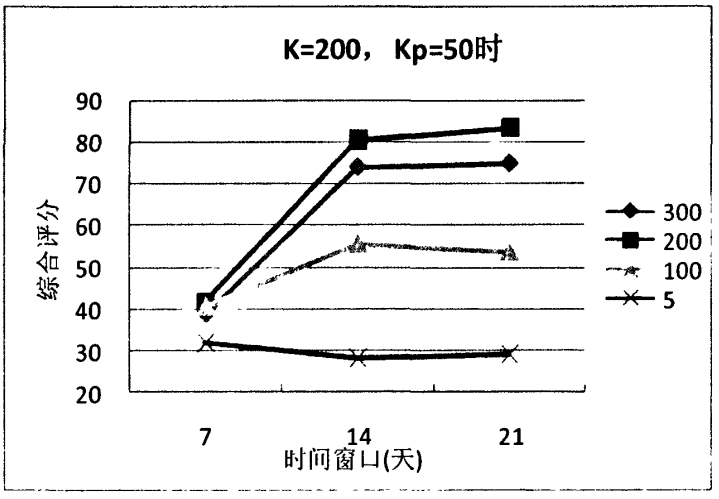

图 4-1 K=200，$K_{profile}$=50 时间窗口对系统综合评分的影响

在图 4-2 中，可以看到在 $K=200$ 时，无论 $K_{profile}$ 取值多少，系统综合表现都优于 K 的其他取值时的情况。K 取到 300 时，可以看到系统综合评分相差很少，可见虽然在计算相似度时，取了更多的关键词进行计算，但是对推荐准确度的影响基本可以忽略，而带来的性能损失却使得系统整体表现变差。

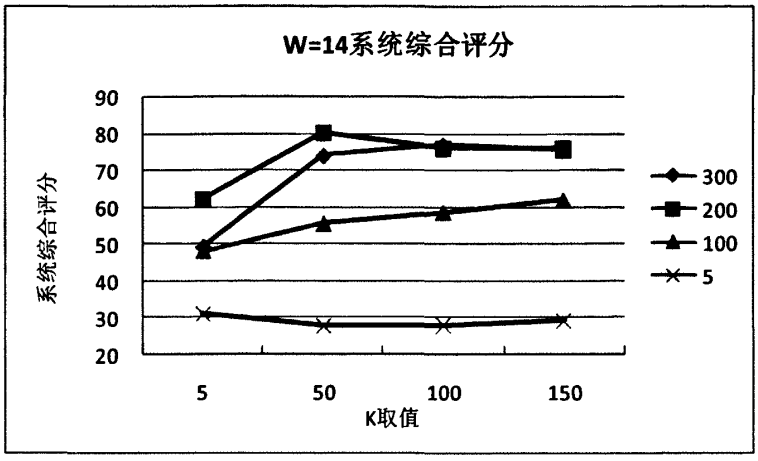

图 4-2 W=14 时，K 和 $K_{profile}$ 取值相互作用对系统综合评分的影响

4.5 本章小结

本章介绍了 EagleNews 系统中所采用的推荐推荐算法为基于内容的推荐算法，以及系统采用此算法的原因。为了提高本系统的推荐性能，本章提出了控制新闻的时间窗口、控制文档向量维数参数和控制用户模型向量维数参数的方法，来实现在推荐是对计算量的控制，从而实现提高推荐系统性能的目的。为了能够使系统在推荐效果和推荐性能上取得平衡，本章提出了资源自适应的推荐算法。最后，本章通过实验数据的分析，给出了系统运行的最佳参数设定。

第5章 EagleNews系统的设计与实现

个性化新闻推荐系统能给用户在网上冲浪带来极大的便利，节省了用户的时间，提高了用户获取新闻资讯数据的效率，有着广泛的应用前景。然而，目前全世界范围内成熟的新闻推荐系统并不多见，且这些系统皆是为西方语言所设计，又多为商业系统，不能直接为我国所用。但是，作为残疾人信息无障碍课题的一个分支，研发独立自主知识产权的新闻推荐系统又势在必行。

EagleNews个性化新闻推荐系统，是一套通用的个性化新闻推荐系统，能够提供高效、稳定的新闻推荐服务，并且有着良好的数据接口定义，可以满足不同的应用需要。

5.1 EagleNews推荐系统简介

EagleNews个性化新闻推荐系统，是完全自主知识产权的通用新闻资讯推荐系统平台，是完全由EagleLab实验室项目组成员研发系统。该系统目前已经部署到万维网之上，稳定运行。为EagleNews网络新闻终端，手机终端软件以及Web页面等多种应用提供稳定的服务。

5.1.1 特点分析

EagleNews个性化新闻推荐系统的设计有诸多考虑，具备了很多优异的特性：

通用的新闻资讯推荐系统

EagleNews个性化新闻推荐系统是一个网络服务系统，运行于互联网之上，通过网络为用户提供服务，所以可以兼容各种各样能够接入到网络中的设备，包括个人电脑、专用终端、个人数字助理（PDA）以及手机等多种多样的设备。

该系统对新闻资讯进行了抽象，新闻资讯都以文档的形态在系统中出现，所以，任何可以抽象成文档的数据，都可以通过该系统进行推荐，如网络新闻，博客，电子书，通知等等，几乎囊括了所有文字类的信息种类。

高性能、资源自适应

作为一款准商业级推荐系统，该系统有着较高的性能，能够满足大量用户同

时访问使用。并且，系统可以自动监测负载情况，可以根据系统负载情况来调配计算资源，是系统在推荐精度和推荐速度之间取得平衡。

非交互式兴趣发掘

该系统是在用户使用过程中，根据用户的使用习惯来“默默”地学习用户习惯的，并不像一般的商业系统一样，需要通过用户交互来完成用户模型的建立。

用户兴趣学习和更新

该系统建立好用户兴趣模型后，并非一成不变，而是随着用户的使用不断更新，能够敏感感受到用户兴趣的变更，并将这种变化反映到推荐结果当中。

简单的通讯协议

该系统采用HTTP协议作为系统与用户通信的协议，该协议为目前互联网上最普遍采用的协议。这就使得系统具备着非常良好的扩充能力，可以根据应用需要来扩展成各类应用。

通用数据交换格式

该系统采用XML文件格式作为数据传输格式，并且XML符合Atom标准，Atom标准是一种在互联上非常流行的用于数据传输和交换的格式。使得系统可以与各种应用实现对接。

5.1.2 系统体系结构

EagleNews采用分层体系架构设计，是为了保持最大的兼容性和可扩展性。图 5-1所示的是系统层次结构图。

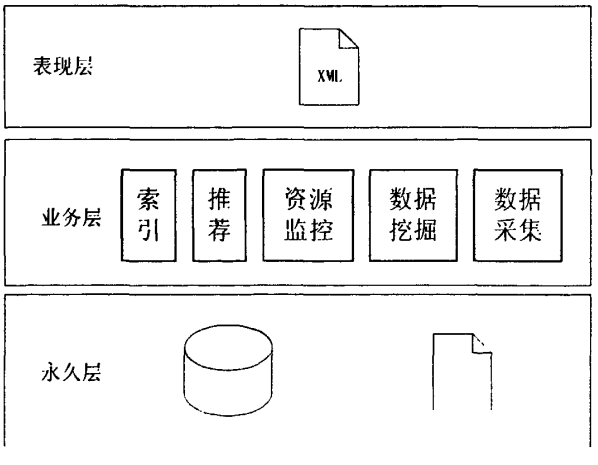

图 5-1 EagleNews系统层次结构图

整个系统总共分为三层，归纳起来可以分为表现层，业务层和永久层。表现层直接为系统的用户提供服务，这里的用户是个广义的概念，代表接入到系统中的个体，可能是专用终端，也可能是浏览器等，无论是什么，系统对外输出的数据采用标准的 XML 数据交换格式。这种格式有利于兼容和扩展系统。

业务层是模块化设计的，由系统中的几个主要重要功能分割成的模块组成。其中最重要的模块就是数据挖掘模块和资源监控模块，前者决定了系统的推荐效果，而后者则决定了系统的效率。

永久层主要用来存储数据。整个系统的数据通过该层来对业务层抽象。像新闻资讯数据和其他相关文件数据都是对业务层透明的，这也是为了日后系统容量扩大后，便于更换永久层，以便扩展系统而考虑的。

5.2 系统整体设计

如图 5-2 所示为 EagleNews 新闻推荐系统的结构化分析图，该图显示了系统各模块之间的协作关系以及系统的运行原理。

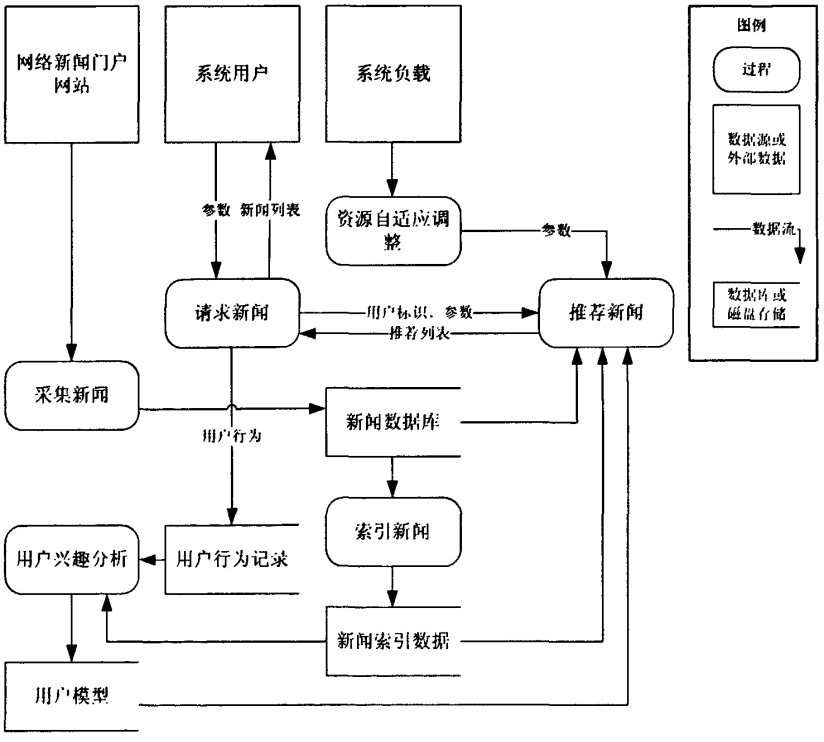

图 5-2 整体系统结构化分析图

EagleNews 系统需要完成的任务主要有两种类型：事件响应型和定时运行型。

事件响应型任务平时并不运行，只在特定事件发生时才运行，如用户请求的响应；定时运行的任务，在系统后台周期性地运行，比如新闻采集、索引生成和用户数据分析等任务。

系统需要使用到的外部数据有三个来源，第一个也是最重要的是各类网络新闻资讯网站，这类网站提供了系统的数据源；第二个是用户，用户的使用习惯是通过分析用户操作记录而来，所以用户也是系统运转的重要数据来源；第三个是运转状态，由于要使系统能够高效运行，必须时刻关注系统运行的物理环境（某台具体的服务器）状态，包括CPU负载、内存占用等等。

5.3 通信设计

EagleNews是一个运行于互联网上的推荐服务，为了使该系统尽可能紧凑、内聚，以得到最大的可扩展性，可重用性以及健壮性，在设计时，对系统提供的服务进行了精炼。该系统所需要完成的任务归纳为如下几个：

- 记录用户数据
- 建立用户模型
- 响应用户请求

除此之外，其他的诸如结果表现形式、界面、验证登陆等等问题，都不应放到系统内部去考虑，而是应该以独立子系统的形式与EagleNews联结。如此一来，EagleNews的功能内聚性就更高，设计上也被大大简化了。

根据以上原则所确立的系统边界，EagleNews系统被设计为通过通信协议与外部系统交互。

由于推荐系统提供的服务是新闻资讯信息，这些信息的载体是文本，又是结构化程度很高的文本，所以系统选用XML格式作为数据传输格式。

图5-3所示的为系统返回的推荐新闻列表，该列表使用XML格式，符合Atom 1.0标准中的Atom供稿格式（Atom Syndication Format），该标准吸取了各种RSS版本的使用经验，解决了RSS 2.0遇到的问题，是目前互联网通行的供稿交换格式。使用该格式作为EagleNews系统的数据传输格式，为今后使用该系统构建其他应用打好了基础。

```
<?xml version="1.0" encoding="gb2312"?>
<document>
  <item>
    <docID>0001</docID>
    <title><![CDATA[奥巴马当选美国首位黑人总统]]></title>
    <link>http://news.eastday.com/w/20081105/u1a3963743.html</link>
    <pubDate>2008-11-05 12:35:52</pubDate>
    <author><![CDATA[袁野]]></author>
    <origin>www.163.com</origin>
    <category><![CDATA[政治]]></category>
    <content>
      <![CDATA[<P>东方网月日消息：美国总统大选终于在北京时间点出现重要分水岭，
    据美国有线电视新闻网报道，民调一路领先的民主党总统候选人贝拉克-奥巴马不负众望，以297：
    选举人票数大幅领先对手共和党总统候选人麦凯恩，同时获得超过总统大选所需的最低票数票，
    如果<P>　　他在竞选中以"变革"为主题，强调结束伊拉克战争、实现能源自给、停止减税政策
    和普及医疗保险等，并承诺实现党派团结、在国际上重建同盟关系、恢复美国领导地位。
    <BR></P></p> ]]>
    </content>
    <keywords><![CDATA[美国 大选]]></keywords>
    <score><![CDATA[0.87939563478]]></score>
  </item>
</document>
```

图 5-3 EagleNews系统返回给客户端的推荐新闻列表

由于获取服务不需要复杂的交互，传输的内容又是纯文本，所以系统采用HTTP协议作为通信协议。HTTP有许多优点：成熟、简单、通用，再各个变成语言平台中都有对应的实现，可以很大的降低开发成本。

表 5-1 查询字符串及其含义

查询串	样例值	描述
Userid	6428523	用户 ID，该数值由用户使用的客户端来生成
Catid	1	用户阅读的分类ID
Num	10	请求的文章数量
Catlist	/	请求分类列表
Docid	53ae8973f2394	请求或者发送的文章ID
Total	100	发送的文章的总长度
Ratio	30	用户阅读的比列

系统外部与系统进行交互需要传输的数据主要有这么几类：

- 用户信息——用户ID（由终端软件生成）、终端类型
- 请求信息——请求分类ID、文章数量、文档ID

- 用户使用记录——访问时间、访问的文档ID、文档长度、阅读比例

以上数据比较短小，而且结构也相对简单，所以系统中采用query string的形式进行传送，具体的传输协议设计如表 5-1。

5.4 各模块设计

EagleNews系统基本上按照其需要完成的任务分割系统模块，各个模块之间通过接口调用来进行协作。系统主要有如下几个模块：新闻采集、新闻索引、Web响应、新闻推荐、用户模型生成。

5.4.1 新闻采集模块

新闻采集模块在系统中的名字是Crawler，其主要职责是从新闻资讯网站采集新闻内容，并将它们都导入到数据库中。该模块完成的任务，属于定时运行型任务。在系统中，其体现形式为，一个单独的，定时执行抓取任务的系统进程。

- 新闻采集模块的工作流程

新闻采集模块运作时，首先读取配置文件，从中取得需要抓取的目标新闻资讯网站。然后为每个目标新闻资讯网站创建一个独立的线程，每个线程首先下载目标网站的新闻频道主页，从中提取每个分类最新新闻的URL地址。接着检索系统中的数据库，如果URL地址在系统数据库中不存在，说明是新的新闻，以前并没有下载过，将其下载下来，保存在配置文件指定的临时目录中。之后，依次打开临时目录中的文件，保存在目录中的文件，是包含一篇新闻的HTML纯文本文件，从中提取出该篇新闻的标题、发表日期、原始URL、分类信息等等元数据，将这些数据一并存储到数据库中。Crawler的工作流程如图 5-4所示，右侧的框中，显示的是每个线程的工作流程图。

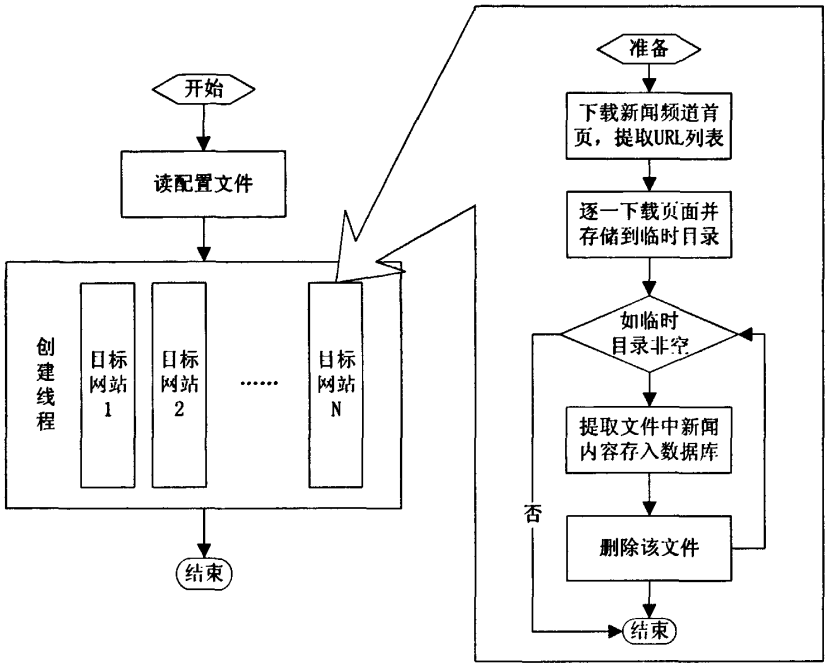

图 5-4 Crawler 工作原理流程图

- 模块设计中的关键问题

1. 内容提取

互联网上的新闻资讯类网站，普遍采用内容管理系统（CMS）来搭建，内容管理系统的特点是，将其管理的内容都以结构化的形式存储在数据库或其他存储中，然后根据页面模板生成页面供用户访问，其页面多是动态构建。鉴于此特点，最简单，最快速的内容抽取方法，就是分析网站的页面模板，建立针对模板的内容抽取方法。

EagleNews 系统中，正是使用这种方法来进行新闻抽取的。

2. 抓取频率

新闻资讯类网站，数量繁多，在采集数据的过程中，如何设定采集频率，是一个问题。如果采集频率过高，会给系统带来更大的负载，此外，这种做法对与提供内容的网站也并不礼貌，可能会使采集程序遭到拒绝访问。而采集频率过低，则会错过一些新闻。

在 EagleNews 系统中，每个新闻资讯网站的采集频率被写入到了配置文件中，每个网站区别对待。采集频率被分成了几个等级，对于更新频繁的网站，设为最高等级，每个小时采集一次。而对于更新并不频繁，如政府、机构类的网站，则

设为最低级，每天甚至更久采集一次。等级一共有四种，分别应对不同类型的网站，而目前系统中采用的只有最高和最低两种。

5.4.2 新闻索引模块

新闻索引模块在系统中的名字叫做 Indexer，其主要职责是为新闻采集模块插入到数据库中的新闻建立索引，以便个性化推荐模块和用户模型建立模块调用。该模块负责的任务也属于定时运行型。在系统中，是一个单独运行的进程。

由于使用基于内容的推荐算法，所以，事先对新闻文档建立索引，可以大幅提高推荐时的运算效率。索引采用倒排索引的形式。

5.4.3 Web 响应模块

Web 响应模块在系统中的名字叫做 WebServer，其主要职责是响应用户请求，并处理与用户之间的交互数据。该模块负责的任务属于事件响应型，仅在用户请求到来时才运行，在系统中，以守护进程的形式运行。

EagleNews 系统的 Web 响应模块采用 JAVA 2 EE 的 Servlet 技术开发。Servlet 是一种在启用 JAVA 技术的 Web 服务器上用以扩展服务器能力的编程接口技术。Servlet 通过创建一个框架来扩展服务器的能力，以提供 Web 上进行请求和响应服务，当客户端发送请求到服务器时，服务器将信息发送给 Servlet，并将其返回内容传回给客户端。

系统要响应的用户请求有很多种，为了使 Servlet 能够更健壮地运行，在系统中采用了 Action 框架，通过这套方法，使得 Servlet 仅作为系统与 Web 服务器的接口，只承担最简单的工作，这样的设计使得 Servlet 实现了最小化，而且，此框架采用工厂模式，做到了最大程度兼容软件的变化。见图 5-5 所示，HTTPServlet 通过调用 ActionFactory 来得到实际接受用户响应的 Action（查询新闻使用 QueryNews，请求分类列表使用 CategoryList，发送反馈使用 Feedback），然后通过 Action 接口进行调用，这样的结构使得实际运行的程序对 Servlet 彻底透明。

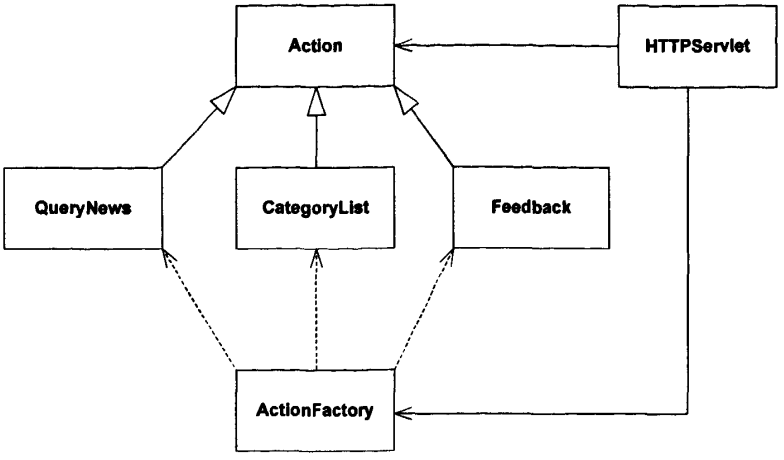

图 5-5 Web响应模块采用的Action框架的UML类图

5.4.4 新闻推荐模块

新闻推荐模块在系统中的名字叫做Recommender，其主要职责是根据输入的参数给出一个新闻推荐列表。该模块负责的任务属于事件响应型。在系统中并不独自运行，而是由Web响应模块调用。

在第4章中已经介绍过本系统所采用的推荐算法，为了系统的冷启动问题，在新闻推荐模块中还需要提供一个按照时间检索新闻的简单推荐算法。于是，在实现新闻推荐模块的时候，也采用了基于接口的设计。

图 5-6 显示的是新闻推荐模块的类图，在这个设计中，Recommender 是一个抽象类，该类有一个静态方法 getRecommender()用于创建实际包含推荐算法的推荐器对象，该方法接受的参数为用户的请求数据。使用这样的设计，系统具体调用的推荐算法是哪个，就对 QueryNews 这个 Action 来说透明了。

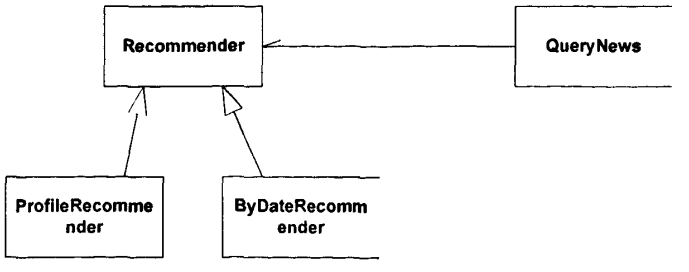

图 5-6 新闻推荐模块的UML类图

这样的设计为日后改进推荐算法提供了便利，对于一个推荐系统来说，推荐

算法是要不断改进的，所以在一开始设计的时候，就要将今后产生的变化考虑在内，并对其进行封装。

在getRecommender()方法被调用时，首先判断当前发出请求的客户端用户是否在系统中包含用户模型，如果没有用户模型，则自动创建ByDateRecommender对象，将新闻按照时间从新到旧顺序推荐给用户。而如果系统已经存在了该用户的用户模型，则创建ProfileRecommender对象，使用第4章中介绍的算法进行推荐。

由于使用的是基于内容的推荐算法，一般基于内容推荐算法存在的问题，本系统也会遇到，在本系统中，我们采用了两个比较简单的方法，来减轻了推荐算法本身缺陷带来的影响，取得了较好的效果。

5.4.4.1 冷启动问题

在一个新用户刚开始使用系统时，系统中并不存在用户的用户模型，也就没有向用户进行推荐的依据。这个时候对于用户来说，推荐系统并不起作用。这个问题被称之为是冷启动问题。

在EagleNews系统中，也存在冷启动问题。为了避免这个问题，推荐算法一般有两种处理方法，对于已经有用户模型的用户，就按照用户模型的内容，给用户推荐新闻，而对于没有用户模型的用户，则将新闻按照时效性排序，推荐给用户。这样的设计，也是基于一个平凡的假设，即更新的新闻总是要优于教旧的新闻。

在使用过程中，用户会逐渐表现出对一些新闻的关注，以及对另一些新闻的不关注，经过t天后，就会形成较能反应用户兴趣的用户模型。经过对实际系统的观察，t一般设定为2天较为适宜。

5.4.4.2 过度特化问题

在使用基于内容推荐算法的系统中，推荐时，仅依据用户模型进行推荐，而用户模型仅依据用户的历史使用数据来生成，因此，容易出现推荐内容仅与用户兴趣相似，而与用户不符的突发事件，可能也是用户可能关注或感兴趣的内容，

则不会得到推荐。这个问题称之为是过度特化问题。

在EagleNews系统中，在对推荐列表进行排序时，不仅依据其与用户兴趣吻合度，我们还添加了时效性加权。用户兴趣吻合度和时效性的权重比例为 p，这样对于最新的新闻，仍有机会进入到推荐列表中，由于最新的新闻仅与时间有关，其内容不受限制，故这样的方法，可以避免过度特化的问题。

5.4.5 用户模型更新模块

用户模型更新模块在系统中的名字为 ProfileBuilder，其主要职责是对 Web 响应模块记录的用户使用记录进行分析，为每个用户建立一个用户模型，或者更新已经存在的用户模型的数据，使得用户模型能够及时反映用户兴趣的变化。该模块负责的任务属于定制运行型，在系统运行时，也以独立进程的形式定时运行。

该模块的结构相比其他模块，较为简单，其主要部分是两个静态函数，一个负责更新用户新闻分类列表，另一个用户更新用户模型。

5.4.6 资源自适应模块

该模块在系统中的名称为 SystemInspector，其主要职责是定时检查系统的运行状态，随时根据系统 CPU 占用内存使用情况来调整系统的运行参数。该模块负责的任务属于定时运行型，在系统运行时，由一个单独的进程来运行。

该模块通过修改一个系统全局对象 Config 的属性，来实现与其他系统进程的通信。系统当前状态会实时地记录到 Config 对象的属性之中。

5.5 系统展示

EagleNews 系统是一个系统服务，所以访问该系统服务必须使用客户端系统，由于该系统与客户端的通信基于 HTTP 协议，而数据交换，采用的是 XML 文件，所以，系统的客户端构造也就相对容易，目前有多种可供选用的客户端。

图 5-7 展示的是基于该系统构建的 Web 个性化新闻推荐系统，用户可以在这里阅读 6 个来源，超过 20 个频道的网络新闻，并且网络新闻实时更新。

图 5-7 EagleNews 系统 Web 客户端界面

图 5-8 展示的是基于该系统构建的应用，网络搜音机，该设备为移动终端设备，通过 GPRS 方式连接互联网，使用语音合成技术将系统推荐的新闻朗诵给用户收听。图 5-9 是在手机上的个性化网络新闻推荐终端。

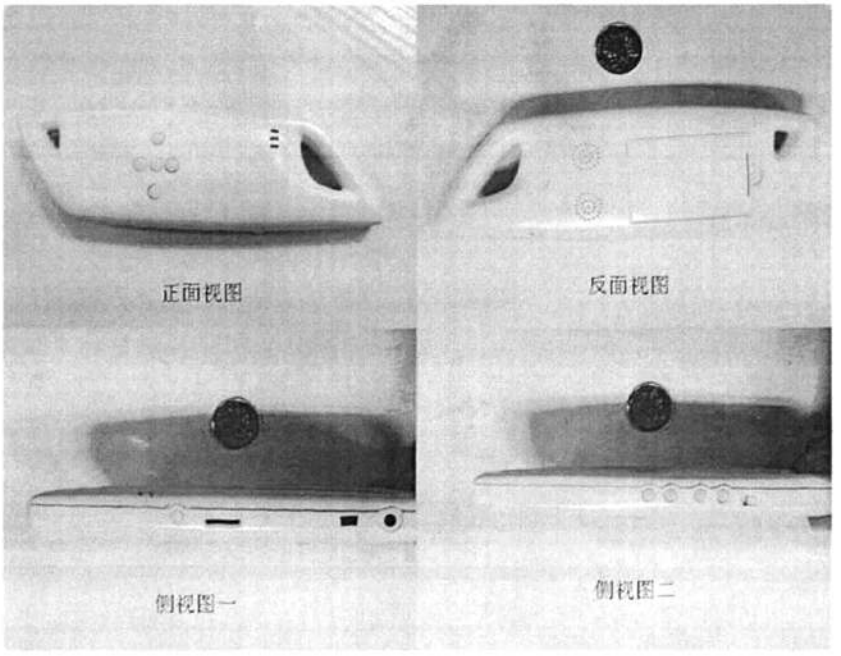

图 5-8 视障人群专用个性化网络新闻收听终端

图 5-9 智能手机上的个性化网络新闻推荐终端

5.6 本章小结

本章介绍了 EagleNews 的系统架构，以及该架构的特性：EagleNews 是一个高性能、资源自适应的通用个性化新闻推荐系统，该系统使用隐式用户数据收集方式，在不给用户使用带来任何压力的情况下，学习用户的使用习惯，形成动态的个性化推荐结果，除此之外，该系统使用通用通信协议，和通用标准数据交换格式，便于扩展和整合到其他应用中。除此之外，本章还介绍了该系统的详细设计和各个模块的设计要点。最后，展示了基于该系统构建的应用，基于 web 的个性化新闻阅读网站，和为视障人群设计的网络新闻阅读终端。

第6章 总结和展望

6.1 论文工作总结

随着信息爆炸时代的来临，互联网用户越来越难以精确有效地从互联网获得信息资讯。因此，催生了个性化推荐系统的发展。到现今，个性化推荐系统已经进入到了人们生活的方方面面。

我国互联网普及率不断提高，网络用户数量激增，其中网络新闻的用户增长速度更是迅猛，然而，国内现在除了少数的新闻门户网站之外，完善的个性化网络新闻推荐系统几乎没有。现有的系统一般通过与用户交互形成静态用户模型为用户推荐新闻，这样的系统无法动态学习用户的使用习惯，并将学习结果反映到推荐结果中，不能取得良好的使用体验。

本文设计了个性化网络新闻推荐系统，并且使该系统能够适应商业应用的需要，有着较好的推荐性能和推荐效果。学术界于个性化推荐系统的性能问题研究较少，普遍关注推荐系统的推荐效果，如全面性和准确率，而在实践中，个性化推荐系统的性能，从一定程度上决定了该系统的价值。而本文研究了个性化推荐系统的性能瓶颈所在，提出了资源自适应的推荐算法改进，使系统能够在一定负载之下，推荐效果和系统性能之间取得平衡。

除此之外，本文提到的系统中，为每个用户建立了用户模型，并通过分析用户的使用习惯，不断更新用户模型，使得用户模型能够反映用户的兴趣变化，实现真正的个性化推荐。本文研究了用户模型构建时，长期兴趣和短期兴趣的特点，提出了时间衰减淘汰算法，使得系统可以仅使用一个用户模型，就使推荐结果同时反映用户的长期兴趣和短期兴趣，节约了系统资源，提高了系统性能。

最后，本文介绍了基于上述方法的系统设计。和各个系统模块设计与实现的关键点。

6.2 未来研究工作展望

在本文的系统之中，采用了隐式的用户信息收集方式，使用该种方式虽然能够减轻用户的压力，提高使用体验。但是这种方式却在灵活性上有所欠缺。并且

该方法是通过启发式的算法来学习用户的兴趣的，难免产生误差。

在实际推荐系统的用户中，不乏愿意参与用户模型构建的用户，未来的系统设计，可以形成显示用户建模与隐式用户建模相结合的方式，这其中可能会涉及到用户模型表示方法的重新设计，和用户模型更新策略的重新设计。

除此之外，还可以考虑在系统中添加允许用户表达评分的接口，使用户可以有选择地对阅读过的新闻进行评分，这样做，可以更准确地把握用户真正的偏好。这其中可能涉及到评分方式和评分值的设定，以及如何在更新用户模型时利用此数据的研究工作。

本文采用的资源自适应算法，虽然可以使得系统可以响应比以前更多的请求，但是并没有改变算法的本质，随着数据量进一步的增大，系统性能还会再次遇到瓶颈。近年来，事件探测和追踪的研究正在兴起，在系统中融合事件探测和追踪技术，使得新闻能够按照事件进行聚类，仅将同一事件的最新进展推荐给用户阅读，这样的方式可以进一步降低待推荐新闻数量，有效提高系统性能，也因为减少了重复内容的推荐，而提高了用户体验。

参考文献

[1] Information overload. [M/OL]. http://en.wikipedia.org/wiki/Information_overload.

[2] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Trans. on Knowl. and Data Eng., 2005, 17(6):734–749.

[3] Chu W, Park S T. Personalized recommendation on dynamic content using predictive bilinear models[C]// WWW '09: Proceedings of the 18th international conference on World wide web. New York, NY, USA: ACM, 2009: 691–700.

[4] Wang J, de Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]// SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2006: 501–508.

[5] Song X, Tseng B L, Lin C Y, et al. Personalized recommendation driven by information flow[C]// SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2006: 509–516.

[6] Das A S, Datar M, Garg A, et al. Google news personalization: scalable online collaborative filtering[C]// WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM, 2007: 271–280.

[7] Singhal A. Modern information retrieval: A brief overview[J]. IEEE Data Engineering Bulletin, 2001, 24(4):35–43.

[8] Belkin N J, Croft W B. Information filtering and information retrieval: two sides of the same coin?[J]. Commun. ACM, 1992, 35(12):29–38.

[9] Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer[M]. 1989.

[10] Baeza-Yates R A, Ribeiro-Neto B. Modern Information Retrieval[M]. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[11] Mooney R, Bennett P, Roy L. Book recommending using text categorization with extracted information[C]// Recommender Systems. Papers from 1998 Workshop. Technical Report WS-98. vol 8. 1998.

[12] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine learning, 1997, 27(3):313–331.

[13] Rich E. User modeling via stereotypes[J]. Readings in intelligent user interfaces, 1998, 329–341.

[14] Goldberg D, Nichols D, Oki B, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):70.

[15] Resnick P, Iacovou N, Suchak M, et al. Grouplens: an open architecture for collaborative filtering of netnews[C]// CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work. New York, NY, USA: ACM, 1994: 175–186.

[16] Konstan J A, Miller B N, Maltz D, et al. Grouplens: applying collaborative filtering to usenet news[J]. Commun. ACM, 1997, 40(3):77–87.

[17] Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use[C]// Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co. New York, NY, USA. 1995: 194–201.

[18] Shardanand U, Maes P. Social information filtering: algorithms for automating "word of mouth"[C]// Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co. New York, NY, USA. 1995: 210–217.

[19] Breese J, Heckerman D, Kadie C, et al. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. vol 461. San Francisco, CA. 1998.

[20] Claypool M, Gokhale A, Miranda T, et al. Combining content-based and

collaborative filters in an online newspaper[C]// Proceedings of ACM SIGIR Workshop on Recommender Systems. Citeseer. 1999.

[21] Pazzani M. A framework for collaborative, content-based and demographic filtering[J]. Artificial Intelligence Review, 1999, 13(5):393–408.

[22] Billsus D, Pazzani M J. User modeling for adaptive news access[J]. User Modeling and User-Adapted Interaction, 2000, 10(2-3):147–180.

[23] Tran T, Cohen R. Hybrid recommender systems for electronic commerce[C]// Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00. vol 4. 2000.

[24] Smyth B, Cotter P. A personalized television listings service[J]. 2000.

[25] Basu C, Hirsh H, Cohen W. Recommendation as classification: Using social and content-based information in recommendation[C]// Proceedings of the National Conference on Artificial Intelligence. JOHN WILEY & SONS LTD. 1998: 714–720.

[26] Burke R. Hybrid recommender systems: Survey and experiments[J]. User Modeling and User-Adapted Interaction, 2002, 12(4):331–370.

[27] Garrigós I, Gómez J, Cachero C. Modelling dynamic personalization in web applications[J]. Lecture notes in computer science, 2003, 2722:472–475.

[28] Popescul A, Ungar L, Pennock D, et al. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments[C]// 17th Conference on Uncertainty in Artificial Intelligence. vol 458. Citeseer. 2001.

[29] Schein A, Popescul A, Ungar L, et al. Methods and metrics for cold-start recommendations[C]// Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM New York, NY, USA. 2002: 253–260.

[30] Condli M, Lewis D, Madigan D, et al. Bayesian mixed-effects models for recommender systems[C]// Conference SIGIR-99 Workshop on Recommender Systems: Algorithms and Evaluation. Citeseer. 1999.

[31] Ansari A, Essegaier S, Kohli R. Internet recommendation systems[J]. Journal of

Marketing Research, 2000, 37(3):363–375.

[32] Google news homepage 2010. http://news.gogole.com.

[33] Yahoo! news homepage 2010. http://news.yahoo.com.

[34] Luo G, Tang C, Yu P S. Resource-adaptive real-time new event detection[C]// SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2007: 497–508.

[35] Gauch S, Speretta M, Chandramouli A, et al. User profiles for personalized information access[G]// The Adaptive Web. vol 4321. Heidelberg, Berlin: Springer, 2007: 54–89.

[36] Kim H R, Chan P K. Learning implicit user interest hierarchy for context in personalization[C]// IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces. New York, NY, USA: ACM, 2003: 101–108.

[37] Moukas A. Amalthaea: Information discovery and filtering using a multiagent evolving ecosystem[J]. Applied Artificial Intelligence, 1 July 1997, 11:437–457(21). http://www.ingentaconnect.com/content/tandf/uaai/1997/00000011/00000005/art00004.

[38] Magnini B, Strapparava C. User modelling for news web sites with word sense based techniques[J]. User Modeling and User-Adapted Interaction, 2004, 14(2-3):239–257.

[39] Trajkova J, Gauch S. Improving ontology-based user profiles[C]// Proceedings of RIAO 2004. Citeseer. 2004: 380–389.

攻读硕士学位期间主要的研究成果

发表论文：

[1] 唐朝. 资源自适应的实时新闻推荐系统[J]. 计算机工程与设计, 2010，已录用.

软件著作权：

[1] 陈伟, 陈纯, 唐朝, 卜佳俊, 张利军, 梁雄君. 智能网络新闻广播系统软件: 中国, 2009SR00749 [P]. 2009-01-06.

致谢

在我硕士论文完成之际，对给予我指导的卜佳俊教授、王灿老师、陈伟博士以及课题组和实验室的所有成员表示真诚的感谢。卜老师渊博的知识、宽广的视野、严谨的治学态度，都给我留下了很深的印象。王老师待人宽厚和蔼，也给了我很多研究上的指导，让我获益匪浅。

其中特别要感谢的是陈伟博士，师兄踏实的作风，深厚的学术修养，都给我留下了深刻的印象，并且他始终耐心地指导我论文的写作工作，提出了许多有益的意见和建议，我能顺利完成论文，师兄功不可没。

还要感谢跟我一起研究同一课题的毛薪硕士、梁雄君硕士、李辉硕士、张海燕硕士，整个课题得以顺利开展，取得成果，都依赖于你们的工作和努力，这才有了我得以工作的基础，非常感谢你们对我工作的支持和帮助。

最后要感谢我的室友王俊峰硕士，和我同级的姜干新硕士、嵇存美硕士，感谢你们帮助我修改论文、提出意见，与你们在一起奋斗的日子将成为我难以忘却的记忆。

仅以此文献给我的父母亲，是你们始终站在我的背后，给予我最强有力的支持，我的一切成果与荣誉都属于你们。

唐朝

2010 年 1 月 26 日于求是园