**密级： 保密期限：**



**硕士学位论文**



**题目：基于大数据平台的中文文本分析系统的设计与实现**

**学 号： 2014111527**

**姓 名： 袁佳露**

**专 业： 计算机科学与技术**

**导 师： 许长桥**

**学 院： 网络技术研究院**

**2016年11月28日**

独创性（或创新性）声明

本人声明所呈交的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。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电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了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名： 日期：

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校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，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密在 年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名： 日期：

导师签名： 日期：

**基于大数据平台的中文文本分析系统的设计与实现**

摘　要

进入21世纪，互联网技术的发展突飞猛进，互联网已经与各行各业息息相关，形成了互联网+的大格局，同时，连接在互联网上的各种设备在源源不断产生着数据，造成数据的大爆炸式产生，为大数据时代的到来奠定了基础，这其中包括大量的文本信息，这些文本信息以日志、评论、文章等形式呈现在互联网上，由于互联网与人们的生活越来越紧密，并且网络舆情对社会热点的影响也越来越大，如何分析网络观点、预测网络情绪并正确引导网络舆情成为当今社会乃至全世界亟需解决的问题。其中对文本分析技术的研究是解决这一问题的关键点，由于中文的特殊性，对中文文本的分析更加困难，中文文本分析技术主要包含文本采集、文本清洗、文本标注、文本训练、文本分类这5大类技术，这5大类技术中包含了许多关键的算法和系统，通过对这些算法的优化和对系统的整合可以满足对文本分析任务的需求，其中最具代表性的任务就是对文本倾向性的分析，文本倾向性分析是指通过采集、标注、训练、分类等手段分析文本的立场、观点、情感、情绪等主观信息，从而对文本倾向性做出褒贬划分。这一技术涉及了文本分析技术的方方面面，并可以应用到评论系统、舆情系统、问答系统、推荐系统等多个领域。同时，随着技术的不断成熟，开源社区的繁荣，研究人员对海量数据的处理不再困难，各种大数据平台的出现为处理海量数据提供了方便，尤其是2009年推出的大数据计算框架Apache Spark，并在2014年成为Apache基金会的顶级项目，其不断的完善为大数据处理带来了新的曙光。

本论文基于Spark大数据平台，并以互联网上的评论信息作为研究对象，构建一套针对中文文本倾向分析的大数据分析系统，通过利用文本分析相关技术，对评论信息进行采集、过滤、标注、训练、分类等操作，最终提供一套分析评论文本正负倾向性的解决方案，并通过利用LDA算法和doc2vec算法相融合的方式提高文本倾向分析的准确率和召回率。本文主要包括以下几个方面：

首先，本论文分析了课题的研究背景和国内外研究现状，并针对中文文本分析场景提出了可行性研究，阐述了关键技术和主要思路。其次，针对中文文本分析领域详细系统的介绍了相关算法和技术，同时，还介绍了大数据平台的特点和架构，并分析了基于大数据平台开发系统的优势，以及如何在大数据平台上构建文本分析系统。然后，在次基础之上设计了中文文本倾向分析系统的总体架构，并提出了基于情感标注和标签融合的文本标注算法和基于doc2vec和LDA的文本表示算法，最后根据系统设计完成各个模块的代码编写和系统整合，并提出总结和分析。

**关键词：** Spark 文本倾向分析 doc2vec LDA神经网络 机器学习

**Research and Implementation of multicast group membership management protocol based on the New nETWORK ARCHITECTURE**

**ABSTRACT**

Multicast technologies help multicast communication service, such as remote video conference and online games, etc., to reduce the bandwidth resource, network congestion, and delay. Internet Standard Organization has published numerous IP Multicast Standard in recent years. But the IP multicast has not been deployed in large-scale. This is because the current multicast technologies adopt the open multicast service model, which does not support identity authentication. The legality of the multicast member is difficult to be guaranteed. Moreover, the mobility management for mobile terminals is also lacked. Therefore, this paper adopts the novel network architecture design to research the multicast member management technologies and proposes a multicast member management protocol. The contributions of this paper include: (1)Design the multicast user management protocol workflow;(2)Define the multicast user management protocol packet format;(3)Implement the multicast user management protocol function module;(4)Test and evaluate the multicast user management protocol.

To test the feasibility and performance of this mechanism, this paper establishes the network topology and installed related software. The function and consistency are also tested. The results demonstrate that the proposed multicast management protocol is able to achieve the access control, management and maintenance, etc. We believe the proposed protocol is capable of providing basic service assurance and reference value.

**KEY WORDS:** identifier separate mapping, multicast user, multicast management, novel network.

**目 录**

[第一章 绪论 1](#_Toc446211579)

[1.1 项目背景 1](#_Toc446211580)

[1.2 研究现状 2](#_Toc446211581)

[1.2.1 国外研究现况 2](#_Toc446211582)

[1.2.2 国内研究现况 5](#_Toc446211583)

[1.3 主要工作 6](#_Toc446211584)

[1.3.1 研究内容 6](#_Toc446211585)

[1.3.2 创新之处 8](#_Toc446211586)

[1.3.3 应用前景 9](#_Toc446211587)

[1.4 论文结构 10](#_Toc446211588)

[第二章 组播相关技术介绍 11](#_Toc446211589)

[2.1 组播工作的机理 11](#_Toc446211590)

[2.2 组播的地址划分 12](#_Toc446211591)

[2.3 IPv4组播管理协议IGMP 14](#_Toc446211592)

[2.4 IPv6组播管理协议MLD 15](#_Toc446211593)

[2.5 现有组播代理技术 16](#_Toc446211594)

[2.6 传统组播的弊端 17](#_Toc446211595)

[2.7 新型网络架构 18](#_Toc446211596)

[2.7.1 新型网络架构分析 18](#_Toc446211597)

[2.7.2 术语定义与实体介绍 19](#_Toc446211598)

[2.8 分离映射思想 20](#_Toc446211599)

[2.8.1 组播标识分离映射概念 20](#_Toc446211600)

[2.8.2 分离映射查询过程 21](#_Toc446211601)

[2.9 本章小结 22](#_Toc446211602)

[第三章 基于新型网络的组播成员管理协议设计 23](#_Toc446211603)

[3.1 组播成员管理协议设计目标 23](#_Toc446211604)

[3.2 协议总体工作流程 24](#_Toc446211605)

[3.3 组播报文格式设计 25](#_Toc446211606)

[3.3.1 基本首部格式 26](#_Toc446211607)

[3.3.2 组播成员管理消息格式 28](#_Toc446211608)

[3.3.3 组播消息分类 29](#_Toc446211609)

[3.4 组播安全认证机制 33](#_Toc446211610)

[3.5 组播加入流程 34](#_Toc446211611)

[3.6 组播离开流程 36](#_Toc446211612)

[3.7 组播代理工作机制 37](#_Toc446211613)

[3.8 组播成员管理协议功能模块划分 38](#_Toc446211614)

[3.8.1 用户层通信接口模块 38](#_Toc446211615)

[3.8.2 组播管理服务器模块 39](#_Toc446211616)

[3.8.3 组播消息处理模块 40](#_Toc446211617)

[3.8.4 数据存储模块 40](#_Toc446211618)

[3.9 本章小结 41](#_Toc446211619)

[第四章 基于新型网络的组播成员管理协议实现 42](#_Toc446211620)

[4.1 协议内核模块的运行机制 42](#_Toc446211621)

[4.2 数据包的发送与接收 43](#_Toc446211622)

[4.2.1 数据包的发送过程 43](#_Toc446211623)

[4.2.2 数据包的接收过程 44](#_Toc446211624)

[4.3 用户层通信接口实现 46](#_Toc446211625)

[4.3.1 MMS接口配置实现 46](#_Toc446211626)

[4.3.2 应用层通信接口实现 46](#_Toc446211627)

[4.4 组播管理服务器模块实现 47](#_Toc446211628)

[4.4.1 初始化接口阶段 48](#_Toc446211629)

[4.4.2 服务建立阶段 48](#_Toc446211630)

[4.5 组播消息处理实现 48](#_Toc446211631)

[4.6 数据存储模块实现 49](#_Toc446211632)

[4.6.1 数据库安装与配置 50](#_Toc446211633)

[4.6.2 数据表设计 50](#_Toc446211634)

[4.6.3 数据表操作 51](#_Toc446211635)

[4.7 协议相关结构体定义 52](#_Toc446211636)

[4.8 本章小结 55](#_Toc446211637)

[第五章 系统搭建及验证测试 57](#_Toc446211638)

[5.1 系统环境与配置说明 57](#_Toc446211639)

[5.2 测试工具介绍与开发 59](#_Toc446211640)

[5.2.1 工具介绍 59](#_Toc446211641)

[5.2.2 wireshark插件开发 60](#_Toc446211642)

[5.3 基于新型网络的组播成员管理协议测试 62](#_Toc446211643)

[5.3.1 安全认证测试 62](#_Toc446211644)

[5.3.2 协议功能测试 63](#_Toc446211645)

[5.4 本章小结 71](#_Toc446211646)

[第六章 结束语 72](#_Toc446211647)

[6.1 论文总结 72](#_Toc446211648)

[6.2 下一步研究工作 73](#_Toc446211649)

[参考文献 74](#_Toc446211650)

[致谢 77](#_Toc446211651)

# 绪论

## 研究背景及意义

进入21世纪以来，互联网已经成为人们生活工作中不可缺少的工具，人类进入了信息化时代，在过去的10年里计算机为人类提供了巨大的生产力，提高了生产效率，随着计算机科学的发展，尤其是互联网的发展，每天产生的数据量以指数级的速度增长，尤其是近几年社交网络、电子商务平台、自媒体的崛起，使得人们对数据的重视程度越来越高，数据已经成为新型的生产资料被人们发掘、利用，并且在实际的生成环境中带来效益，人类已经不可避免的从信息化时代步入了大数据时代[4]，但是大数据时代的到来既给我们带来机遇也给我们带来了挑战，如何对海量数据进行存储、过滤、处理和分析一直是工业界和学术界有待解决的问题，尤其在这些大数据中80%的数据是非结构化的文本数据，其中包括以社交媒体为代表的微博、博客、日志，以虚拟交易平台为代表的商品描述、用户评论，以新闻媒体为代表的时事新闻、讨论等海量文本信息，通过对这些信息的处理和分析可以实现情感分析、舆情分析、商品推荐、金融预测、用户行为描述、广告精准投放等具有实用价值的应用。本文针对互联网上大量需要处理的文本数据，设计并实现一种基于大数据平台的中文文本分析系统，旨在为需要处理海量文本，尤其是中文文本的用户提供一种可行的解决方案。

大数据对于全球还是比较新的课题，各个国家都有自己的大数据研究项目，尤其是以Google为代表的新型互联网公司，对大数据技术的探索和发展产生了推动的作用，Google作为搜索引擎巨头公司，在2003年开始陆续发表了三篇有关大数据的论文，分别论述了GFS[1]、MapReduce[2]、Bigtable[3]三大技术，这三篇论文揭开了大数据技术的神秘面纱，成为工业界对大数据理论的实现标杆。

同时，受到Google的启发，Apache基金会于2005年秋天引入了Hadoop项目，Hadoop作为开源项目和其稳定的架构设计迅速成为了大数据平台的佼佼者，可以说Hadoop已经成为了大数据的代名词，其最大的核心就是实现了一个高容错性的分布式文件系统HDFS[5]和MapReduce计算模型，其中前者为海量数据存储带来解决方案，后者为海量数据计算提供了支持。但是随着Hadoop在工业界的广泛使用，其为企业带来便利的同时也暴露出自身的一些局限性，例如抽象层次低、上手难度大；只提供map和reduce两种操作，表达力不足；复杂的计算需要多个mapreduce过程，job之间的依赖关系由开发者自己管理；时延高，只适合计算密集型任务；对于迭代式数据处理性能差等。

正在大家努力寻求可行的解决方案的时候，Spark[6]作为新一代的大数据框架正悄悄地走进人们的视线，Spark是UC Berkeley AMP lab开源的类Hadoop MapReduce的通用并行框架，其不但补充了Hadoop的不足，而且还集成了很多优秀的组件，例如其增加了数据库和数据框架库、流处理库、机器学习库和图形库。通过引入这些库，用户可以执行更加复杂的任务，并且spark将中间结果保存到内存中而不是磁盘上，这使得spark的效率也非常高，可以想象spark未来发展的前景是非常乐观的。但是，由于Spark技术出现的比较早，人们对spark的应用仍然处在探索阶段，如何利用spark解决现有的大数据问题仍然值得我们去研究和探索。

文本分析相对来说发展的比较成熟，现有的文本分析算法也非常多，效果也经历了工业界的考验，文本分析是自然语言处理发展的必经之路，首先对于自然语言来说，文字是表达语言的重要形式之一，计算机如何正确的理解文本内容，就相当于其可以正确的理解自然语言，在文本分析的初期，计算机理解文本的方式是通过词典映射，间接的“理解”其意思，其实是通过最大字符串匹配获得结果，这种方式其实是没有解决计算机理解自然语言的根本问题，随着人们不断的研究，基于概率统计的算法被应用到文本分析领域[7][8]，并且产生了很好的效果，文本分析进入了新的阶段，近年来机器学习、深度发掘[9][10]等新的领域发展迅速，人们又找到了文本分析新的突破口，现阶段文本分析的主要思路是结合词典和概率统计算法对语料进行初级处理，之后利用机器学习分类算法对其进行训练，最终得到比较理想的结果。

但是基于机器学习的文本分类算法对训练文本的依赖性很高，而且需要对文本的特征进行提取，不同的特征提取方式会导致结果的精确度不同，常用的的特征提取方式是采用TF-IDF作为文本的特征值进行训练，效果虽然可以满足要求，但是TF-IDF算法没有考虑文本之间上下文语境，准确度仍然有待提高，本文以文本倾向分析作为切入点采用LDA作为文本特征提取的优化算法，并结合神经网络中的doc2vec算法对文本进行训练，完成文本分析任务。采用神经网络的算法考虑了文本上下文之间的关系，可以提高文本倾向分析的精度。

同时，处理海量的文本信息，需要进行大量的计算任务，如何将这些任务并行化是文本需要解决的另一个问题，本文通过利用Spark大数据平台，对文本进行分析和处理，其中包括对文本的切分，存储，训练和分类等文本分析涉及的过程。利用大数据平台的计算能力，采用LDA作为文本特征提取算法，并结合最新的doc2vec算法对文件进行倾向分析，设计并实现针对中文的文本倾向分析系统，并大大的提高了准确率和效率。

## 国内外研究现状

文本分析是自然语言处理过程中最重要的环节，其中文本分析任务最首要的难题是如何表示文本，让计算机更好的理解文本。经过研究者的不断努力，文本表示算法已经有了突飞猛击的发展，从最初的基于词典的表示，到基于上下文的表示，最后到基于词向量的表示，文本表示算法已经满足现阶段的需求。其次需要解决的问题时如何获取文本特征，由于文本表示的特征向量往往具有高维性和稀疏性，这样的表示方法增加了处理文本的难度和成本，如何从高维特征中提取出最具代表性的特征就是文本分析需要解决的第二个难题。最后是对特征表示后的文本进行分类，完成任务需求，分类任务相对于比较容易，因为基于机器学习的分类算法已经非常成熟，但是如何使用特定的分类算法需要根据具体的需求来分析，这样才能达到理想的效果，并且进行算法模型的训练时间往往比较长，如何提高训练效率也是需要考虑的问题。

针对以上提到的三个问题，本文以文本倾向分析为切入点来阐述文本分析领域研究现状。主要通过两个方面进行综述：一、文本倾向分析研究综述，二、基于大数据平台的文本分类。

### 文本倾向分析研究综述

文本倾向分析是指通过对文本的处理获得该文本的正负情感倾向性，文本倾向分析是文本分析的重要应用领域，通过对文本倾向分析的研究可以应用到整个文本分析过程中涉及的关键技术和重要算法，文本倾向分析概念的形成经历了一个漫长的探索过程，最初是以“意见挖掘(opinion mining)”的概念在2003年被Dave发表在 www大会上的论文提出，其认为理想的意见挖掘工具是可以处理一个给定主题的搜索结果集，生成一个产品属性（质量，功能等）的列表，并汇总他们每个人的意见（褒义，中立，贬义）。与“意见挖掘”同时期出现的概念还有“情感分析(sentiment analysis)”，情感分析是由Das、Chen以及Tong在2001年发表的论文中提出的概念，用来形容在市场情绪分析领域，自动分析预估性文本和预测性评论的方法。在2002年的ACL大会上和2003年的EMNLP大会上，情感分析这一概念再次被多为学者作为论文的主要论点进行发表。之后几年意见挖掘、情感分析、倾向分析成为了文本分析领域和自然语言处理领域的热点研究课题，并成为文本分析领域的主流研究方向。

在国内，文本倾向分析也获得了学术界的关注，各种以文本倾向性分析为主要主题的计算机会议也纷纷在国内举办，其中比较有代表性的会议有2008年，由中文信息学会信息检索专委会主办的中文倾向性分析评测（Chinese Opinion Analysis Evaluation, COAE2008）会议，该会议提出了关于中文倾向性分析的主要任务，阐述了中文倾向性分析的难点和要点。由于中文与英文的不同，没有固定的分割符区分单词，而且中文存在歧义导致中文分词与英文分词的难度不在一个级别，而文本倾向分析又依赖于文本切分和处理的精确性，所以中文的倾向性分析相较于英文来说更加具有挑战性。

可以发现，2001年以后的近几年是文本分析领域发展的重要时期，这期间在各大顶级会议期刊上发表了近百篇有关情感分析和意见挖掘的论文，同时，文本倾向分析的研究也逐渐被应用到生产环境中为人们的生活带来便利，产生这一现象的主要原因包括机器学习领域在自然语言处理中的应用，以及互联网的发展产生了大量的可用来研究的语料数据，同时，人们对软件服务质量的需求提升也促进了该领域的发展。

但是基于机器学习的文本分析算法需要训练标注好的文本获得训练模型，这种方式对文本的特征提取方式依赖很大，最近提出的基于神经网络算法的方式进行文本分析是无监督的训练方式，而且训练的文本模型的准确度更高，本文采用了doc2vec神经网络模型作为文本倾向分析任务的主要模型，并结合LDA主题模型对文本特征进行提取，产生了不错的效果。

### 基于大数据平台的文本分类综述

上一节提到了文本倾向分析领域的发展，其中涉及到如何高效的对文本特征进行分类的问题，随着互联网技术越来越成熟，互联网与人们的生活息息相关，但是随之而来的是以指数级速度增长的信息数据，普通的计算节点已经无法承受如此规模的计算任务，2003年Google公司发表了MapReduce计算模型，该论文阐述了如何使用廉价的计算节点构造计算集群来处理大数据计算任务，2005年Apache基金会开源了Hadoop分布式框架，该框架提供了一套进行MapReduce计算的分布式系统，这使得各个研究机构和公司可以应用大数据计算来处理相关的课题和业务，同时大数据也成为研究热点进入人们的视线。

2009年Spark项目在加州大学伯克利分校AMPLab成立，并在2014年成为Apache的顶级项目，其出色的运行效率以及丰富的操作迅速成为大数据计算中的佼佼者，并被应用到各个领域中，在文本分类领域，大数据平台也被应用到其中作为提高训练速度的工具，同时，企业和用户对文本处理任务的需求越来越多，也促进了以文本处理为核心的大数据平台的推出，比较具有代表性的是2016年Google推出的Cloud Natural Language API，以及2016年微软推出的Natural Language Processing Tools，这两个工具包都是针对NLP领域的并行计算工具，其中容纳了许多自然语言处理相关的算法，包括文本表示、特征提取、文本分类等相关算法。

但是，这两个工具包并没有提供针对中文文本分析接口，本文采用Spark大数据平台实现了针对中文文本的倾向性分析系统，利用spark的分布式计算能力，完成中文文本的切分、过滤、特征提取以及模型训练等过程。

## 本文的研究内容

(1).文本数据预处理研究

* 针对特定文本数据的标签过滤提取

为了获取真实的互联网数据进行模型训练，常常采用互联网爬虫的方式摄取文本数据，但是通用的爬虫对文本的处理只是简单的过滤掉html标签，保留其正文部分，这种摄取方式打乱了文本的关联性，而且增加了语料的噪声，训练的模型往往准确率低，满足不了实际的应用需求。针对以上问题，本系统设计了一种带标签过滤的定制化爬虫，这种爬虫对于特定的文本内容，只抽取与该主题相关的文本信息，并对特定标签进行过滤，保留与主题相关的特征项，这种摄取方式可以保证主题不会受到冗余信息的干扰，而且通过对特征标签的分类，可以细粒度的判定主题的特征。

* 文本数据的存储

互联网上80%的数据是文本数据，而且是非结构化的文本数据，现有的关系型数据库以及无法满足这些海量数据的存储，而且简单的将文本数据存储到本地磁盘也是不现实的，因为本地文件系统是针对现有的操作系统设计开发的，大多不支持分布式存储，而且本地文件系统的容量有限，可扩展性和可维护性差，为了解决以上问题，需要将海量文本数据存储到分布式文件系统上，HDFS是hadoop的分布式文件系统，满足对海量数据的分布式存储，而且容错性和稳定性良好，非常适合存储非结构化的文本数据，同时，HDFS可以兼容其它大数据平台系统，为之后的处理分析提供了便利条件。

* 文本语料的预处理

Spark是新一代的大数据平台，其天生优势是可扩展性强，并且兼容HDFS。相对于HDFS而言，Spark的RDD技术是更加抽象的支持容错和并行的数据结构，可以让用户显示地将数据存储到磁盘和内存中，并能控制数据的分区。同时，RDD还提供了一组丰富的操作来操作这些数据，例如map、flatMap、filter等转换操作，除此之外，RDD还提供了诸如join、groupBy、reduceByKey等更方便的操作，以支持常见的数据运算。

(2).针对文本分析的中文语料分词算法的研究与优化

中文分词是中文文本处理的第一阶段，由于中文的特殊性，词与词之间是没有分割符间隔的，这就使得中文分词过程相对于其他文字分词异常的繁琐，目前主流的分词算法是基于词库获得切分结果，由于中文分词受限于语料和词库，所以切分效果差强人意，而且分词算法是计算密集型任务，目前大多分词器都是基于单机环境设计实现的，分词效率受限于单个节点的处理能力，如果可以将分词并行化，不仅可以构建足够庞大的词库和语料库，使得分词更加准确，而且可以提高单个结点分词效率达到快速分词的效果。

中文分词并行化需要将分词中使用的算法进行并行化处理，例如最大匹配算法、字典树算法、HMM算法等，这些算法都是分词过程中解决不同问题的有利工具，如何将分词并行化，就是解决如何将这些算法并行化的问题。首先，本课题需要对中文分词中用到的各种算法进行深入研究，确定各算法的优缺点，其次，需要根据实际情况将这些算法进行并行化处理。最后，将并行化后的算法移植到大数据平台下，并通过实验验证其效果。

(3).针对文本倾向分析相关算法的研究与改进

文本倾向分析涉及到自然语言处理的诸多算法，如何让计算机理解自然语言一直是学术界探讨和研究的问题，由于自然语言的表达方式非常丰富，机器无法识别文本的语义，而仅仅是将文本看作数据处理，然而语义识别可以应用到诸多领域，对计算机科学的发展具有深远意义，目前语义识别算法大多都是潜层语义识别，利用机器学习的相关算法分析语义相关度，效果在某些特定领域还是比较理想。例如通过语言模型进行文本特征抽取获得文本的特征，再通过相应的算法进行训练最后可以得出很多语义相关性的应用。本系统针对文本的语言模型提出采用融合的文本表示方法来解决文本倾向分析问题，通过利用doc2vec模型获得文本特征向量并进行训练，来完成文本倾向性分析任务。

Spark大数据平台自身包含一个机器学习库MLIB，其中包含许多文本分析相关的机器学习算法，主要由以下部分组成： 基本的统计学算法、分类和回归算法、聚类算法、降维算法。本课题采用LDA算法对文本进行特征提取，并结合doc2vec算法对文本倾向分析进行优化，提高准确率。

(4).基于大数据平台完成中文分析系统的设计与实现

中文文本分析系统是搭载在spark大数据平台上，利用大数据平台的分布式计算能力和丰富的组件构建一个专门提供中文文本分析解决方案的系统，为用户提供命名实体识别、文本关键字提取、相似度分析、舆情和情感分析等文本处理里常用的功能，并且满足用户对大数据存储和访问的需求。

本课题需要熟练的掌握Spark大数据平台的使用和原理，并且可以在其基础上进行二次开发，充分理解Spark的spark streams、spark SQL、spark MLIB、spark Graphx库，尤其是Spark RDD和spark MLIB，并且掌握spark的任务下发和管理模型，spark API编程等技术要点，同时，需要对中文文本处理相关算法非常熟悉，可以根据spark的特点以spark的编程接口实现这些算法，这些都是本课题实现大数据中文文本分析系统的基础。

## 本文的组织结构

根据以上阐述的内容，本文的组织结构安排如下:

第一章 绪论部分。本章主要对文本分析理论的研究和大数据平台背景进行阐述，并分析了国内外在相关领域的研究成果和本课题的创新之处，最后结合研究背景归纳了本文的主要研究内容。

第二章 中文文本分析算法及关键技术。本章主要是对中文文本分析算法和关键技术的介绍，详细的介绍了和本课题相关的文本分析算法的细节，通过本章节的介绍可以为后面的章节打下理论基础。

第三章 应用大数据平台处理文本数据。本章主要介绍Spark大数据平台的特点及内部原理，通过本章的介绍，阐述了为什么要使用大数据平台处理文本数据，以及如何使用大数据平台处理文本数据来提高计算效率。

第四章 中文文本倾向分析系统的设计。本章介绍了中文文本倾向分析系统的总体设计思路，以及总体框架设计。通过对文本倾向分析系统进行需求分析，总体框架设计，倾向分析算法设计三个小结来阐述整个系统的设计过程。

第五章 系统实现及验证测试。本章节给出了整个系统的实现过程，通过对各个模块的分析以及核心代码展示来阐述系统的实现过程。其中包括文本预处理模块的实现，文本存储模块的实现，文本分析模块的实现。并对系统进行验证，给出结果分析。

第七章 结束语。本章是对整个课题的总结，并给出对研究课题的进一步展望。

# 中文文本分析算法及关键技术

## 文本分析相关知识点介绍

文本分析中涉及了许多自然语言处理的专业术语，这些术语是研究文本的基础，本小结通过三方面来讲解这些关键知识点，其中包括语言模型及文本表示方法、文本特征及特征抽取方法、文本分类及分类模型介绍。这三个方面是进行文本倾向分析过程中必须经历的三个关键过程，通过对文本的一系列处理，最终可以得到训练后的模型，之后可以通过系统提供的接口，根据具体的需求对训练模型进行应用，完成实际的任务需求。

### 语言模型及文本表示方法

(1). 语言模型

语言模型是用来对自然语言进行建模的方式，传统的语言模型是统计语言模型（Statistical Language Model）,它是一个表示语言片段的概率分布函数，其数学表达形式如下：

其中W==表示由T个词按顺序构成的语言片段，则p(W)代表这些词组合在一起的概率。根据贝叶斯公式，可以将p(W)链式的分解为：

统一将每个词的上下文记作Context，则可得到(1)式的表达形式。根据对Context的划分策略不同，可以形成不同的语言模型，常见的有n-gram模型、n-pos模型、决策树模型、最大熵模型、最大熵马尔科夫模型、神经网络语言模型等方法，对于文本的不同建模方式会有不同的效果，同时各语言模型也各有特点，以下是对各语言模型的介绍：

* N-gram模型

n-gram模型认为某个词出现的概率与其前面n个词有关，最极端的情况是n=1的时候，即某个词出现的概率只与该词本身有关。这种语言模型称为上下文无关模型，即:

当n>=2时是称为上下文相关模型，在一般应用中取n=2或n=3，即Bigram或Trigram。n-gram模型的优点是考虑了前n-1个词的因数，而这n-1个词在自然语义上有很强的意义，同时采用Bigram或Trigram的方式可以大大简化计算规模，提高效率。但是n-gram语言模型自身也有一定的局限性，例如：n-gram模型依据语料的关系，语料不足训练的结果一般不理想，而且这种模型忽略了词之间的相似关系，只考虑了词与上下文的关系而没有考虑词与词之间的关系，其次n-gram模型在有些元组没有出现过的情形下会出现统计概率为0的情况，这会导致整个语言序列的概率为0，出现这种情况往往需要进行修正才能获得准确的结果。

* N-pos模型

n-pos模型是基于n-gram模型的基础上衍生的一种语言模型，n-pos模型是基于这样的假设：单单考虑前n个词不足以表示当前词的特征，而自然语言中的词语搭配往往是根据词的语法功能决定的，所以n-pos将当前词的前n个词按照语法功能进行分类，由这些词来决定当前词的概率，该分类叫做Part-Of-Speech，即n-pos算法的由来，n-pos模型的条件概率公式如下：

c是词性映射函数，表示将单词映射为其所属的词性类别，如果一段语言序列有T个词，K种词性分类，则可以将n-gram的条件概率求解过程从可能变成种，大大的提高了计算效率，而且这种效率的提高也不会影响准确率的下降。

* 决策树模型

决策树模型是为了解决n-gram模型中对相似的词类搭配重复求解导致的效率损失问题，构造一颗决策树表示当前词的上下文，由对根节点提问的不同回答进入子节点，直至叶子节点，从而得到当前词上下文的分布信息。为构造决策树，需要预先定义一个关于上下文信息的问题集和一个评论问题的优劣函数。

决策树模型的优点是该模型是一种通用的语言模型，n-gram模型、n-pos模型都可以用决策树模型表示出来，而且决策树中的分布信息不是固定好的而是根据训练语料库得到的，但是，由于决策树模型的抽象性导致这种模型的构造难度偏大而且时间复杂度和空间复杂度都很高。

* 最大熵模型

最大熵模型的基本思想是：对一个随机事件的概率分布进行预测时，若概率模型需要满足一些约束，则在满足约束的情况下对未知的情况不做任何假设，这种情况下，概率分布最均匀，得到的概率分布的熵也最大。最大熵语言模型的概率分布公式如下所示：

其中是参数，Z(context)为归一化因子。

* 神经网络语言模型

神经网络语言模型(Neural Network Language Model)是近几年比较流行的语言模型，即通过神经网络算法来拟合语言模型的方式求概率分布的最大似然估计，NNLM采用的是Distributed Representation来表示词向量，神经网络语言模型与n-gram、n-pos、决策树模型、最大熵模型最大的区别是前者是基于统计的语言模型，后者是基于深度学习的语言模型，这些模型所关注的结果都一样，但是建模思路不尽相同。

常用的NNLM模型有C&W的SENNA、M&H的HLBL、Mikolov的RNNLM，其中比较经典的是Bengio等人在2001年发表在NIPS上的文章《A Neural Probabilistic Language Model》，该文章用一个三层的神经网络模型来构建语言模型，这三层网络分别是输入层，即以词向量首尾相连得到一个很长的向量作为输入向量，第二层是隐藏层，使用d+Hx计算所得，d是一个偏执项，然后使用tanh作为激活函数，最后一层是输出层，共有|V|个节点，其中|V|表示词表的大小，每个节点表示下一个词为i未归一化log概率。最后使用 softmax 激活函数将输出值 y 归一化成概率。整个模型的计算公式为:

式子中的 U是隐藏层到输出层的参数，整个模型多数的计算集中在 U和隐藏层的矩阵乘法中。式子中还有一个矩阵 W，这个矩阵包含了从输入层到输出层的直连边。直连边就是从输入层直接到输出层的一个线性变换，如果不需要直连边的话，将 W置为 0 就可以了。

(2).文本表示方法

文本表示方法是指将文本向量化，转换成可以计算的数学矩阵，一般是将文档作为矩阵中的行(如上图)，单词作为矩阵中的列，而元素就是需要数字化的表示方式，在机器学习领域常用的表示方法是Bag-Of-Word(BOW)模型和Bag-Of-N-gram模型，而在深度学习中常用的模型是Word Representation 或Word Embedding, 后者通常叫做词向量模型。以下是对文本表示方法的介绍：

* 词袋模型

词袋模型是进行文本分类中经常用到的文本表示方式，因为这种方式方便快捷，易于理解，但是同时词袋模型忽略了词语之间的顺序，所以词袋模型一般应用在那些对词序要求不高的文本分析领域，词袋模型通常需要一个词典保存文档集合中所有出现过的单词，并对单词进行编号，然后文档可以表示成一个一维向量，其中向量维度为词典的大小，向量元素一般单词的出现的频率(也可以用TF-IDF表示)，以下通过例子来讲解词袋模型的原理。

假设有两个文档，文档1为“我 爱 北京 天安门， 我 是 中国人”，文档2为“她 喜欢 看 电影， 她 是 美国人。”则根据文档1和文档2可以构造一个词典：{1:“我”，2:“爱”，3:“北京”，4：“天安门”，5：“是”，6：“中国人”，7：“她”，8：“喜欢”，9：“看”，10：“电影”，11：“美国人”}，根据词典大小可以将文档表示成为一个11维的向量，其中每个元素是单词在词典中出现的频率。

文档1：[2,1,1,1,1,1,0,0,0,0,0]

文档2：[0,0,0,0,2,0,1,1,1,1,1]

为了提高词袋模型的实用价值可以将文档中的元素用单词的TF-IDF值表示，这样可以提高单词和文档的特征关系。

* Ngram词袋模型

以上提到的词袋模型是最简单的文本表示方式，其没有考虑单词与单词之间的搭配关系，这种表示方式丢失了很多文本特征，在文本分类过程中使用这样的模型得到的准确率往往不高，Ngram词袋模型是对这种模型的优化，即使用n-gram语言模型得到单词的条件概率，用该条件概率表示单词和文档，这种表示方式保留了单词之间的搭配关系，但是无法获得单词和文档之间的关系，即无法获得语义相近的单词，丢失了文档与文档的特征，模型不够平滑。为了使得模型更加平滑，还需要采用平滑处理来优化模型。

* 词向量模型

词向量是指将自然语言中的词用数学化的向量表示，这样做的好处是可以对词进行数学公式计算，把计算机无法理解的文字数字化，常见的对词向量的划分有one-hot representation，就是用一个只有一项为1其他项全为0的长向量表示一个词，向量的维度N为词典的大小，向量中为1的项对应该词在词典中的索引，例如对于“中国”这个词，假设其在字典中的索引为2，这”中国”可以表示成如下的词向量：

这种词向量表示方法虽然简单，但是其只标识了单个词的信息，向量十分稀疏，而且向量的维度很高，这样会导致维数灾难，不适合作为计算机运算的输入，这种表示方式还有一个问题是向量中没有词与词之间的关系，即词的上下文关系，所以在训练过程中会丢失许多特征信息，其数学意义不大，所以很少被使用。

另一种对词向量的表示是Distributed Representation，其基本的思想是：通过训练将自然语言中的词映射成为一个长度固定的向量，该向量维度一般比较短，这样所有的词向量可以构成一个向量空间，同时由于向量长度较短，也有利于对其进行相应的数学运算从而研究它们之间的关系。这种词向量表示方式比较符合计算机输入参数的要求，且自带平滑，保留了词的上下文特征和词与词之间的特征，是比较理想的文本表示模型。

### 文本特征及特征抽取

文本特征是表示文本的基本单位，文本的特征抽取是文本挖掘领域中重要的研究课题，对于文本来说其特征一般都具备一定的特性。首先，文本特征项需要能够标识文本的主要内容。其次，特征项需要代表文本的特征属性，即与其他文本的特征有明显区别。最后，特征项的维度不宜过多，容易导致维度灾难。在中文领域，中文文本由字作为基本单元组成，但是字所包含的特征不足够表示文本的特征，相对于单个字来说，词和短语所包含的信息更加丰富，所以一般采用词或短语作为文本的基本特征单元。如果用词或短语作为文本的特征项，随着文档的规模越来越大，特征向量包含的特征值就会非常巨大，这样会导致维度灾难，无法进行计算。这时就需要借助特征抽取技术来完成文档的特征抽取(或文档的稀疏表示)，特征抽取需要完成的任务是在不丢失文本核心信息的情况下，用更少的特征表示文本，即将原来稠密矩阵用稀疏矩阵进行表示。

特征抽取的方式一般有3种：（1）通过映射或变换的方法减少特征的维度，（2）通过排序获取最具代表性特征值，（3）用机器学习的方法训练文本获取特征模型。以下是基于统计的特征提取方法：

1. TF-IDF:

TF-IDF(term frequency–inverse document frequency)算法是用来衡量单词权重的统计方式，其中TF指为单词频率，IDF为逆文档频率，即该词出现在文档中频率的倒数，该参数用于计算该词区分文档的能力。TF-IDF之所以能作为衡量单词权重的特征是因为其基于这样的思想：如果单词在一篇文档中出现的频率很高而在其他文档里出现的概率很低，那么说明这个单词对这篇文档很重要，也就是该单词能作为这篇文档的特征。TF-IDF通常用来提取文档的关键词特征，即通过排序获取TF-IDF值高的单词作为文档的关键词来代表文档的特征。

1. 互信息（Mutual Information）

互信息(Mutual Information)是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。互信息的定义如下：

其中X,Y为随机变量，其联合分布为p(x,y),边际分布为p(x),p(y),互信息I(X,Y)是联合分布p(x,y)与乘积分布p(x)p(y)的相对熵。

使用互信息理论进行特征抽取是基于如下假设:在某个特定类别出现频率高,但在其他类别出现频率比较低的词条与该类的互信息比较大。通常用互信息作为特征词和类别之问的测度，如果特征词属于该类的话，它们的互信息量最大。由于该方法不需要对特征词和类别之问关系的性质作任何假设，因此非常适合于文本分类的特征和类别的配准工作。

1. 信息增益方法（Information Gain）

信息增益是信息论中的一个重要概念, 它表示了某一个特征项的存在与否对类别预测的影响, 定义为考虑某一特征项在文本中出现前后的信息熵之差。某个特征项的信息增益值越大,贡献越大, 对分类也越重要。信息增益方法的不足之处在于它考虑了特征未发生的情况。特别是在类分布和特征值分布高度不平衡的情况下, 绝大多数类都是负类, 绝大多数特征都不出现。此时的函数值由不出现的特征决定, 因此, 信息增益的效果就会大大降低。信息增益表现出的分类性能偏低。因为信息增益考虑了文本特征未发生的情况，虽然特征不出现的情况肿可能对文本类别具有贡献，但这种贡献往往小于考虑这种情况时对特征分值带来的干扰。

1. 遗传算法（Genetic Algorithm, GA）

文本实际上可以看作是由众多的特征词条构成的多维空间,而特征向量的选择就是多维空间中的寻优过程,因此在文本特征提取研究中可以使用高效寻优算法。遗传算法(Genetic Algorithm, GA)是一种通用型的优化搜索方法,它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素,复制出最佳代码串,并使之一代一代地进化,最终获得满意的优化结果。在将文本特征提取问题转化为文本空间的寻优过程中,首先对Web文本空间进行遗传编码,以文本向量构成染色体,通过选择、交叉、变异等遗传操作,不断搜索问题域空间,使其不断得到进化,逐步得到文本的最优特征向量。

1. 主成分分析法（Principal Component Analysis，PCA）

主成分分析法不是通过特征选取的方式降维的，而是通过搜索最能代表原数据的正交向量，创立一个替换的、较小的变量集来组合属性的精华，原数据可以投影到这个较小的集合。PCA由于其处理方式的不同又分为数据方法和矩阵方法。矩阵方法中，所有的数据通过计算方差一协方差结构在矩阵中表示出来，矩阵的实现目标是确定协方差矩阵的特征向量，它们和原始数据的主要成分相对应。在主成分方法中，由于矩阵方法的复杂度在n很大的情况 以二次方增长，因此人们又开发使用了主要使用Hebbian学习规则的PCA神经网络方法。

主成分分析法是特征选取常用的方法之一，它能够揭示更多有关变量\_丰要方向的信息。但它的问题在于矩阵方法中要使用奇异值分解对角化矩阵求解方差一协方差。

1. n-gram算法

n-gram算法的基本思想是将文本内容按字节流进行大小为N的滑动窗口操作,形成长度为n的字节片段序列。每个字节片段称为gram,对全部gram的出现频度进行统计,并按照事先设定的阈值进行过滤,形成关键gram列表,即为该文本的特征向量空间,每一种gram则为特征向量维度。由于n-gram算法可以避免汉语分词的障碍,所以在中文文本处理中具有较高的实用性。

特征提取是为了获取文本的特征，需要根据特定场景需求选择特定的特征提取方式，对特征的提取不局限于单一的方法，往往需要通过多种方法对基础语料进行处理才能获得比较理想的特征项，本课题采用特征抽取的第三个方法，即通过机器学习方法训练文本获取特征模型，本文采用LDA模型作为文本特征提取算法，LDA算法是以N-gram算法为基础的文本表示算法，其可以将文本进行稀疏表示获得文档与主题模型，该模型可以作为文档特征进行文档分类任务。

## 基于主题模型和文本向量的文本倾向分析算法

本节主要介绍本课题使用的文本倾向分析算法，通过上一个章节的介绍，对文本分析领域的基本概念有一定的了解，文本分析需要对文本建模，并提取特征，之后需要对文本进行训练获取模型，然后通过文本分类算法对文件进行分类，获取文本的正负倾向性。本文采用LDA模型获取文本主题作为文本特征的一部分，然后结合神经网络模型训练获取词向量，并训练得到word2vec模型，最后通过随机梯度下降算法对文本进行分类，以下是对这三个算法的原理介绍。

### 基于LDA的文本特征提取算法

隐含狄利克雷分布（Latent Dirichlet Allocation，简称LDA）是一种主题模型，该模型可以将文档中的主题以概率分布的形式给出，从而可以通过分析文档抽取出文档的主题，我们利用LDA算法的这一特性可以作为文本特征抽取的一种方法，因为对于文档来说，其主题就是该文档的特征，同时，LDA算法是基于N-gram词典模型的，即其认同一篇文档是有一组词构成，词与词之间没有先后关系。LDA的主要思想是：

给定一篇文档，该文档的生成方式如下：

* 从Dirichlet分布中取样生成文档i的主题分布
* 从主题的多项式分布中取样生成文档i第 j 个词的主题
* 从Dirichlet分布中取样生成主题对应的词语分布
* 从词语的多项式分布中采样最终生成词语

其中，Dirichlet分布是多项式分布的共轭先验概率分布。

假设有一篇文档W，其由N个词组成，记作W=(w1,w2,w3,…,wn)。用p(wn)表示词wn的先验概率，则生成文档W的概率为：

其图模型为：



图中灰色部分w表示可观测变量，N表示一篇文档中共包含N个单词，M表示有M篇文档。则可以得到LDA算法的概率模型图如下所示：



其中，其中表示词分布，表示主题分布，是主题分布的Dirichlet分布的参数，是词分布的Dirichlet分布的参数，N表示单词总数，M表示文档总数。

LDA算法根据先验知识确定文档的主题分布，然后从该主题分布中抽取主题Z，然后根据先验知识获得在主题Z下的词语分布，最后从主题Z对应的词语分布中抽取单词w，重复以上过程N词，就生成了文档W。是Dirichlet分布的两个参数，即LDA算法通过Dirichlet分布来估计文档的主题分布和词分布，获取文档的概率。

由上所述，我们知道基于N-gram词袋模型的文本是服从多项式分布的，而Dirichlet分布正好是多项式分布的先验分布，同时LDA模型是从pLSA主题模型思想上衍生出的模型，所以在深入了解LDA模型之前我们需要了解pLSA模型和Dirichlet分布。

1).概率潜在语义分析(probabilistic latent semantic analysis，简称pLSA)

pLSA是一种文本主题模型,文本主题模型的主要思想是：一篇文章由多个主题构成，多个主题出现的概率不同，每个主题中包含多个单词，每个单词出现的概率也不同，则生成一篇文章的过程是首先选择一个主题，然后从该主题选择一个词，重复执行N，则可以生成一篇包含N个词的文章。上述过程抽象后就是PLSA模型，则可得到PLSA模型描述如下：

定义：

表示从海量的文档中选取某篇文档的概率。

表示在给定文档的情况下出现词的概率。

是一个可计算的概率，对于每个词语，表示该词语在该文档出现的频率除以文档中词语的总数。

表示在给定文档的情况下包含主题的概率。

表示在给定主题的条件下出现某个词的概率。

则根据以上定义可以得到生成一篇文档的步骤如下：

以概率选择一篇文档

以概率从主题分布中选择一个隐含的主题类别。

以概率从词分布中选择一个单词。

LDA模型是在PLAS模型的基础上增加了两个Dirichlet先验估计，LDA模型生成文档的方式是：

以概率选择一篇文档。

从Dirichlet分布中以为参数取样生成该文档的主题分布。

从主题分布中选择文档第j个词的主题类别。

从Dirichlet分布中以为参数取样生成主题为对应的词语分布。

从词分布中选择一个单词。

接下来，我们就需要知道为什么选择Dirichlet分布作为主题分布和词语分布的先验概率分布。

2).Dirichlet分布

在正式了解Dirichlet分布之前，首先我们从最简单的二项分布入手，二项分布即重复n次独立的伯努利试验。在每次试验中只有两种可能的结果，而且两种结果发生与否互相对立，并且相互独立，与其它各次试验结果无关，事件发生与否的概率在每一次独立试验中都保持不变，则这一系列试验总称为n重伯努利实验，当试验次数为1时，二项分布服从0-1分布。二项分布的概率密度函数为：

其中，对于,是二项式系数，记为C(n,k)。

把二项分布推广到多个互斥事件就得到了多项分布，多项分布的概率密度函数为：

而Dirichlet分布的概率密度为：

其中:

，（即Gamma函数）。

由证明可得(证明略)：

其中Dir表示Dirichlet分布，MultCount表示多项分布，即上式为Dirichlet-Multinomial 共轭，也就是说Dirichlet分布和MultCount分布为共轭关系。由共轭先验的定义可知，Dirichlet分布是MultCount的共轭先验分布，同时，基于N-gram的文本主题模型其实质是一个多项式分布，所以LDA模型采用Dirichlet作为主题模型的先验概率分布，来估计主题参数。

综上所述，使用LDA模型可以获取文章的主题分布，而文章主题分布是标识文章特征的更加高级的特征项，我们可以使用LDA模型提取出文本的主题作为特征值，这些特征值可以作为文本分类任务的参数，完成有关文本分析的需求。

### 基于word2vec的文本表示算法

上一节中提到了NNLM模型，word2vec是基于NNLM模型训练语言模型的算法，Word2vec算法是通过神经网络语言模型拟合自然语言概率模型，从而获得计算语言概率模型的高效方法，在训练过程中，word2vec会获得语料库中每个词的词向量，该词向量是Distributed Representation类型的词向量，其维度由输入参数决定，一般在100-500维。根据词向量我们可以得到一个词向量空间，该词向量空间包含了词本身的特征以及词与词之间的特征，可

本节主要阐述word2vec的算法原理，word2vec中主要运用了CBOW模型(Continuous Bag-of-Words Model)作为其语言模型，其中CBOW模型是在已知当前词的上下文,,,的前提下预测当前词的概率。由上一节可知自然语言的语言模型为（1）式所示，而基于神经网络的语言模型通常取其对数似然函数作为目标函数：

(2)

word2vec构造了一个三层的神经网络：输入层、投影层和输出层，这三层的含义分别是：

(1).输入层：包含Context()中2c个词的词向量，即：

v(),…,v(),v(),v(),v(),…,v()。

(2).投影层：将输入层的向量做求和累加，即:

(3).输出层：输出层对应一课Huffman树，Huffman树是以语料中各词的词频为权值构造而成的，其叶子结点为语料中的词。

构造了以上神经网络模型后，主要的问题是解决如何通过以上神经网络模型来拟合条件概率，word2vec将其视作一个二分类问题，利用逻辑回归的方法求每个结点的概率，即从Huffman的根结点到叶子结点的过程，将左结点分为负类，右结点分为正类，利用逻辑回归，易知，一个结点被分为正类的概率是：

显然，其被分为负类的概率就是：

其中，为sigmoid函数,其作为神经网络的激活函数，为词向量，为参数向量，其代表Huffman树中的非叶子结点向量。

根据以上定义就可以得到计算的算法：对于词典D中的任意词w,Huffman树中必定存在一条从根结点到词w的路径。路径上存在个分支，将每个分支看做一次二分类，每一次分类就产生一个概率，将这些概率乘起来，就是所需的。

于是，可以得到条件概率的公式为：

其中:

表达式中，表示路径中第j个结点对应的Huffman编码，表示路径中第j-1个非叶子结点对应的向量。

将上式带入对数似然函数(2)，可得：

其中，记L(w,j)为：

则上式即为CBOW模型的目标函数，使得该函数最大化，当超过某个阈值时，就可以用来判断语言模型是否是自然语言，Word2vec通过采用随机梯度上升法来计算该目标函数。

基于神经网络语言模型的训练，其是无监督的训练算法，不需要进行人工识别，只需要将文本切分成单个词语，将切分后的语料进行训练就可以得到语言概率模型，大大节约了训练成本，而且神经网络语言模型通过滑动窗口的模式保留了词语上下文的关系，这样对语义分析也有很重要的意义。

同时，对于word2vec神经网络语言模型中涉及的参数，都是由样本训练得到的，这有别于通常的机器学习算法，需要在训练样本前给定参数，所以词向量和概率模型都是在训练中得出，词向量作为训练过程中的副产品对解决自然语言领域中的问题有极大的参考作用，同时，其无监督的训练方式和高效的训练速度，也是word2vec算法受到人们青睐的重要原因。

## 文本分析的关键技术

文本分析是一个综合学科，其覆盖了许多知识点，除了前几节提到的有关算法方面的内容，还包含文本采集技术、文本切分技术、文本标注技术等，这些技术是进行文本分析的基础，掌握这些关键技术，在处理复杂的文本分析任务时会游刃有余，不仅可以提高效率而且还可以灵活的应对不同的需求，本节通过三个方面来对文本采集技术，文本切分技术和文本标注技术进行阐述。

### 文本采集关键技术

文本采集通常是采用爬取技术爬取网络上的数据进行分析，或者是收集自有的日志数据、数据库数据进行分析，网络爬虫的优点是可以根据需求收集各种类型的文本数据训练模型，缺点是网络上的文本数据大多是没有进行标注的文本，而且存在大量的网络用语，需要进行清洗才能满足训练要求。

网络爬虫一般分为通用爬虫和定制爬虫两种类型，通用爬虫一般是为搜索引擎等系统设计的爬虫，其目的是快速的爬取互联网上的内容，而定制爬虫一般是针对某一特定领域或网站进行爬取，目的是获取特定的资源，在效率方面定制爬虫不如通用爬虫那样高效，但是在质量上，定制爬虫比通用爬虫更加出色。

作为文本采集任务，一般采用的是定制爬虫来采集文本信息，在效率容忍的范围内可以获取到符合需求的训练语料，定制爬虫一般都具有对特定的网页内容进行过滤的能力，这样可以精确的获取的想要的文本内容，达到初次过滤文本的目的，例如对于网站的评论信息，定制爬虫可以将评论主题，评论时间等有关评论的文本爬取下来，并对个字段分别保存，而通用爬虫一般只是单单的爬取整个网页，或者是过滤掉HTML标签，这样获取的语料带有许多无关信息，对模型训练或文本进一步分析都是额外的噪声，所以文本采集任务更适合使用定制爬虫。

### 文本切分关键技术

文本切分关键技术是针对中文文本而言的，因为中文文本没有分隔符来区分文本之间的单词，所以需要进行分词处理，将整个文本切割成由单词组成的文本序列，中文分词技术是自然语言处理里非常重要的技术，同时，也是进行文本分析的前提条件，常见的分词算法是基于词典和规则的最大匹配算法，这类算法一般需要有强大的词库用来进行分词，但是，对于词典中不存在的词，这类算法就会失去作用，而且对于存在歧义的文本序列，这类算法同样不能达到满意的效果，于是，出现了基于概率统计的分词算法，其中最经典的是使用HMM算法进行分词，HMM算法可以消除对歧义词语切分不准的问题，而且还可以切分出未登录的词，所以现代的中文分词程序都是结合两种方法进行分词，而且分词效果也十分出色，一般能达到98%的准确率。

### 文本标注关键技术

文本标注一般是在文本分类训练过程中对语料进行标注，形成熟语料，从而可以训练所需的模型，文本标注任务根据需求的不同标注方法也不尽相同，一般的标注方式有三种：人工标注，基于特征词典的文本标注，基于标签的文本标注，以下是对三种标注方式的解释：

(1).人工标注：顾名思义，人工标注就是采用人工的方式对文本进行标注，这种方式的标注成本很高，但是标注效果好，一般是对标注要求很高的需求才会使用人工标注的方式。

(2).基于特征词典的文本标注：基于特征词典的标注一般是采用一个特征词典，根据分词匹配结果计算文本的最终标注结果，特征词典一般包含特征的级性，采用线性叠加的方式计算文本片段，计算结果就是文本的标签值，采用特征词典的文本标注适用于标签比较少的请求，例如情感倾向标注就适合使用该标注方式。

(3).基于标签的文本标注：标签标注是通过文本在源文本中的标签特性进行标注，一般依赖原文本的标签值，例如在爬取评论信息时，根据评论的星级以及点赞次数，可以标注该评论文本的倾向和情感强度，这种标注方式依赖源文本的标签，如果原文本没有明显的标签就需要结合人工识别的方式指定特定的标签作为标注特征，再结合特征词典的方式会达到很好的效果。

# 应用大数据平台处理文本数据

文本处理任务是一个计算密集型任务，而且通常需要训练的语料库非常庞大，动辄几十G，上百G的语料需要训练，而且随着互联网的更新迭代速度的加快，数据的产生以指数级的规模不断的扩大，单机环境已经无法满足文本数据的处理需求，所以十分需要借助分布式计算能力来提高计算效率，随着大数据技术的发展，大数据计算平台越来越成熟，同时也越来越受到业界的青睐，使用大数据平台来处理海量文本数据已经成为许多企业首选的解决方案，本节主要通过介绍Spark大数据平台来阐述使用大数据平台的优势，并给出如何使用Spark处理文本数据的步骤。

## Spark大数据处理平台

Spark是一个围绕速度、易用性和复杂分析构建的大数据处理框架。最初在2009年由加州大学伯克利分校的AMP Lab开发，并于2010年成为Apache的开源项目之一。Spark作为现今最流行的内存分布式计算框架，其优点远远不仅仅是作为大数据的计算中心这么简单，它已经形成一套完整的大数据处理生态系统，其中包括Spark Streaming:用来处理实时数据，Spark SQL:用类SQL语句对Spark数据进行查询，Spark MLIB:可扩展的机器学习库,包括常用的机器学习算法,Spark Graphx:支持图计算和并行图计算的Spark库。

Spark大数据计算框架是建立在MapReduce基础之上的分布式计算框架，但是相较于Hadoop简单的Map和Reduce操作，Spark提供了一个分布式内存抽象数据集RDD,针对RDD还提供了更加丰富的算子操作，这些操作被划分为转换（Transformations）和动作（Action）。转换操作是对RDD进行的用来返回一个新的RDD的算子，包括map、reduce、union、filter、reduceByKey、partitionBy、join、count、sample等，动作是在数据集上经过一次计算后返回的结果。所有的Spark应用都是由驱动程序组成，这些驱动程序用来并行执行用户提交的计算任务，计算任务又由各个单向的转换操作组成，在Spark中，所有RDD的转换都是惰性求值的，旧的RDD数据经过转换后生成新的RDD集合，每个RDD又可以有多个分区，这样每次RDD转换相当于生成一个结点到另一个结点的有向无环图(DAG),由多个DAG构成一个Spark任务，同时，Spark对于DAG任务进行优化处理，确定阶段、分区、流水线、任务和缓存，数据按照DAG的方向流动，并当遇到Action类型的算子后进行实际运算，这样数据的流转都在内存中进行，大大提高了计算效率。下图展示了Spark的计算模型。



## Spark大数据平台的优势

* MapReduce

MapReduce是一种抽象的编程模型，通过将复杂的数据处理分解为Mapper和Reducer两个过程，然后将这些分解后的计算job分发到几十到几百台服务器上分布式的运行，同时对外隐藏了并发、分布、故障恢复、集群通信等细节问题，从而可以处理大数据集的计算问题。

MapReduce过程是将复杂问题抽象化，隐藏抽象处理的过程，这些抽象的细节被称为shuffle过程，shuffle过程是MapReduce过程中非常重要的一环，当问题被分解为以Mapper和Reducer构成的有向无环图(DAG)时，各个job的最终执行过程是受shuffle过程影响的，如果分解合理会大大提高执行的效率，反之，即使经过了MapReduce过程可能也不会达到预期的效果。下图中map和reduce中间的过程就是shuffle过程。



所以，MapReduce也存在一定的局限性，首先MapReduce的抽象层次比较低，只提供Mapper和Reducer两个操作，对程序员不够友好，很难上手写MapReduce程序。其次，一个复杂的任务需要分解为多个job进行计算，job的管理需要开发者自己维护，如果设计的不够合理会导致效率下降，而且处理的实际逻辑隐藏在shuffle过程中，不受开发者的管理，出现问题需要修改MapReduce过程，维护成本很高。最后，Reducer需要等待所有Mapper执行结束后才能执行，且中间结果需要保存到HDFS上，没有完全利用集群的内存资源。

针对MapReduce的局限性，研究者在MapReudce的基础上进行了许多改进，出现了如Pig，Cascading，JAQL，OOzie，Tez等技术，Spark是其中比较新的一种技术，而且基于其出色的表现能力，逐渐成为独立的大数据计算框架。

* 从MapReduce到Spark

Spark是基于MapReduce的大数据内存计算框架，其主要特点是有一个抽象的分布式内存对象RDD(Resilient Distributed Dataset)，RDD作为Spark编程模型的核心，具有很多丰富的操作，这些操作可以分为两类：Transformations和Action。Transformations包括map, flatMap, filter, union, sample, join, groupByKey, cogroup, ReduceByKey, cros, sortByKey, mapValues等，Action包括collect，count，save，lookupKey等。

Spark作为新型的计算模型相对于Hadoop来说具有很多优点，首先，Spark是基于内存的抽象计算模型，在性能上比Hadoop的MapReudece有很大的提升，2014年10月，Spark完成了Daytona Gray排序测试，该测试结果显示Spark用相对于Hadoop1/10的计算资源完成该项测试的耗时比Hadoop的MapReduce过程的耗时快3倍。以下是该次测试的实际结果对比：



其次，Spark中的RDD都是惰性求值的，一个RDD会转换成新的RDD，RDD之间不是强依赖关系，对于非依赖的RDD计算，并不需要等待其运行结束就可以执行其他运算。所以Spark对迭代式计算有更好的支持。最后，Spark具有丰富的扩展，其不单单是一个内存计算模型，还包含交互式（Spark SQL），流式（Spark Streamng），机器学习（MLLIB）,图形计算（GraphX）等模块，所以Spark已经成为一个大数据计算平台，许多计算任务都可以在Spark上进行运行。下图是Spark模块的构成：



## 如何使用Spark大数据平台处理文本数据

以上两节分别介绍了Spark大数据平台和Spark的优势，本节针对如何使用Spark大数据平台处理文本数据这一问题进行阐述，根据第二章介绍的有关文本分析的知识可知，文本分析主要的计算任务是需要对文本语料进行预处理和模型训练。文本预处理包括文本切分、文本过滤、文本标注等，文本模型训练包括文本特征训练、文本特征抽取、文本分类模型训练等。通常情况下，这些文本任务都是通过不同的模块进行分别处理，而基于Spark平台的支持，这些任务可以结合在一起完成。

首先，对于初始语料的处理，Spark提供了针对Java、Scala、Python的接口，可以选择其中的任何一种语言对语料进行处理，而且Spark也提供了对语言内部数据结构和RDD相互转换的接口，所以，可以通过现有的语言编写好对语料的处理逻辑，然后根据Spark提供的接口，将处理逻辑转换成Spark内部的数据结构，即RDD内存模型，然后进行分布式的处理。下图是以初始语料集处理为例，展示如何将文本处理任务运行到spark平台上：



初始语料可以保存在HDFS上，从HDFS中将语料文本导入Spark系统，构造初始语料RDD，通过map操作将语料RDD切分成词序列保存到新的RDD里，这里记作RDD1，Spark的map操作可以通过函数将RDD映射到新的RDD上，这里的映射函数就是通过其他语言实现的分词模块，形成的RDD1实际上是一个列表，其中每个元素是一个token对象，该对象包含三个属性：word,offset,nature。其中word指单词本身，offset为该词在语料中的位置，nature指该词的词性。根据nature我们可以通过filter算子获得具有特定特征的词语的集合RDD2，然后，我们通过停用词词典将切分后的语料过滤，剔除无用的停用词，例如“的，是，了”等，形成待训练的新语料RDD3，用来后续的文本分析任务。

以上，就是对初始语料的处理，之后我们可以通过利用Spark提供的MLLIB库对语料进行训练获取语言模型，MLLIB库提供了丰富的机器学习算法，其中就包括第二章提到的用于进行文本分析的算法，对于没有实现的算法，我们同样可以使用外部的机器学习库来训练语料，然后通过将其转换成RDD在进行其他任务。

通过以上的描述，我们知道了使用Spark大数据平台处理文本是非常方便的，同时，基于Spark出色的表现力，还可以提高文本处理的效率，具有开发方便，性能卓越的优点。

# 基于电影评论的文本倾向分析系统的设计

本章以1~3三章为基础，设计并实现一套基于电影评论的文本倾向分析系统，根据前几章的讲解，文本倾向系统一般涉及文本的获取、文本的切分、文本的标注、文本的训练以及文本的分类，本系统就是围绕文本分析的这5方面设计的。同时，在系统设计过程中，采用基于情感词典和标签的融合标注算法来对训练语料进行标注，并设计了基于doc2vec和LDA的文本倾向分析算法，该算法向较于其他文本倾向分析算法，不仅提高了准确度，而且是无监督的分析算法，同时，本系统借助Spark大数据平台，将算法采用并行化的方式运行在集群上，提高算法的并行能力和训练速度。以下是对系统流程的描述：

首先，该系统通过定制化网络爬虫从豆瓣网站上爬取电影的评论信息作为训练语料，评论信息包括评论发表的日期、星级、点赞数、评论文本，并将这些信息以结构化数据的形式保存到数据库中。然后，通过标签（包括星级，点赞数）对评论文本进行正负性分类，对于无法判断正负倾向性的评论，采用情感词典的方式对其进行正负性分类。分类后的文本分为正向评论和负向评论，然后对正负评论文本进行分词处理，去掉停用词，并采用LDA算法对文本的主题进行提取，作为整个文本的主题特征。之后，对文本进行Word2vec模型训练获取word2vec模型，并结合文本的主题特征训练doc2vec模型，最后采用SGD分类算法训练分类器模型，并得到最终的文本倾向分析模型。最后，通过实验验证该模型的准确度和速度，最终实验表明采用LDA和dov2vec算法可以提升文本倾向分析的准确度，并通过使用Spark并行化算法提升模型训练的速度。

基于以上的系统流程，本章以如下结构组织来阐述整个系统的设计。

4.1节是基于文本倾向分析系统总体框架设计，主要讲解整个文本倾向系统的总体设计框图和各模块设计细节。

4.2节是基于doc2vec和LDA的文本倾向分析算法的设计，该节主要讲解文本倾向系统的倾向分析算法，该算法是系统的核心算法。

4.3节是中文文本倾向分析系统的并行化实现，主要讲解如何在spark上并行化文本倾向分析的各个环节。

## 文本倾向分析系统总体框架设计

IDGMP协议运行在新型网络架构底层通信协议IDMP基础之上，为了说明IDGMP的运行环境，本文有必要对IDMP的工作原理作以简要介绍。IDMP数据包的工作流程设计如**错误! 未找到引用源。**所示。



图4‑1 IDMP数据包的工作流程

IDMP在设计之初，采用内核模块的加载机制，主要是因为内核的模块化加载非常便于管理和维护，用户可以根据需求将IDMP协议动态地添加到linux网络内核[32]中，而不需要在操作系统启动时就直接运行协议栈；此外，内核模块设计的另一好处就是当用户不需要该项功能时可以对模块进行自由卸载，以释放内核资源。用内核模块不仅保证了linux内核代码的最小化、减少造成内核代码的膨胀度而且内核模块的动态加载和卸载也增强了调试的灵活性。当开发者每次修改完内核模块代码后，只需要重新编译内核模块即可，从而避免了对整个内核代码的编译工作。

IDMP协议作为新型网络通用的基础协议栈，需要预先安装到网络中的所有设备上，才会保障新型网络架构中所有实体的基本通信能力；其次为了使ASR支持分离映射功能，所有的ASR设备还需要安装分离映射模块。分离映射模块属于新型网络中的又一个关键模块，它主要负责底层通信协议报头中的标识的映射的查询、封装与分离等操作。如果接入网络层的数据进入核心网络层中时，分离映射模块就会对基础协议栈报头进行重新封装；如果核心网络层的数据进入接入网络层中，映射分离模块就会负责基础协议栈报头的卸装工作。

## 数据包的发送与接收

### 数据包的发送过程

IDGMP消息数据包基于IDMP协议进行传输，主要数据发送过程如图4‑2所示。idmp\_push\_pending\_frames函数首先读取输出队列中的数据，然后为数据构建IDMP首部，设置IDMP首部中各字段（包括扩展选项），之后读取数据包的目的路由，最后对通过NF\_IDMP\_LOCAL\_OUT检查的数据包交由dst\_output函数。dst\_output函数被调用后，将目的路由表项中的网关地址封装为待output的数据包，然后再交接给idmp\_output函数。如果数据包大于MTU，则调用idmp\_fragent函数进行分片，否则调用idmp\_output2函数将数据包的协议类型设为ETH\_P\_IDMP，最后通过netfilter的NF\_IDMP\_POST\_ROUTING检查后将数据包交由idmp\_output\_finish函数通过邻居子系统将IDMP数据包输出到网络设备。



图4‑2IDGMP数据包的底层发送流程

### 数据包的接收过程

当设备网卡接收到数据包后，检查数据包类型，如果是ETH\_P\_IDMP，则调用idmp\_rcv函数对数据包进行处理。IDGMP数据包的发送过程如图4-3所示。首先检查数据包合法性（协议版本号、数据包长度、校验和计算和目的标识是否为自身的标识等）。如果其中某项不满足，则丢弃该数据包，并返回网络控制报文消息。将合格的数据包交付给函数idmp\_rcv\_finish检查是否存在该数据包的路由表项缓存。如果没有，则调用idmp\_route\_input函数为其查找输入路由表项缓存。如果路由表项缓存查找失败，则丢弃该数据包，并发送网络控制报文消息。如果找到路由缓存，则调用dst\_input函数决定将数据包输入到本地或者转发，前者调用idmp\_input()函数，后者调用idmp\_forward()函数。idmp\_inpu函数调用钩子函数NF\_HOOK对数据包进行本地输入的检查，将合格的数据包交付idmp\_input\_finish函数。idmp\_input\_finish函数负责向上层协议交付数据。



图4‑3 IDGMP数据包的底层接收过程

## 用户层通信接口实现

### MMS接口配置实现

在Unix/Linux[33]的正式的项目上，程序员通常会使用解析命令行函数getopt()或者getopt\_long()来获得输入的参数和选项。getopt()函数只支持短选项的命令行解析，而getopt\_long()不仅支持短选项也支持长选项的命令行解析。本文MMS配置接口的实现，采用getopt\_long()函数解析命令行参数，使用man getopt\_long命令，得到其声明如下：

int getopt\_long(int argc, char \* const argv[],const char \*optstring, const struct option \*longopts,int \*longindex)；

如果执行一个命令时涉及多个选项和参数，通过while循环可以将选项和参数逐个取出，直到返回-1时为止。

MMS配置接口命令常用的包括user、password、confile、database命令，它们本质上就是通过get\_opt\_long()函数解析的，其中long\_options结构如下：

static struct option long\_options[] = {

{"user", required\_argument, 0, 'u'},

{"password", requireed\_argument, 0, 'p'},

{"database", required\_argument, 0, 'D'},

{"version", no\_argument, 0, 0 },

{"confile", required\_argument, 0, 'c'},

{"select", required\_argument, 0, 's'},

};

当通过switch语句判断到某个命令具体的选项时借助optarg变量可以取出参数值，然后调度相应的函数去执行任务即可。比如运行user命令：user –uroot，指定MMS用户名为root。调用函数就会使用MySQL的API接口更改当前数据库的用户名。

### 应用层通信接口实现

接入网络层内终端为组播业务的发起者，同样需要运行组播管理协议IDGMP，为此需要为接入端的用户开发一个配套软件，用以模拟组播业务中常见的请求消息类型。

软件入口main函数首先调用idmp\_sock\_raw\_init函数，创建并初始化套接字接口[34]。为了使用新型网络底层通信协议栈， socket地址簇指定为AF\_IDMP，套接字类型为原始套接字SOCK\_RAW，同时指定协议类型为IDPROTO\_IDGMP，具体做法如下

server\_fd = socket(AF\_IDMP, SOCK\_RAW, IDPROTO\_IDGMP);

执行为idmp\_sock\_raw\_init函数之后，下一句就是封装组播消息。encp\_header函数主要负责组播各类消息的封装处理，函数参数可变，消息封装之后，最终调用sendto()函数。sendto()函数的返回值可以判断所发消息包的大小，单位为字节，通过打印字节数可以检测消息是否被成功发送。

本文为了证明各类消息能被MMS成功得到解析和得到响应，在用户端定义了不同消息发送接口用以验证协议的可行性。下面就不同消息的处理接口进行说明。

1. 组播注册消息发送接口函数register\_group(int flag, char\*msi)

该函数需要传递两个参数，第一参数表明用户角色作为组播源还是组播接收者，第二个参数指定用户所要加入的MSI，返回值为消息认证码

1. 组播加入消息发送接口函数join\_group(int flag, char\*msi, char\*, auth char\* asr\_aid)

该函数需要四个参数，除了最后两个参数外，前两个参数跟第一个接口函数的参数保持一致，其中auth参数由第一个接口函数收到的消息认证码（已缓存到本地）得到，当触发加入消息时会将该参数赋值，第四个参数asr\_aid标识组播成员所加入的MSI所在的ASR的接入标识AID。该函数返回值为true或者false。

1. 组播离开消息发送接口函数leave\_group(int flag, char\*msi, char\*, auth char\* asr\_aid)

该函数与加入消息接口函数的参数含义一致，返回值为true或者false。

1. 组播服务查询接口函数search\_group(char\*msi)

该函数需要传递所要查询MSI，返回值为MSI以及MSI所属的ASR的AID。

1. 组播服务更新接口函数update\_group(int flag, char\*msi, char\* asr\_aid, int timer)

该函数需要传递四个参数，除了前三个参数在前面已经提高外，最后一个参数timer表示组播成员所要申请的服务最新时长，返回值为true或者false。

1. 组播注销消息发送接口函数unregister\_group(int flag, char\*msi, char\* auth)

该函数需要传递两个参数，第一参数表明用户角色作为组播源还是组播接收者，第二个参数指定用户所要注销的组播组MSI，第三个参数为消息认证码，返回值为true或者 false。

发送接口函数定义之后，接下来就是建立结构体缓存区（char\* buffer），将相应消息的报头和消息体填入结构体，封装完毕之后，最后统一调用sendto函数将buffer里的内容发送给目标地址。

## 组播管理服务器模块实现

组播管理服务器模块MMS与接入交互路由器逻辑上分开，不过它们皆位于核心网络层的边缘，实际上二者可以部署到同一机器设备上。MMS相当于一台强悍的组播服务器，所有消息的到达都有它负责完成处理，因为MMS主要负责组播成员的认证、管理和维护工作，所以它又离不开数据库服务的支持。MMS的实现过程主要分为两个阶段：初始化接口阶段、服务建立阶段。

### 初始化接口阶段

包括对命令选项参数的读取和设置，信号处理函数的设置，存储MySQL表的初始化，分别调用用户层通信接口模块和数据库存储模块相关函数实现。

1. 根据用户在终端输入的命令选项和参数，调用getopt\_long()函数配置相关变量；
2. 然后，调用数据库接口，创建数据库mms，生成表table.user和table.subscribe。

### 服务建立阶段

MMS服务建立过程包括原始套接字socket的创建、填充、服务端口绑定与监听以及消息处理线程的不断fork，MMS的服务建立过程如图4‑4所示。



图4‑4MMS的服务建立过程

## 组播消息处理实现

组播管理服务器MMS作为整个协议运作的核心模块，肩负所有组播消息的调度重任。当有消息传至MMS后，调度消息线程进行相应处理，其工作流程如图4‑5所示。

1. 当MMS收到组播请求消息时，立刻fork出一个消息处理线程pThread，由其负责提取消息报头中的type类型和code代码字段，判断消息所属类型；
2. 根据消息具体类型可以检测消息发送者来自组播源还是组播接收者，确定flag；
3. 当消息判定为组播注册类消息，则调用is\_exist()函数，判断用户是否曾经已经注册。如果返回结果为true，则证明该用户已经被注册过，程序直接执行echo\_success\_msg()，回复注册成功消息；否则，执行add\_user()方法，将用户添加到user表中，之后再执行echo\_success\_msg()；
4. 针对非注册类消息，需要对所有消息进行消息认证，判断消息来源的合法性。即检测消息报头中是否携带有消息认证码并且是否与用户管理表中已经事先分配的auth字段匹配，如果认证成功，则继续执行，否则返回echo\_fail\_msg()；
5. 消息认证通过后，若消息类型属于加入消息，MMS就会调度add\_subscribe()方法执行subscribe表操作，记录组播成员所订阅组播组的详细情形；若为离开消息，则调度update\_subscribe()方法更新成员的组播服务状态；若为组播注销消息，则调度update\_user()方法更新组播成员的存在状态，即将user表中的对应用户的state字段修改0。
6. 以上各方法执行之后，则回复相应的响应消息于组播成员，该线程调度结束，然后，等待下一个消息请求的到来，重复以上操作。



图4‑5MMS功能调度流程图

## 数据存储模块实现

本文采用的数据库为MySQL数据库[30][31]，由于运行IDMP协议栈的机制所装操作系统为Fedora，下面主要介绍下MySQL在Fedora系统下的安装和使用方法。

### 数据库安装与配置

首先切换为管理员账户，保障具有root操作权限，初次安装可以参照下面步骤进行。

1. yum安装

#yum -y install mysql && yum -y install mysql-server

1. 启动服务
2. #systemctl start mysqld.service

这里有时候也会不成功，死活都起不来，可以试试重启，或者关机一会。

1. #systemctl enable mysqld.service如果成功的话：会出现 ln -s '/lib/systemd/system/mysqld.service' '/etc/systemd/system/multi-user.target.wants/mysqld.service'这样的提示。
2. 数据库的简单配置
3. 登陆：#mysql –u root（这里不用-p先，因为现在还没有密码 ）
4. 创建MySQL用户：insert into mysql.user(Host,User,Password) values ("localhost","allen",password("111111"));
5. 为用户allen创建数据库mms：mysql>cretae database mms;
6. 授权allen拥有mms数据库的所有权限：mysql> grant all privileges on mms.\* to allen@localhost identified by 'allen';
7. 刷新系统权限表：mysql> flush privileges;

### 数据表设计

数据库安装之后，依据组播业务需求分析，创建组播用户管理表user和组播业务订阅表subscribe。

其中，组播管理用户表的sql创建语句如下（效果如图4‑6所示）：

CREATE TABLE `user`(

`id` bigint(11) unsigned NOT NULL AUTO\_INCREMENT COMMENT 'ID',

`aid` varchar(200) NOT NULL DEFAULT '' COMMENT '接入标识',

`msi` varchar(200) NOT NULL DEFAULT '' COMMENT '组播服务标识',

`key` varchar(100) NOT NULL COMMENT '安全秘钥',

`flag` tinyint(11) NOT NULL DEFAULT '0' COMMENT '标志（0：组播源，1：组播接收者）',

`state` tinyint(11) NOT NULL DEFAULT '1' COMMENT '用户状态（0：已注销，1：生效）',

PRIMARY KEY (`id`)

) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='组播用户注册表';

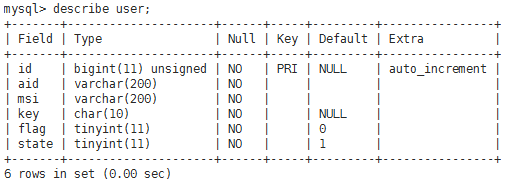


图4‑6组播用户管理表

组播服务订购表的建立语句如下（效果如图4‑7所示）：

CREATE TABLE `subscribe` (

`id` bigint(11) unsigned NOT NULL AUTO\_INCREMENT COMMENT 'ID',

`aid` varchar(200) NOT NULL DEFAULT '' COMMENT '接入标识',

`msi` varchar(200) NOT NULL DEFAULT '' COMMENT '组播服务标识',

`asr\_aid` varchar(200) NOT NULL DEFAULT '' COMMENT '用户所在的ASR的接入标识',

`register\_at` timestamp default CURRENT\_TIMESTAMP COMMENT '组播用户加入时间',

`update\_at` timestamp COMMENT '组播用户更新时间',

`flag` tinyint(11) NOT NULL DEFAULT '0' COMMENT '标志（0：组播源，1：组播接收者）',

PRIMARY KEY (`id`)

) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='组播用户订阅表';

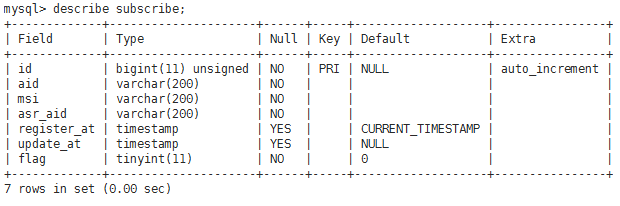


图4‑7组播业务订阅表

### 数据表操作

当MMS服务器收到组播注册消息时，系统会调用add\_user()函数，该函数首先会检测该用户是否已注册或存在记录表项，如果不存在该用户，则执行”INSERT INTO USER VALUES”语句添加新成员记录。用户表user设有auth字段，该字段用于为新注册用户分配安全秘钥。安全秘钥相当于用户密码，该字段会添加到注册响应消息中发给组播用户，以便组播用户在以后的消息请求中始终携带着该秘钥，保证身份的合法性。如果收到组播服务查询请求，则调用search\_subscribe()查询当前组播成员的服务状态信息，主要执行“SELECT \* FROM subscribe WHERE aid=($aid)”语句，并根据当前用户的标识条件实现精确的查询效果。如果收到组播离开请求，则调用update\_subscribe()函数，该函数主要执行“UPDATE subscribe SET update\_at= CURRENT\_TIMESTAMP WHERE aid=($aid)”语句完成用户状态修改。如果组播发起注销请求，则执行update\_user()函数，该函数主要执行“UPDATE user SET status=0 WHERE aid=($aid)”语句，用以修改用户的存在状态。

## 协议相关结构体定义

基于新型网络架构的组播成员管理协议生成的组播通信消息封装在新型网络底层通信协议IDMP之上传输，组播通信消息具体格式如图4-8所示。

图4‑8消息公共报头格式

本文针对不同种类的消息定义了不同的结构体变量，下面就不同消息的结构体定义介绍如下。

组播服务发布消息：

struct idgmp\_broadcast\_msg

{

\_\_u8 type;

\_\_u8 code;

\_\_u16 checksum;

\_\_u16 len;

\_\_u16 reserve;

struct in\_id start;

struct in\_id end;

};

组播成员注册消息：

struct idgmp\_register\_msg

{

\_\_u8 type; //消息类型

\_\_u8 code;//消息代码

\_\_u16 checksum;//消息校验和

\_\_u16 len;//消息体长度

\_\_u16 reserve;//保留位

struct in\_id aid;

struct in\_id msi;//注册组播组，组播服务标识

};

组播成员注册成功响应消息：

struct idgmp\_register\_reply

{

\_\_u8 type;

\_\_u8 code;

\_\_u16 checksum;

\_\_u16 len;

\_\_u16 reserve;

\_\_u8 key[8];

};

组播成员注销消息：

struct idgmp\_unregister\_msg

{

\_\_u8 type;

\_\_u8 code;

\_\_u16 checksum;

\_\_u16 len;

\_\_u16 reserve;

struct in\_id aid;

struct in\_id msi;

\_\_u8 auth[8];

};

组播成员加入或离开消息：

struct idgmp\_join\_msg

{

\_\_u8 type;

\_\_u8 code;

\_\_u16 checksum;

\_\_u16 len;

\_\_u16 reserve;

\_\_u64 auth;

struct in\_id aid;

struct in\_id msi;

struct in\_id asr\_aid;

\_\_u8 auth[8];

};

组播服务查询消息：

struct idgmp\_search\_msg

{

\_\_u8 type;

\_\_u8 code;

\_\_u16 checksum;

\_\_u16 len;

\_\_u16 reserve;

struct in\_id msi;

};

其中id结构体的定义如下：

/\* IDGMP ID \*/

struct in\_id {

union {

uint8\_t in\_id8[16];

uint16\_t in\_id16[8];

uint32\_t in\_id32[4];

} id\_u;

#define s\_id8 id\_u.in\_id8

#define s\_id16 id\_u.in\_id16

#define s\_id32 id\_u.in\_id32

};

将以上结构体定义放入header.h文件中。当使用某个结构体时，只需要在c文件中引入头文件并作相应结构体变量声明即可。例如封装一个组播源的注册消息，可以按照以下操作进行：

struct idgmp\_register \*msg=NULL;

//映射信息

char \*aid = "13333-2-2-1~8";

char \*rid = "10000-50000-10~1";

char \*msi = "10000-60000-10~1";

msg=(struct idgmp\_register\*)malloc(sizeof(struct idgmp\_register));

memset(msg, 0, sizeof(struct idgmp\_register));

msg -> type = 2;

msg -> code = 1;

msg -> len = 16;

msg ->msi = msi;

在处理多字节变量时，由于不同机器对字节处理顺序会有不同的做法，为了保证目的设备能够正常解析，需要将主机字节序转化成网络字节序。

常用的c函数有htons()、ntohs()，htonl()、ntohl()。其中htons()针对short类型变量，htonl针对long类型变量，如果在c文件中使用它们还需要引入头文件<arpa/inet.h>。有时候还会需要处理64位的long long类型字段数据，此时可以单独封装htonll()、ntohll()函数，具体做法如下：

htonll(uint64\_t val) {

return (((uint64\_t) htonl(val)) << 32 ) + htonl(val >> 32 );

}

uint64\_t ntohll(uint64\_t val) {

return (((uint64\_t) ntohl(val)) << 32 ) + ntohl(val >> 32 );

}

针对不同消息宜选用不同结构体进行封装，消息封装完毕之后，就可以使用sendto()或recvfrom()函数执行相应的操作了。

## 本章小结

本章节基于新型网络架构对组播成员管理系统的实现过程给予了详细说明与介绍，涉及内容包括协议内核的运行机制、数据包的收发过程、用户层通信接口、组播管理服务器模块实现和数据库设计等。在协议内核模块的实现机制中介绍了IDMP数据包的工作流程和套接字接口类型，描述了IDGMP数据包的发送和接收原理以及实现过程。在用户层接口实现小节中介绍了用户层配置接口和通信接口的实现机理，为下一步配合组播管理协议的验证工作提供了配套软件方案。本章节特别介绍了组播管理服务器模块的实现，将其分为三个阶段，并对每个阶段实现过程、工作流程、工作原理给出了详细介绍。紧接着详细设计了有关组播管理方面的数据表操作，讲解了协议中使用的基本表的设计与管理。最后，详细定义了IDGMP有关消息的数据结构体。

# 系统搭建及验证测试

本章将对基于新型网络的组播成员管理协议的设计与实现进行原型系统的搭建和实际验证测试。系统搭建主要介绍了系统平台环境与网络拓扑结构，测试工作主要主要包括组成员安全认证测试和协议功能性测试。组成员安全认证测试的目的是为了验证组播成员消息的合法性；协议的功能性测试的目的一方面是为了验证协议消息是否符合协议规范，另一方面主要是验证组播管理服务器运行逻辑和数据库操作结果是否符合预期，是否满足实际场景需求。验证过程中，采用Wireshark抓包工具进行辅助测试。

## 系统环境与配置说明

系统测试环境所采用的网络拓扑结构如图5‑1所示。测试环境包含两台测试终端一台CSR，两台ASR，一台MMS。



图5‑1测试网络拓扑

其中，Source终端充当组播源的角色，Receiver终端充当组播接收者的角色；ASR3和ASR1分别作为两个终端所属的接入交换路由器。MMS负责集中管理和维护所有的组播成员的服务状态。CSR主要负责核心网络层数据的转发和路由操作。整个网络拓扑符合新型网络架构要求。

测试过程中，Source作为组播源向组播管理服务器发送组播注册请求消息，途径接入交换路由器ASR3和核心交换路由器CSR3到达组播管理服务器MMS。由于请求消息依赖新型网络底层通信协议IDMP，需要单独配置单播路由，才可保证数据的完整性到达。同理，Receiver作为组播接收者同样需要配置单播路由表项。当MMS收到对端请求消息后执行相应操作，记录组播状态，更新数据库，为后期组播路由树的建立提供基本的数据保障。

测试环境中，各主机硬件环境及软件配置详见表5‑1。

表5‑1测试设备及仪表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **设备名称** | **硬件** | **软件** | **数量** |
| 1 | 接入网终端 | CPU：Inter(R) Core(TM) i3 CPU T 1350 @ 1.86GHz；  RAM：2G；  硬盘：250G；  网卡：一块 | 操作系统：Fedora Linux version 2.6.28 (gcc version 4.3.2) | 2 |
| 2 | 核心交换路由器 | CPU：Inter(R) Core(TM) i5-3470 CPU @ 3.20GHz；RAM：4G；  网卡：三块； | 操作系统：Fedora Linux version 2.6.28 (gcc version 4.3.2) | 1 |
| 3 | 接入交换路由器 | CPU：Inter(R) Core(TM) i3 CPU T 1350 @ 1.86GHz；  RAM：2G；  网卡：两块，eth0连接核心网络层，eth1连接接入网络层； | 操作系统：Fedora Linux version 2.6.28 (gcc version 4.3.2) | 2 |
| 4 | 组播管理服务器 | CPU：Inter(R) Core(TM) i5-3470 CPU @ 3.20GHz；  RAM： 8GB；  网卡：一块；  硬盘：1TB； | 操作系统：Fedora Linux version 2.6.28 (gcc version 4.3.2)  数据库：MySQL(version 5.1) | 1 |

各个网络设备的接口配置见表5‑2。

表5‑2网络设备接口配置表

|  |  |  |
| --- | --- | --- |
| 设备名称 | 接口 | 分配地址 |
| CSR1 | eth0 | 10000-40000-10~1 |
| CSR1 | eth1 | 13333-2-2-1~1 |
| CSR1 | eth2 | 10000-40000-30~1 |
| ASR1 | eth0 | 10000-40000-10~3 |
| ASR1 | eth1 | 13333-2-2-1-10~1 |
| ASR2 | eth0 | 10000-40000-30~2 |
| ASR2 | eth1 | 13333-2-2-1~2 |
| Source | eth0 | 13333-2-2-1~8 |
| Receiver | eth0 | 13333-2-2-1~4 |
| MMS | eth0 | 13333-2-2-1~6 |

## 测试工具介绍与开发

### 工具介绍

本文在对协议测试中采用Wireshark作为测试工具，并对其进行插件扩展以解析新提出的协议。Wireshark是一款著名的网络嗅探抓包工具，通过抓包可以撷取从机器网卡设备上传输通过的网络消息包，详细展示网络各层协议字段数据等内容。Wireshark默认支持大量的网络协议，例如 Wireshark 1.6.2版本已支持1170种协议，包括Diameter、GTP、FTP、SCTP等协议。

Wireshark协议解析流程如图5‑2所示，首先Wireshark侦听网络接口并捕获数据包，随后提取数据包内容，根据协议规范对数据包内容进行移位操作提取各个地段的数据，最后打印出各个字段的数据。

虽然Wireshark具有强大的协议解析功能，但上述都是互联网通用协议，然而对于自主研发的新型通信协议，Wireshark无法解析，导致开发调试或者测试分析效率较低。目前让Wireshark也能解析自定义协议的方法主要有两种：修改Wireshark源码，或基于Lua语言开发Wireshark协议解析插件。本文采用后者，即采用Lua语言[35]开发协议解析插件。

安装Wireshark之后，默认会在安装路径下生成一个初始脚本文件init.Lua，该文件是Wireshark启动时第一个加载的Lua脚本。开发者可以手动配置该文件，激活Lua语言解释器、调用dofile函数加载自定义Lua插件文件，从而实现自主协议的解析操作。



图5‑2Wireshark协议解析流程

Wireshark按照与配置如下所示：

a) 源码安装lua：

b) 安装wireshark：(安装成功后的wireshark应用程序在/usr/local/bin/下)

c) 添加idmp.lua文件到wireshark:

安装过程中可能出现的问题及解决方法：

a) 在安装lua过程中可能会出现：readline/readline.h: no such file or directory 或readline/history.h: no such file or directory

解决：yum install readline-devel

b) 在安装wireshark时可能会出现：pcap.h：no such file or directory

解决：yum install libcap-devel

c) 在安装wireshark时可能会出现：gtk+-version >=2.4.0 的要求

解决：yum install gtk+extra-devel.i386

d) 在linux的root下启动wireshark时，可能会出现权限不允许是情况

解决：修改/etc/selinux/下的config文件，把SELINUX=enforcing修改为SELINUX=disabled保存。重启linux操作系统。

### wireshark插件开发

基于Lua语言开发的Wireshark协议解析插件，首先通过标识类型指示位F确定数据包是否来自接入网络层。IDMP协议规定，标识类型指示位F为“0”是接入网络层的数据包，采用接入网络层数据包格式进行解析；为“1”是核心网络层的数据包，采用核心网络层数据包格式进行解析。以上完成基本首部解析，然后通过下一首部确定扩展选项，进行相应扩展选项的解析。若扩展选项为IDGMP协议，则按照IDGMP协议规范进行下一步解析。若选项后应再无数据需进行解析，解析过程直接结束。IDGMP协议解析过程如图5‑3所示。



图5‑3 IDGMP协议解析流程

## 基于新型网络的组播成员管理协议测试

基于新型网络的组播成员管理协议的测试主要包含组播安全认证测试和消息功能性测试两大块内容，其中消息功能性测试又主要包括组播注册与注销测试、组播服务发现测试、组播加入与离开测试以及组播更新测试。组播安全认证测试是组播注册时由组播管理服务器分配安全秘钥后供其他几类消息合法性认证的前提性测试，虽然不属于本文消息类型定义规范，但为了满足其他类型消息的功能性测试目的，有必要单独作为一项测试内容。

### 安全认证测试

除了组播注册消息之外的大部分类型消息都需要进行安全认证测试。基于新型网络的组播成员管理协议的消息安全认证的设计思想是利用组播成员发送组播注册消息时由组播管理服务器分配的安全随机数作为秘钥发送给终端用户，由终端用户缓存到本地，之后，终端用户在发送组播类型消息时会使用该秘钥对所发消息进行加密处理，加密后的数据将作为消息认证码附加到消息尾端，最后和真正的消息内容一并发送到组播服务管理器。组播服务管理器收到该消息后，首先从组播成员管理列表中取出事先分配好的安全秘钥相同，其次，利用相同的算法对除去消息认证码的部分消息载荷加密计算，将计算结果和接收提取到的消息认证码进行比对，如果二者一致，说明该消息数据合法，执行接下来的任务操作，否则，回复非法请求。该思想实际上属于共享秘钥机制，将安全秘钥key和消息载荷data作为哈希函数H(data|key)的输入最终计算而得到消息认证码，具体算法还可随数据安全等级而定。

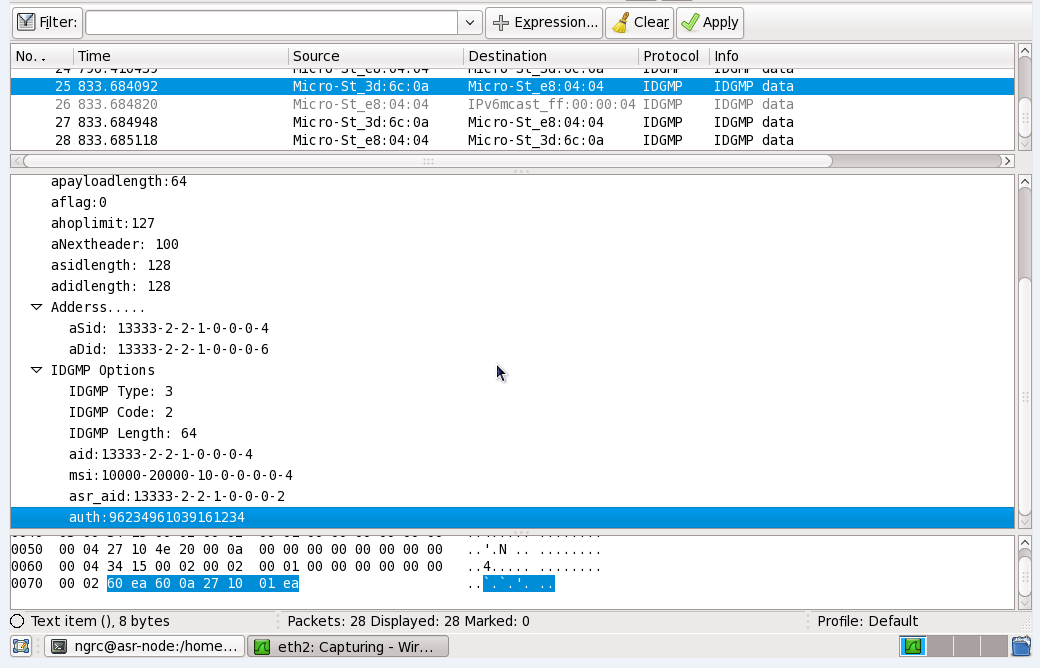


图5‑4消息安全认证分析

本文设定的消息认证码为64位，由MMS分配的auth作为秘钥k，发送消息时对所发消息载荷data通过加密后得到m，data和m作为最终消息一并送到组播管理服务器MMS完成认证。如图5‑4所示展示了一个利用Wireshark工具撷取分析所看到的携有消息认证码的组播消息效果图。

### 协议功能测试

基于新型网络的组播成员管理协议的功能性测试主要目的之一是验证IDGMP协议的消息的完整性、一致性，另外一个重要目的就是验证MMS在收到组播成员消息后的逻辑功能处理情况，检测能否按照协议的要求实现组播成员的管理与维护工作。本文将从以下四类消息进行功能性测试：(a) 组播服务发现测试；(b) 组播注册与注销测试；(c) 组播加入与离开测试；(d) 组播更新与维护测试。

测试采用如图5‑1所示的拓扑图。其中，设备Source作为组播源，设备Receiver作为组播接收者，设备MMS作为组播管理服务器。

1. 组播服务发现测试

组播服务发现测试的内容主要包括组播服务发布消息测试和组播查询消息测试。组播服务发现的测试目的主要是为了验证组播服务是否能被准确发布并被潜在的组播成员后期查询和订阅。

由本文第三章节组播消息分类可知，组播发布消息的类型为type=0x01，消息代码为code=0x01，图5-1中的Source向MMS发布该类型消息后，通过wireshark采撷数据包解析消息内容如图5‑5所示。

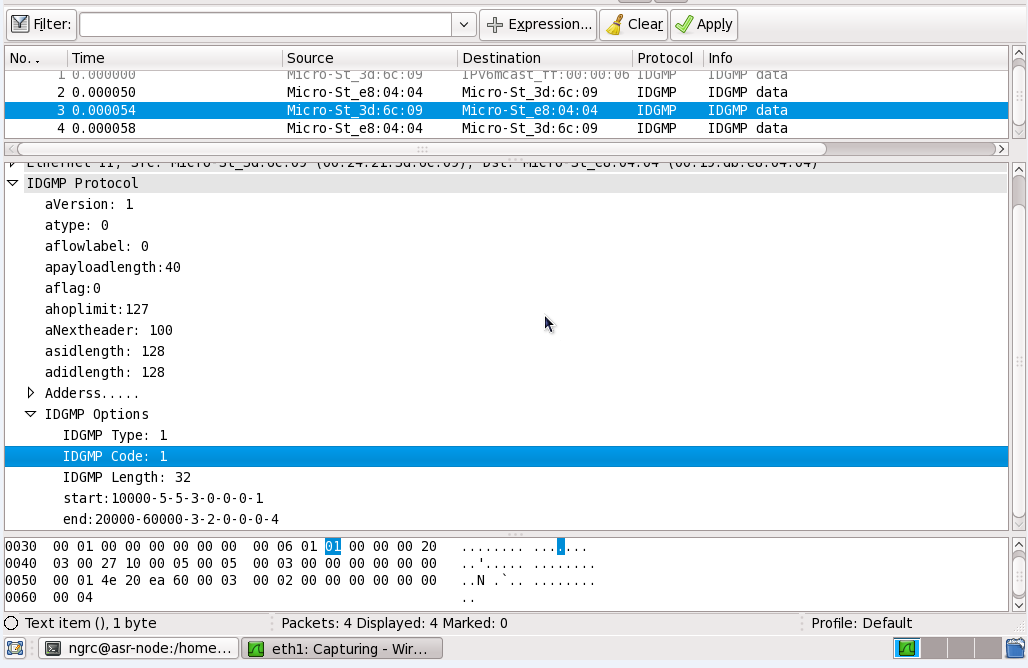


图5‑5组播发布消息解析

从抓包图中可以观察到组播服务范围的起始地址和终止地址分别为10000-5-5-3~1和20000-60000-3-2~4。

当MMS建立组播服务范围后，组播源向MMS发送组播查询消息，以确定某组播服务是否存在。通过wireshark抓包后的消息解析如图5‑6所示。

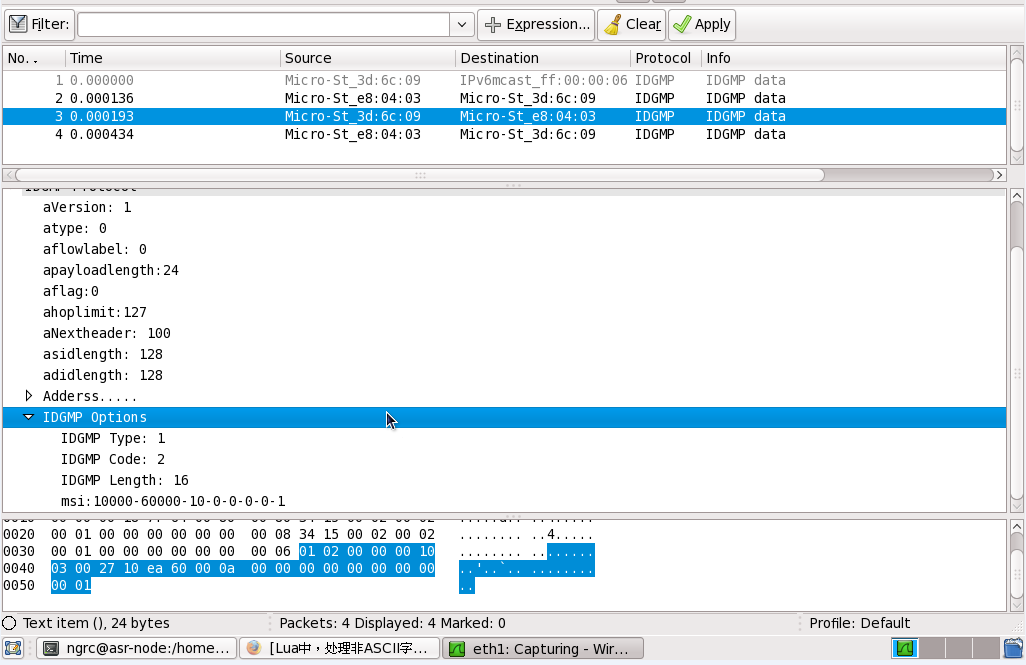


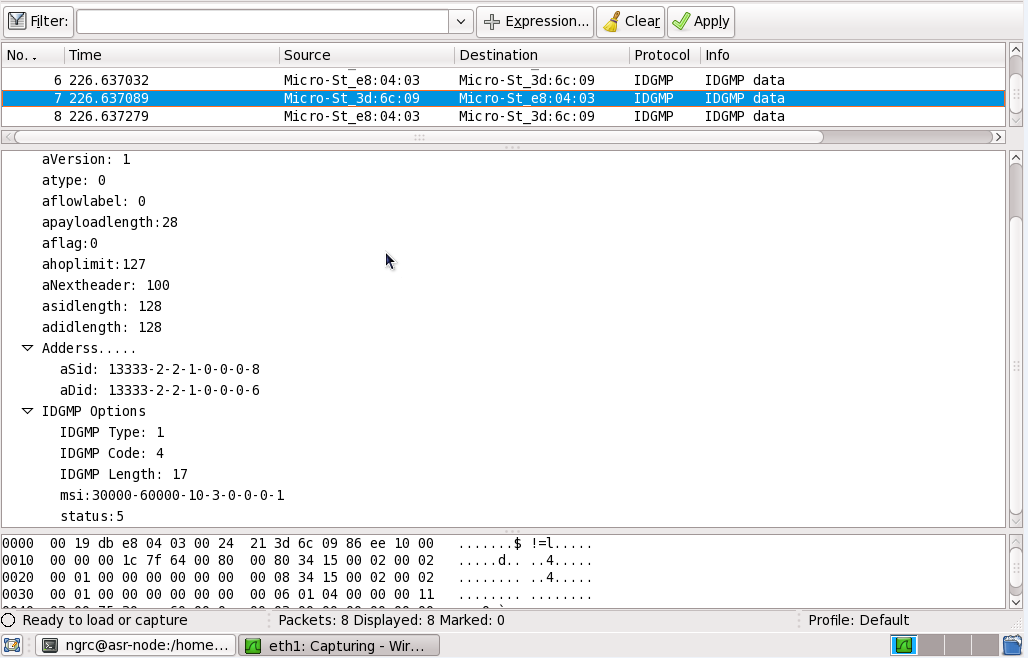
图5‑6组播查询消息解析

图5‑7组播查询失败消息解析

从上图中可以看出，组播源发送了一个消息类型type=1，code=2的组播查询消息，所要查询的组播服务标识为：30000-60000-10-3~1。如果组播服务存在，组播源将会收到确认响应报文，否则会收到一个携带有组播错误码的组播查询失败响应消息。当组播管理服务器MMS收到组播查询消息后，发现该用户所要查询的组播服务标识为30000-60000-10-3~1，已经超出了自身服务管辖范围，于是乎，MMS向此用户回复了消息类型type=1，code=4的消息，经查表3‑1发现，该消息为组播查询失败消息，并携带有status字段，值为5，经查表3‑5可知，此失败情形属于组播服务不存在。具体内容经wireshark解析如图5‑7所示。

1. 组播注册与注销测试

组播注册消息的功能性测试是为了验证组播注册成功之后是否分配安全秘钥，MMS是否在数据库表user中新增成员记录等。组播注销消息的测试在于验证MMS是否更新组播成员的申请状态。

本文仅就组播源的注册消息请求作以测试案例进行分析，组播接收者的注册请求在此不做介绍。首先使Source向MMS发送一个type=0x02，code=0x01的注册报文（见图5‑8所示），若注册成功，Source会收到成功注册报文（见图5‑9所示）以及MMS新增组播成员记录。具体测试通过wireshark解析结果如图5‑8所示。

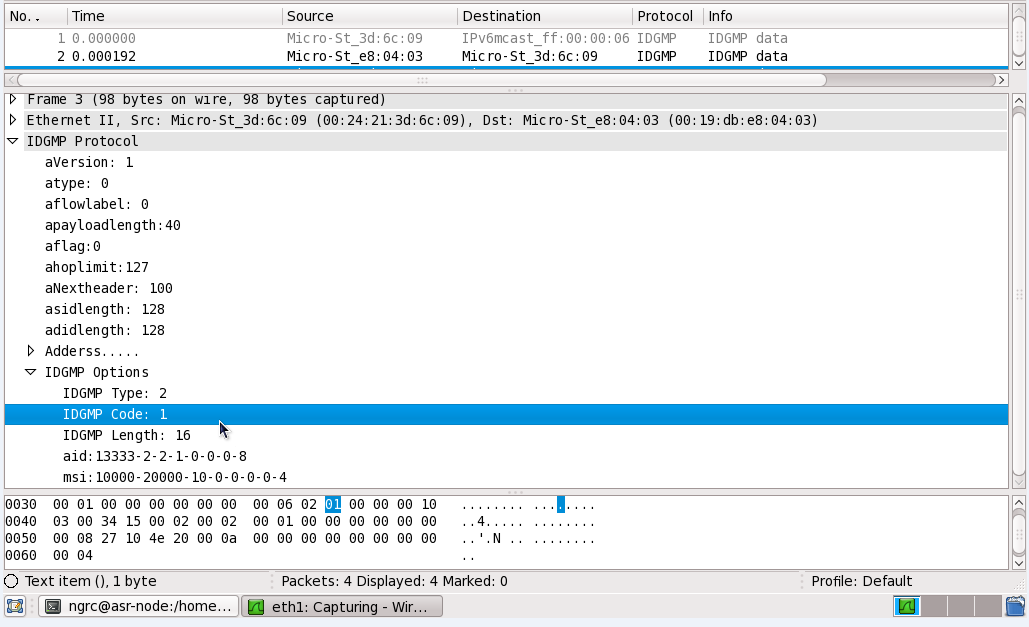


图5‑8组播源注册消息解析

当MMS收到来自组播源（接入标识为：13333-2-2-1~8）的注册报文时，若查询其已经注册，则直接回复注册成功响应报文。否则，若判定为初次注册，则MMS就会调用安全随机数函数生成8个字节长度的秘钥，然后，按照协议规范，生成针对该组播源的成功注册响应消息，返回给组播源（从图5‑9中可见返回消息中带有key字段，将由该用户保持到本地，以便将来生成非注册类消息使用），同时，将该成员的基本注册信息以及刚刚分配的秘钥保持到组播用户管理表中。具体结果通过wireshark抓包解析如图5‑9所示。

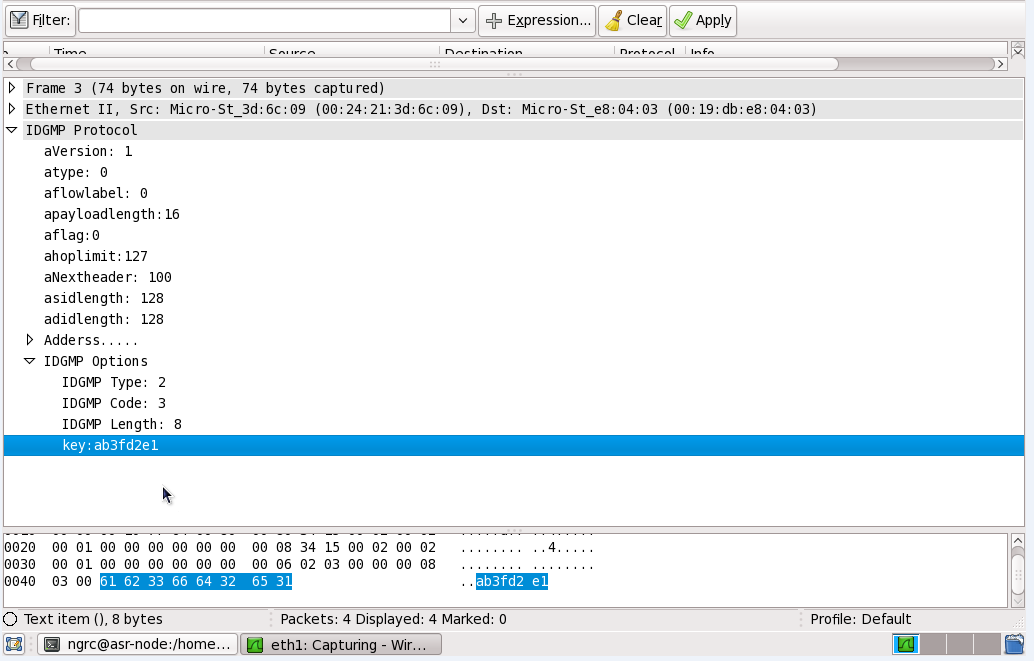


图5‑9组播源注册成功响应消息解析

当组播源第一次注册某组播组，MMS初次收到该组播源的注册报文时，将其身份标识、注册申请的组播服务标识等信息记录到本地组播成员管理表中（从图5‑8中可以发现组播源接入标识为：13333-2-2-1~8，注册申请的组播服务标识为：10000-20000-10~4），同时将MMS为其分配的秘钥存储到key字段，将该用户的角色标记为0即（flag=0），用户激活状态为1，经通过MySQL数据库查询可得如图5-10结果。

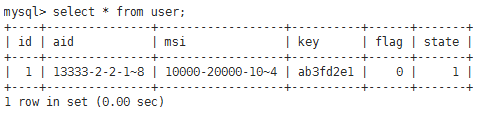


图 5‑10组播用户管理表查询结果

从图 5‑10中显示结果可以分析当前新增用户角色为组播源（因为flag=0），组播成员接入标识为：13333-2-2-1~8，所注册的组播组为：10000-20000-10~4。从state状态可以看出当前用户为激活状态。

当组播源Source向组播管理服务器MMS发送一个type=0x02，code=0x06的组播源注销报文（经wireshark抓包解析如图5‑8所示），若注销成功，Source会收到成功注销报文，同时MMS更组播成员的用户状态。若注销失败，Source会收到了一个含有组播错误码的消息报文。从解析结果可知，组播源的接入标识为13333-2-2-1~8，欲注销的组播服务标识地址为10000-20000-10~4，同时携带有长度为8字节的消息认证码。

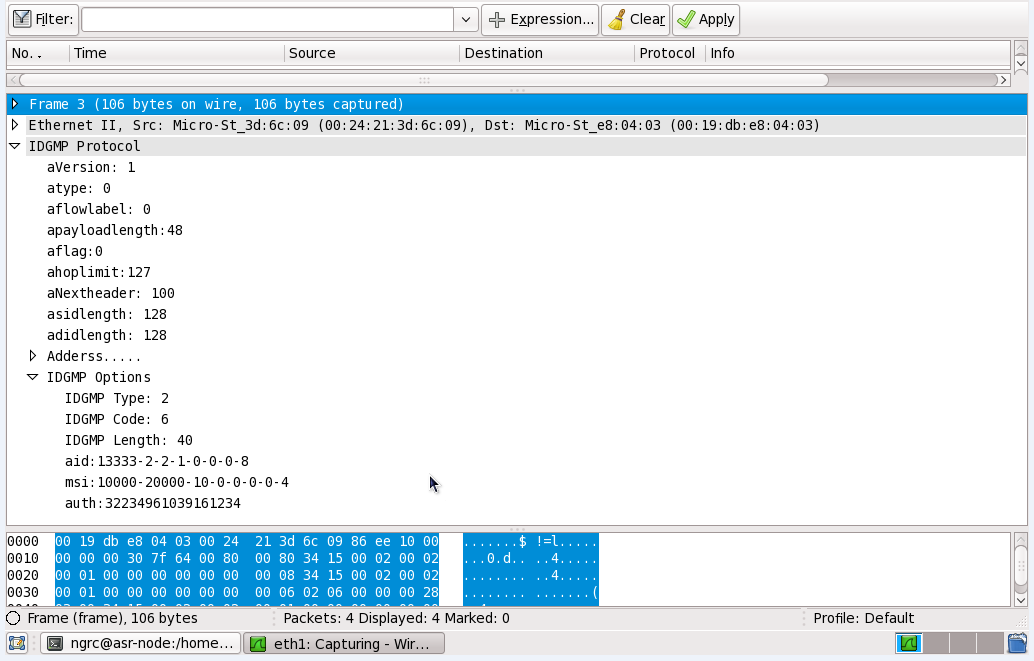


图5‑11组播源注销消息解析

接入标识为13333-2-2-1~8的组播源发送组播源注销消息后，MMS首先对该消息认证合法性认证，经判定该消息合法且为组播源注销消息类型，接着，MMS查询组播用户管理表，检测是否存在相对应的数据记录，delete\_user()函数返回true，表明确实存在该用户，于是乎，执行该用户的状态更新操作，将state字段值设为0。数据库结果如图 5‑12所示。由state=0可知当前组播成员已经成功解注。

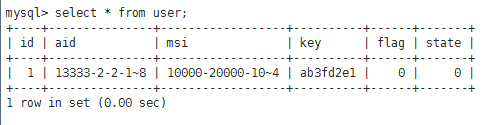


图 5‑12组播成员管理更新查询

1. 组播加入与离开测试

组播成员加入与离开的功能性测试是为了验证程序根据协议定义在消息首部字段的数据包的不同情形下的处理。当组播成员未授权申请加入组播组时，组播管理服务器返回加入失败消息；当组播成员成功加入组播组时，组播管理服务器返回加入成功消息。当组播成员发起组离开请求时，组播管理中心更新相应组播状态，同时MMS更新组播业务订阅表记录。

此次测试只针对组播源，组播接收者同理。首先Recevier向MMS发送一个不携带消息认证码的组播加入报文（如图5‑13所示）。

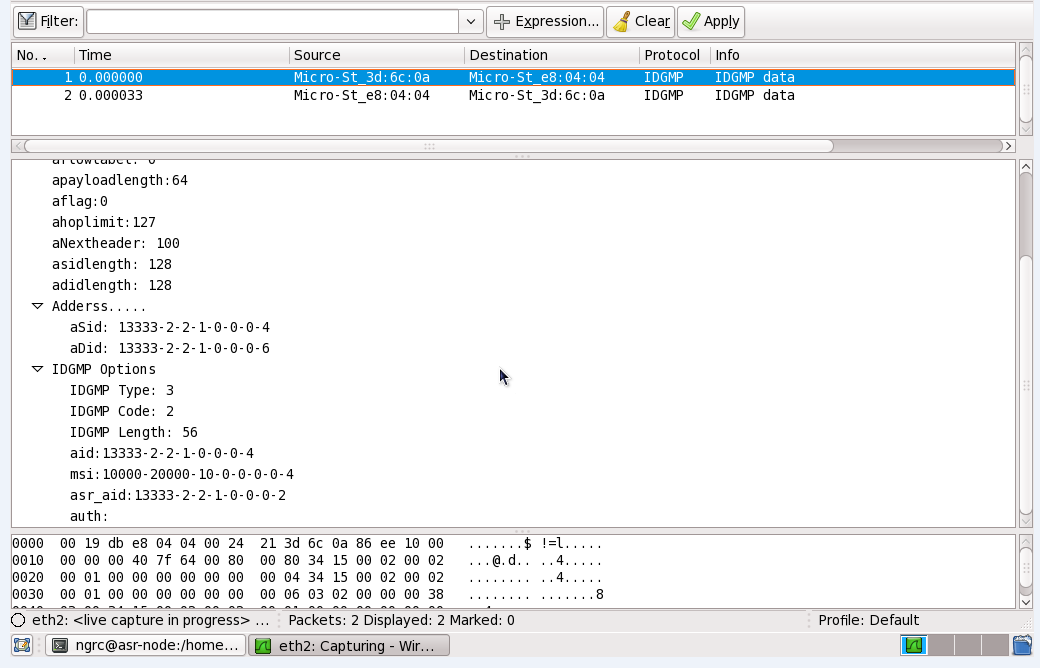


图5‑13组播加入消息解析（未含消息认证码）

MMS收到该报文后首先判断组播加入消息的合法性，经验证合法后，继续执行消息处理函数，解析消息字段，判断该成员所发送的消息类型为组播加入消息类型，紧接着解析其申请的组播服务标识，经再次判断msi字段值不在本地服务所管辖的服务范围内。根据协议规范，MMS应回复加入失败响应消息。在接入网络层内的source端，收到该回复消息后，通过wireshark抓包解析得到如图5‑14所示结果，从解析结果可以看出status字段值为1，按照表3‑5组播消息状态码 中的规定可知，属于服务不存在错误，符合预期。

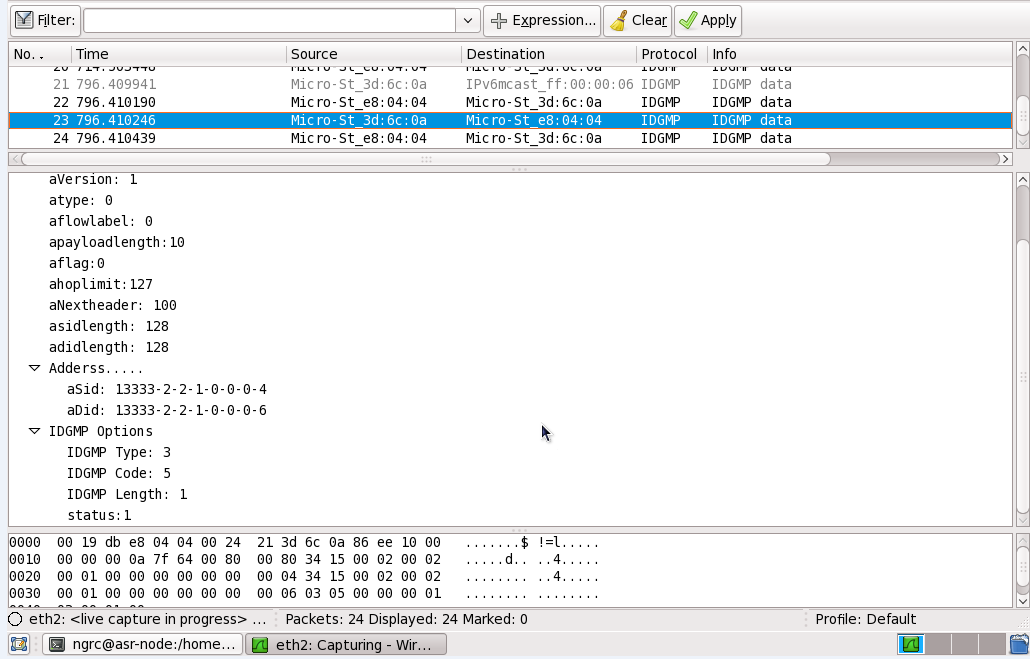


图5‑14组播成员加入失败响应消息解析

同理，当组播接收者向MMS发送一个携带有消息认证码的type=0x03，code=0x02组播加入报文（如图5‑15所示）。MMS收到该报文后，同样进行消息的安全认证，若判断合法后，接下来调用加入消息处理函数，执行有关成员加入的逻辑功能部分的操作。从所发消息中可以看出组播接收者Recevier（接入标识为：13333-2-2-1~4）预备加入的组播服务标识为10000-20000-10~4，组播接收者所在的接入交换路由ASR的标识地址为13333-2-2-1~2，MMS在此消息得到正确处理后，在本地组播业务订阅表中新增组播的订阅记录，同时给Reveiver回复一个成功加入报文。MMS初次收到组播接收者加入消息报文后的本地wireshark抓包解析结果如图5‑15所示。

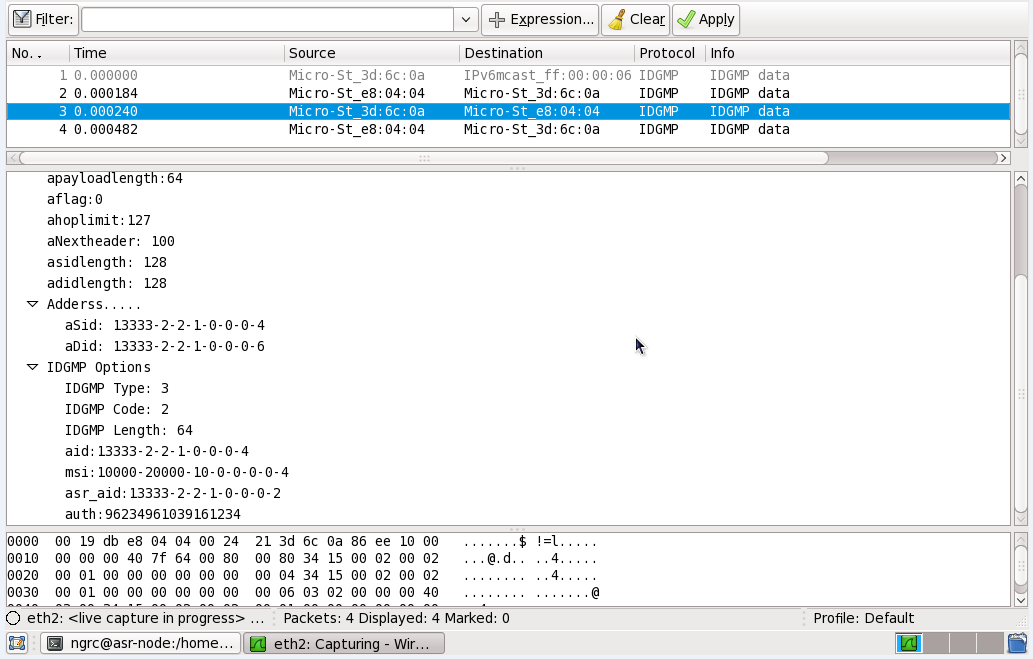


图5‑15组播源加入消息解析（含有消息认证码）

当组播接收者（接入标识为：13333-2-2-1~4）成功加入组播组时，MMS调用add\_subscribe()函数在组播业务订阅表中插件新的组播状态记录，如图5‑16所示。

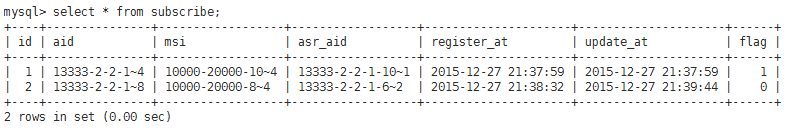


图5‑16组播业务订阅查询结果

当组播接收者发送（type=0x03,code=0x08）组播离开请求报文时，wireshark抓包解析如图5‑17所示。从解析结果可知，接入标识为13333-2-2-1~4的组播接收者欲申请离开的组播组地址为10000-20000-10~4，该接入标识所在ASR的接入标识为13333-2-2-1~2，携带有auth消息认证码，MMS经判定该消息合法后，调用离开消息处理函数，执行组播业务订阅表中的相关成员的记录更新或删除操作。

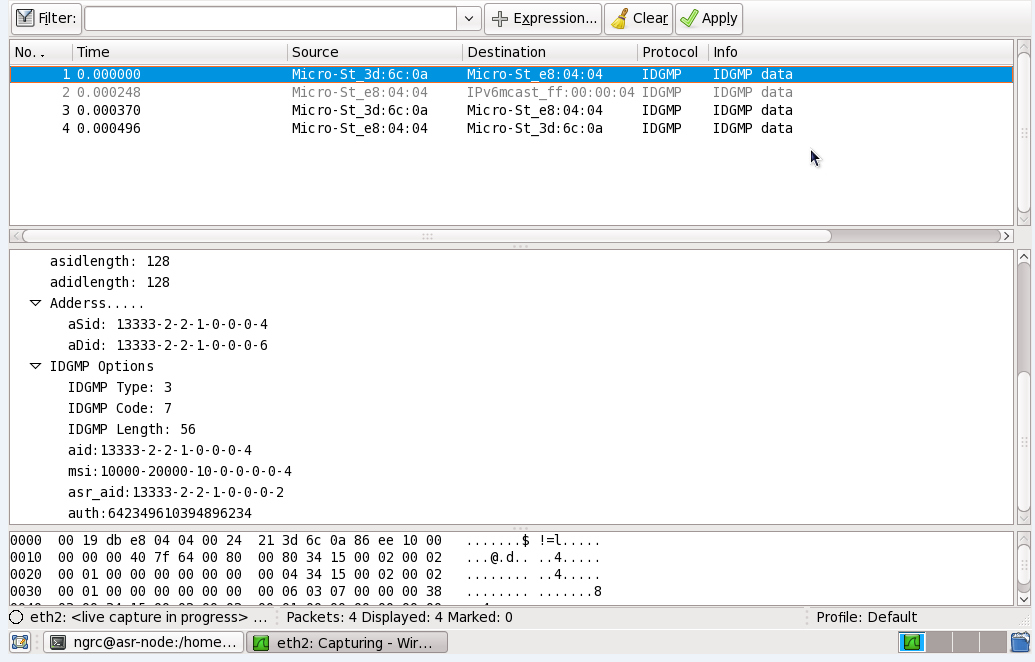


图5‑17组播接收者离开请求报文解析

MMS收到该请求后，首先验证该消息的合法性，认证通过后，MMS则调度update\_subscribe()函数执行组播成员的服务状态删除操作，查询结果如图5‑18所示。从表中可以看到第一条记录已经被删除。

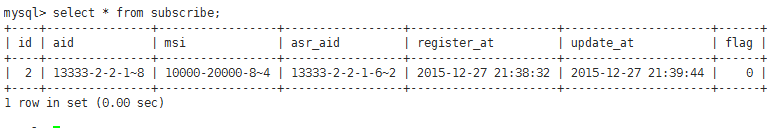


图5‑18组播服务状态删除结果查询

1. 组播更新维护测试

组播更新维护的测试主要验证组播成员服务时间时组播管理服务器的操作正确性。

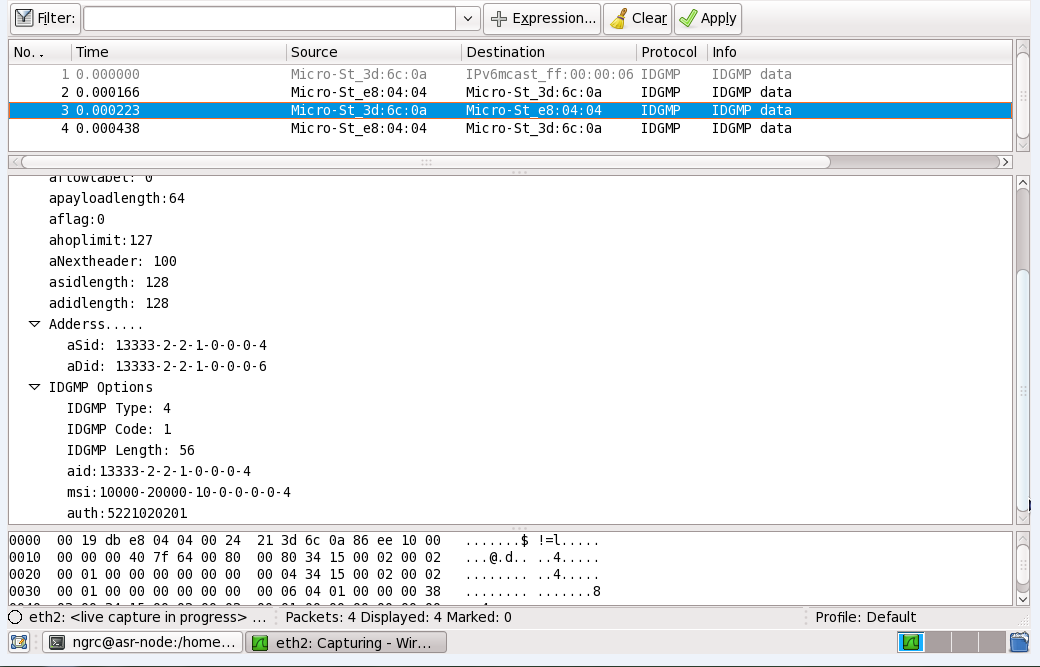


图5‑19组播服务时间更新消息解析

测试时首先使已经过组播管理服务器认证的Receiver设备向MMS发送服务时间更新消息报文，MMS收到该请求消息后经验证合法，回复更新确认消息，同时更新subscribe表中的组播状态记录。测试经wireshark抓包解析如图5-19所示。

MMS回复服务时间成功消息后，紧接着更新相应成员的组播服务时间，即更新update\_at字段值，MMS默认时间更新机制为当前服务器时间的基础上增加100分钟。经查询可见如下结果。

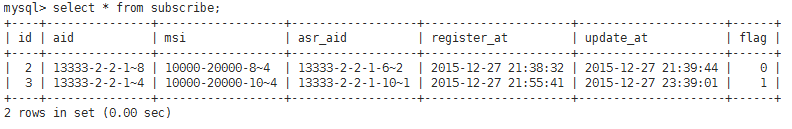


图5‑20组播成员服务时间更新查询结果

## 本章小结

本章节主要介绍了组播成员管理系统验证平台搭建以及所需要的各项环境配置，介绍了自主协议开发测试工具wireshark，讲解了wireshark插件的解析工作流程。系统环境配置之后，对本文所研究的新型协议进行了整体测试。测试内容主要从消息安全认证和功能性测试两块入手。其中功能性测试包含了组播加入与离开、组播注册与注销、组播服务发现、组播服务更新等各类消息的测试，测试内容涵盖了协议所牵涉的各个功能模块，主要属于组播管理服务器对本地数据库的管理和维护等操作。测试过程顺利，基本满足协议规范和要求，充分验证了本文所设计的组播成员管理协议在新型网络架构下的工作稳定性，下一步工作，除了继续完善MMS系统软件之外，还需要在大规模网络环境下运行该协议，保证协议能够工作良好，保障协议的工作性能。

# 结束语

## 论文总结

互联网技术正在日益渗透到人们日常工作、学习和生活的各个方面，逐渐成为传播和获取信息的重要工具之一。借助互联网平台，人类不仅可以获取基本的语音、视频、图像、文件等流量服务，同时还可以享受到众多便捷实惠的衍生服务。比如在线商品交易、语音助手、群组聊天、智能家居、云盘共享等，互联网的便利性和丰富应用极大增强了用户对网络的依赖性。互联网正在发生着深刻的改变，主要表现在其使用模式逐渐由传统基于端的通信模型向以内容的分发与获取转变，其使用主体由传统的固定设备如大型工作站、个人电脑向移动设备如智能终端等迁移，由传统的以人工操作为主的网络设备向以物为中心的传感节点等进行扩展，特别是近几年由于智能手机的普及，移动互联网正以惊人的步伐向前奔跑。面对这些变化，传统的互联网设计理念早已无法满足网络发展的新需求。由于传统互联网在安全设计上的不足，导致各种网络攻击行为和隐私泄漏问题频发，严重影响了网络的正常使用，特别是当互联网应用于特殊领域时，比如在线实施交易，这些弊端将带来严重的潜在威胁。以互联网为平台，各种信息监控、捕获和分析技术也在严重威胁着网络的安全。

传统互联网在设计之初，仅面向抗毁性和连通性的有限基本目标，忽略了对整个网络的安全性、移动性、可控可管性和服务质量支持等问题。随着网络规模的不断扩张和终端实体入网的复杂性、多样性等特征，基于TCP/IP网络架构的网络缺陷将会日益凸显，严重阻碍着信息网络的健步发展，亟待需要研制全新的互联网络架构。近些年，世界各国纷纷开始重新审视传统的TCP/IP体系架构，开展未来网络架构设计，试图解决诸如移动性差、扩展性差、安全性差等问题，以改善网络整体的生态环境。

在此背景下，本文基于我国自主研发的新型网络架构，针对组播管理这一关键技术进行了系统研究，系统通过吸收新型网络架构中标识分离映射的核心思想，研究一种新型的组播成员管理系统，力求从协议创新的角度，在期许丰富新型网络通信协议、填补空缺的同时，寄希望于弥补现有组播通信协议栈的缺陷，在推动新型网络的研究道路上添砖加瓦。

本文主要介绍了组播成员管理协议的工作机制和设计方案，并对系统中各个功能模块给出了详细实现和介绍。重点介绍了MMS的工作机理。MMS是整个协议能够工作的核心，它摒弃了传统路由器分布式管理成员的思想，而是采用集中式的管理方式实现了组播成员的认证授权、加入、离开、注册、注销、服务更新等功能。

IDGMP协议使用C语言开发，实现了MMS的基本框架设计，主要包括协议数据包的收发过程，用户层接口定义，MMS的服务建立阶段和消息处理调度阶段，数据存储模块的设计与实现以及协议相关结构体的定义。之后，按照新型网络的标准，搭建了网络拓扑环境，对本文协议进行最终测试，测试过程主要包括安全认证测试和协议功能性测试。具体包括组播服务发现、注册与注销、加入与离开、服务时间更新等情形。测试结果通过Wireshark抓包解析，证明该协议基本满足对组播成员的管理和控制。

## 下一步研究工作

在广泛阅读有关组播技术方面的最新研究成果和相关资料，以及组播成员管理方面的相关标准后，立足于当前组播管理技术发展的现状，面向未来新型网络架构的前沿性研究，考虑新型网络架构的特点和发展现状，基于新型网络架构的组播成员管理协议可以在以下几个方面开展下一步研究与开发工作：

1. 组播管理服务器的功能扩展

本文在MMS的设计实现时从实际需求出发，满足对组播成员的基本管理和维护功能，但若要实现商业化应用，仍需要进步完善组播管理服务器的服务功能。比如组播数据的加密、安全秘钥分发与更新机制、基于耗时或流量实施计费等。由于新型网络架构中的接入网络层终端设备标识具有可变长的特点，IDGMP消息字段还需要进一步扩展和完善，保障不同长度的接入标识均能申请和享用组播服务。此外，针对MMS对组播成员生成的安全秘钥可以考虑根据用户身份等级差异采用不同强度的加密算法，加强组播消息的安全性。

1. 组播成员管理协议的性能研究

本文所设计的基于新型网络架构的组播成员管理协议主要在简单网络拓扑环境下实现了组播管理的基本功能、在少量用户的情况下完成了简单功能测试，尚未部署到实际生产环境中。如果在大规模网络环境下进行高并发的集中测试，就需要考验MMS的服务性能和消息处理性能。在复杂多用户需求的情况下，协议本身能否正常的工作还需要进一步的研究和测试。

# 参考文献

1. W. Richard Stevens著. 范建华, 胥光辉, 张涛等译. TCP/IP详解 卷1：协议 [M]. 第1版. 北京. 机械工业出版社. 2004. 1-12,15-37, 170-254
2. Luo H, Lin Y, Zhang H, et al. Preventing DDoS attacks by identifier/locator separation[J]. Network, IEEE, 2013, 27(6): 60-65.
3. P. Newman, In Search of The All-ip Mobile Network [J], IEEE Communications Magazine, 2004, 42(1): 3-8.
4. D. Meyer，L. Zhang, K. Fall, “Report from the IAB Workshop on Routing and Addressing,” RFC 4984, IETF, September 2007.
5. D. Clark, R. Braden, A. Falk, and V. Pingali. FARA: Reorganizing the Addressing Architecture. Proc. ACM SIGCOMM Workshop on Future Directions in Network Architecture (FDNA), Karlsruhe, Germany, August 2003, pp. 313-321.IP地址二义性
6. Hongke Zhang and Wei Su, Fundamental Research on the Architecture of New Network -- Universal Network and Pervasive Services[J], Acta Electronica Sinica, Vol.35, No.4, 2007, pp.593-598
7. Ping Dong, Yajuan Qin and Hongke Zhang, Research on Universal Network Supporting Pervasive Services[S], Acta Electronica Sinica, Vol.35, No.4, 2007, pp. 599-606.
8. 杨冬, 周华春, 张宏科. 基于一体化网络的普适化服务研究[J]. 电子学报. 2007.
9. S. Deering, Host Extensions for IP Multicasting, RFC1112 IETF, August 1989.
10. W. Fenner, Intemet Group Management Protocol. Version 2. IETF RFC 2236. November 1997.
11. B. Cain, S. Deering, I. Kouvelas, B. Fenner, A. Thyagarajall. Intemet Group Management Protocol. Version 3. IETF RFC 3376. October 2002.
12. S. Deering, W. Fenner, B. Haberman, Multicast Listener Discovery (MLD) for IPv6. IETF RFC 2710. October 1999.
13. R. Vida, L. Costa, Multicast Listener Discovery Version 2 (MLDv2) for IPv6. IETF RFC 3810. June 2004.
14. A. Conta, S. Deering, M. Gupta, “Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification,” RFC 4443, IETF, March 2006.
15. B. Fenner, H. He, B. Haberman, H. Sandick, Internet Group Management Protocol (IGMP)/Multicast Listener Discovery (MLD)-Based Multicast Forwarding (“IGMP/MLD Proxying”). IETF RFC 4605. August 2006.
16. M. Christensen, K. Kimball, F. Solensky. Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches. IETF RFC4541. May 2006
17. S. Venaas, R. Parekh, G. Van de Velde, T. Chown, M. Eubanks. Multicast Addresses for Documentation. IETF RFC6676. August 2012.
18. H. Krawczyk, M. Bellare, R. Canetti. HMAC: Keyed-Hashing for Message Authentication. IETF RFC 2104. February 1997.
19. Daniel P. Bovet and Marco Cesati著, 陈莉君等译. 深入理解Linux内核 [M], 北京, 中国电力出版社, 2009
20. [美]桑德斯著, 诸葛建伟, 陈霖, 许伟林译. Wireshark数据包分析实战. 第二版. 人民邮电出版社. 2013
21. S. Bhattacharyya. An Overview of Source-Specific Multicast (SSM). IETF RFC 3569. July 2003.
22. C. Hedrick, Routing Information Protocol. IETF RFC 1058. June 1988.
23. D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson，C. Liu, P. Sharma, L. Wei. Protocol Independent Multicast-Sparse Mode (PIM-SM). IETF RFC 2362．June l998.
24. J. Moy, OSPF Version 2. IETF RFC2328. April 1998.
25. D. Waitzman, C. Partridge, S. Deering. Distance Vector Multicast Routing Protocol. IETF RFC 1075. November 1988.
26. David C. Plummer, “48.bit Ethernet Address for Transmission on Ethernet Hardware,” RFC 826, IETF, November 1982.
27. S. Gundavelli, M. Townsley, O. Troan, W. Dec. Address Mapping of IPv6 Multicast Packets on Ethernet. IETF RFC6085. January 2011.
28. IP Version 6 Addressing Architecture. R. Hinden, S．Deering. IETF RFC 2373. July 1998.
29. 王上. 一体化网络接入交换路由器分离映射的设计与实现[D]. 北京交通大学. 2008年.
30. Michael Kofler著, 杨晓云, 王建桥, 杨涛等译. MySQL 5权威指南. 第三版. 人民邮电出版社. 2006
31. Vikram Vaswani. 徐小青, 路晓村等译. MySQL完全手册. 电子T业出版社. 2004
32. Klaus Wehrle Frank Pahlke, Harmut Pitter, The Linux Network Architecture[M],北京，清华大学出版社，2006
33. W. Richard Stevens, Bill Fenner, Andrew M. Rudoff, UNIX Network Programming. Volume 1: The Sockets Networking API. Third Edition. 人民邮电出版社. 2010.56-72,150-173,580-622
34. Neil Horman, Understanding And Programming With Netlink Sockets [M], December 6, 2004, pp. 4-10
35. （巴西）莱鲁萨利姆斯奇（Ierusalimschy，R.）著, 周惟迪译. Lua程序设计. 第二版. 电子工业出版社. 2008
36. 张爱华. 下一代互联网路由体系架构下映射系统的研究与实现[学位论文].北京邮电大学. 2010.p14-15
37. 陈旭明. 一体化网络接入认证方案的设计与实现[学位论文].北京交通大学.2008.p27-44
38. 郭华明. 标识路由关键技术研究[学位论文].北京交通大学.2010.p22-42
39. 朱莹. 一体化网络中基于标识的组播路由机制设计与实现[学位论文].北京交通大学.2009.p17-28
40. 李晓倩. 一体化标识网络身份与位置映射关键技术研究[学位论文].北京交通大学.2013.p32-54
41. 杨凡. 基于ipv6的组播安全技术研究[学位论文].西安电子科技大学.2010.p35-49
42. 蒋玉芳. 数字电视网络传输中的组播密钥管理方案研究.中国科技博览. 2009(24)
43. 关建峰. 基于IPv6的移动组播关键技术研究[学位论文]. 北京交通大学.2009.p26-35

# 致谢

一转眼，研究生两年半的学习生涯就快要结束了，中间有太多值得回味与怀念的故事，不禁感慨时光飞逝的同时，难忘人生又一宝贵的经历。入世校园生活将近二十载，终于要准备出世去进入另一个更广阔的空间发展自我了。此时此刻，我有太多的话想对身边的人说一声谢谢，感谢这两年半来给予我关心和帮助的人。

首先感谢我的导师张宏科老师，感谢张老师在学习和生活上给我们的诸多建议和关怀，张老师对待科研的精神和对学生的严格要求，不断督促着我们努力上进。其次感谢下一代互联网中心主任许长桥老师和指导恩师关建峰老师，感谢许老师给我一次继续深造的机会，他不仅给实验室营造了良好的科研氛围，让我们能够在舒适的实验室环境下安心的学习和成长，同时他那敬业的工作态度深深的感染了我们，成为我们以后工作学习的榜样，在此衷心的感谢许长桥老师。另外，还要特别感谢关建峰老师，感谢关老师对我的栽培和指导，让我在研究生期间得以快速的成长。关老师作为老师，给予我学习工作上莫大的指点和帮助，作为朋友，给予我诸多过来人的建议与经验，让我醍醐灌顶，受益匪浅，在此送上深深的祝福与感激之情。祝愿老师们在以后的工作中再结硕硕成果。

感谢下一代互联网中心实验室的所有兄弟姐妹们，感谢你们的陪伴让我度过了一个快乐轻松的研究生生活。感谢所有2013级的研究生战友们，是你们在我困惑无解时给我关心和帮助，是你们在项目程序调不通时帮助我修复bug，感谢生命中遇到了你们，我会永远珍藏生命路途中这段美好的记忆。

感谢在校园里、宿舍楼、科研楼打扫卫生的叔叔阿姨，是你们提供了清洁的环境，让我们度过了美好的校园生活。

最后感谢所有关心、鼓励、支持我的家人、亲戚和朋友。