复旦大学

硕士学位论文

基于改进K-means算法的Web文档聚类系统的研究与实现

姓名：王钦平

申请学位级别：硕士

专业：通信与信息系统

指导教师：张世永

20070508

中文摘要

本文研究了一种基于改进 K-means 算法的 Web 文档聚类系统，并开发出了一套由网络爬虫、数据清理、中文分词、特征提取、权重计算和聚类分析等模块组成的 Web 文档聚类系统。同时，针对 K-means 算法的主要缺点和不足，本文对 K-means 算法中的关键环节如相似度计算公式，初始聚类中心的选择和新聚类中心的计算方法进行了改进。并且使用 F-measure 评价方法对 K-means 算法整体改进后的聚类效果进行评价，通过实验性能对比说明了改进算法的优越性。

文章对数据挖掘、聚类分析和 Web 挖掘进行了概述和总结，介绍了整个系统的架构。并对网络爬虫、中文分词、英文词干提取、特征提取，权重计算和聚类分析等模块进行了深入的研究。最后，通过开发的由网络爬虫、数据清理、中文分词、特征提取、权重计算和聚类分析等模块组成的 Web 文档聚类系统进行了对比实验，验证了基于改进 K-means 算法的 Web 文档聚类系统在准确性和稳定性方面都有所提高。

关键词：数据挖掘，聚类，Web 挖掘，K-means 聚类算法，向量空间模型

Abstract

This dissertation will study the realization of Web documents clustering system (WDCS) based on the improved K-means algorithm, and design a novel WDCS including several modules such as Web spider, Chinese split, English stemmer, feather selection, weighting calculation and clustering. On the other hand, the dissertation will improve the performance of K-means algorithm by several methods such as optimizing the similarity calculation, upgrading the initial clustering centers and changing the selection of the new clustering centers. By F-measure evaluation method, the simulated results will prove that the performance of the K-means algorithm can be improved.

The dissertation will also summarize the main feather of data mining (DM), clustering analysis (CA) and Web mining (WM), introduce the framework of the WDCS. In addition, the key technologies of Web spider, Chinese split, English stemmer, feather selection, weighting calculation and clustering analysis will be clarified in this dissertation. Finally, using the simulated results of designed WDCS, the improvements for the correctness and stability of the improved K-means algorithm are confirmed.

Key words: Data Mining, Clustering, Web mining, K-means clustering algorithm, VSM

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：王欢军 日期：2007.6.9

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：王欢军 导师签名：张世永 日期：2007.6.9

第一章 绪论

1.1 论文的研究背景和意义

Internet 和数字图书馆等领域的发展导致了信息的“爆炸”。随着以数据库、数据仓库等数据仓储技术为基础的信息系统在各行各业的应用，海量数据不断产生。随之而来的问题是如此多的数据让人难以消化，更无法从表面上看出他们所蕴含的有用信息，更不用说有效地指导进一步的工作。如何从大量的数据中找到真正有用的信息成为人们关注的焦点，数据挖掘技术也正是伴随着这种需求从研究走向应用[1]。

随着 Internet/Web 技术的快速普及和迅猛发展，使各种信息可以以非常低的成本在网络上获得。由于 Internet/WWW 在全球互连互通，可以从中取得的数据量难以计算，而且 Internet/WWW 的发展趋势继续看好，特别是电子商务的蓬勃发展为网络应用提供了强大支持。如何在 WWW 这个全球最大的数据集合中发现有用信息无疑将成为数据挖掘研究的热点[3]。

信息中最主要的信息源就是文本数据，传统的文档和文本处理工具已经不能满足用户的需求。于是在人工智能研究领域结合结构化数据库中的数据挖掘技术，提出了一种有效的、可以充分利用这些文本数据的新的信息处理技术— 文本挖掘(text mining)。文本挖掘是抽取有效、新颖、有用、可理解的、散布在文本文件中的有价值知识，并利用这些知识更好地组织信息的过程。文本挖掘是信息挖掘的一个研究分支，用于基于文本信息的知识发现。文本挖掘是利用智能算法，如神经网络、基于案例的推理、可能性推理等，并结合文字处理技术，分析大量的非结构化文本源，如文档、电子表格、客户电子邮件、问题查询、网页等，抽取或标记关键字概念，文字间的关系，并按照内容对文档进行分类，获取有用的知识和信息。它涉及多个学科领域，诸如数据库、信息位索、信息提取、机器学习、自然语言处理、计算语言学、统计数据分析、线性几何、概率理论，甚至还有图论等知识[3]。

Web 挖掘指使用数据挖掘技术在 WWW 数据中发现潜在的、有用的模式或信息。Web 挖掘研究覆盖了多个研究领域，包括数据库技术、信息获取技术、统计学、人工智能中的机器学习和神经网络等[8]。

聚类是文本挖掘的主要内容之一。它是根据某种相似性准则将样本空间分成多个子空间，使每个子空间内部样本点尽可能相似，不同子空间内样本点之间差异尽可能大。其实质是寻找隐藏在数据中不同的数据模型，是一个无监督学习过程，能够实现样本空间的盲分类。聚类广泛应用于统计、机器学习、模式识别，数据分析等领域，并越来越受重视。聚类可以帮助人们更快的找到所需要的信息，因此，聚类在现实生活中有着很重要的意义。现在，聚类分析已成为一个非常活跃的研究课题[5]。

1.2 相关内容的研究情况

在加拿大蒙特利尔市召开了第一届 KDD （knowledge discover database，是数据挖掘的另一种叫法——"数据库中的知识发现"的英文）国际学术会议[1]，以后每年召开一次。近年来，KDD 在研究和应用方面发展迅速，在电信、银行、商业等领域得到了广泛的应用，SAS, SPASS 等诸多软件都提供了数据挖掘的功能[3]。

数据挖掘中的聚类分析是传统聚类方法（特别是统计学中的聚类方法）的继承和发展。聚类分析作为数据挖掘中的一个重要研究热点，目前，聚类分析的研究主要集中在两个方面：一方面是对聚类分析算法的研究，现在已经提出了很多种算法，比如：K-means、CLARANS、BIRCH、CURE、Chameleon、DBSCAN、OPTICS、DENCLUE、STING 算法等。特别是对文本、多媒体和 Web 信息等复杂数据进行聚类的研究还是方兴未艾。另一方面是聚类分析的实际应用的研究，聚类分析可以作为独立的数据挖掘工具来获得数据分布的知识，也可以作为其它数据挖掘算法的预处理步骤。它在许多领域有着重要的应用，比如金融数据分析，商业零售数据分析，电信和网络数据分析，生物医学和 DNA 数据分析，天文，陆地和海洋地理等科学探测数据分析等领域[5]。

1.3 论文的研究内容和组织结构

本文主要做了下述工作：

（1） 本文对文本挖掘中的文本表示方法、特征提取、权值计算进行了系统的研究，并对网络爬虫、分词、聚类等过程进行了比较详细的阐述。

（2）开发了一整套由网络爬虫、数据清理、中文分词、特征提取、权值计算和聚类分析等模块组成的系统。

（3）对K-means算法进行了有特色的改进。区别于其它论文中只对算法的某一方面进行改进，本文综合了大量参考资料，针对K-means算法的主要缺点和不足，对K-means算法中的关键环节：相似度计算公式，初始聚类中心的选择和新聚类中心的计算方法进行了改进。

（4）使用F-measure评价方法对K-means算法改进前后的系统进行评价，通过F-measure值对比表、对比柱状图和F-measure分布情况说明了改进算法的在准确性和稳定性方面都有所提高。

（5）论文最后除对研究工作进行了总结外，还对今后的研究方向进行了展望。

本文的组织结构如下：

• 第一章“绪论”，首先分析了论文研究的背景和意义，然后介绍了本文相关内容的研究情况，最后介绍了论文的研究内容和组织结构。

• 第二章“数据挖掘与聚类分析”，是对数据挖掘，聚类分析和Web挖掘等本文相关内容的综述。

• 第三章“Web文档聚类系统的总体架构”，首先给出了Web文档聚类系统的总体架构图，然后分别介绍了网络爬虫、数据清理、中文分词、特征提取、权值计算和聚类分析等模块。

• 第四章“K-means算法和基于改进K-means算法的聚类分析模块”，首先介绍了原始的K-means算法，然后对其进行了改进。

• 第五章“基于改进K-means算法的Web文档聚类系统的实验及结果”，首先介绍了聚类结果的评价方法，然后利用本文开发的系统对改进前后的K-means算法的聚类质量进行了实验和比较，并给出了对比数据表和相应的柱状图。

• 第六章“结束语”，对本文的工作进行了总结，同时指出了还需进一步研究的方向。

第二章 数据挖掘与聚类分析

2.1 数据挖掘

2.1.1 数据挖掘的产生

Internet 和数字图书馆等领域的发展导致了信息的“爆炸”。随着以数据库、数据仓库等数据仓储技术为基础的信息系统在各行各业的应用，海量数据不断产生。随之而来的问题是如此多的数据让人难以消化，无法从表面上看出他们所蕴涵的有用信息，更不用说有效地指导进一步的工作。如何从大量的数据中找到真正有用的信息成为人们关注的焦点，数据挖掘技术也正是伴随着这种需求从研究走向应用[1]。

随着 Internet/Web 技术的快速普及和迅猛发展，使各种信息可以以非常低的成本在网络上获得。由于 Internet/WWW 在全球互连互通，可以从中取得的数据量难以计算，而且 Internet/WWW 的发展趋势继续看好，特别是电子商务的蓬勃发展为网络应用提供了强大支持。如何在 WWW 这个全球最大的数据集合中发现有用信息无疑将成为数据挖掘研究的热点[3]。

KDD 这一术语首先出现在 1989 年在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上，1991，1993 和 1994 年又接着继续举行 KDD 专题讨论会。1995 年在加拿大召开了第一届知识发现和数据挖掘国际学术会议，以后每年召开一次。从 1997 年开始，KDD 已经拥有了专门的学术刊物《Knowledge Discovery and Data Mining》。国外在这方面发表了众多的研究成果和论文，并且开发了一大批数据挖掘软件，建立了大量的相关网站。对 KDD 和数据挖掘的研究已成为计算机领域的一个热门课题。我国近几年也逐渐跟上国际步伐，许多计算机、数据库、人工智能、机器学习领域的专家学者投入到 KDD 和数据挖掘的研究中，并已取得了一定的成果[3]。

2.1.2 数据挖掘的定义

知识发现[2] (KDD: knowledge discovery in database)是指从大量数据中提取出可信的、新颖的、有效的，并能被人理解的模式的高级处理过程。这是根据

W. J. Frawley 和 G. P. Shapiro 等人提出的定义。

在信息时代，用户可获得的信息包含了从技术资料、商业信息到新闻报道、娱乐资讯等多种类别和形式的文档构成了一个异常庞大的具有异构性、开放性的分布式数据库。数据挖掘(data mining)是指从大型数据库的数据中提取出人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在的有用信息。而这个数据库中存放的是非结构化的文本数据。结合人工智能研究领域中的自然语言理解和计算语言学，从数据挖掘中派生出了两类新兴的数据挖掘研究领域：网络挖掘和文本挖掘。网络挖掘侧重于分析和挖掘网页相关的数据包括文本、链接结构和访问统计(最终形成用户网络导航)。一个网页里包含了多种不同的数据类型。因此网络挖掘就包含了文本挖掘、数据库中数据挖掘、图像挖掘等。文本挖掘作为一个新的数据挖掘研究领域，目前并没有给出统一的、确切的定义。但是文本挖掘的目的就是把文本信息转化为人可利用的知识。从这一目的出发，有如下 5 种关于文本挖掘的定义[3]：

定义 1 文本挖掘是指对大规模文档集的处理和从文本数据集中提取隐含的知识。

定义 2 文本挖掘是信息处理和信息管理中的重要研究问题。它基于语义学，使用贝叶斯模型、概率理论、向量空间模型、统计模型甚至是图论，从文档中挖掘出知识模式，及短语结构。

定义 3 文本挖掘是使用计算语言学规则从文本中提取信息的研究和应用方法。文本挖掘的关链领域包括特征提取、主题索引、聚类、摘要。

定义 4 文本挖掘是指从自然语言文本中提取模式。可以定义为按照特定目标从文本中提取信息的分析过程。

定义 5 文本挖掘结合数据挖掘的规则、信息提取、信息检索、文本分类、概率模型、线性几何、机器学习、计算语言学去发现文本集中的结构、模式、知识。

根据以上描述并结合数据挖掘的定义，可以总结出文本挖掘的定义[3]为：文本挖掘(text mining)以计算语言学、统计数理分析为理论基础，结合机器学习和信息检索技术从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识。它是一个从文本信息描述到选取提取模式，最终形成用户可理解的信息知

识的过程。

2.1.3 数据挖掘的功能

从数据库中发现隐含的、有意义的知识是数据挖掘的目标。数据挖掘的功能主要有以下六类[4]：

(1) 概念描述

数据库中通常存放大量的细节数据。然而，用户通常希望以简洁的描述形式观察汇总的数据集。这种数据描述可以提供一类数据的概貌，或将它与对比类相区别。此外，用户希望方便、灵活地以不同的粒度和从不同的角度描述数据集。这种描述性数据挖掘称为概念描述。

(2) 关联分析

关联分析就是从大量数据中发现项集之间有趣的关联或相关联系。随着大量数据不停地收集和存储，许多业界人士对于从他们的数据库中挖掘关联规则越来越感兴趣。从大量商务事务记录中发现有趣的关联关系，可以帮助许多商务决策的制定。

(3) 分类和预测

分类和预测是两种数据分析形式，可以用于提取描述重要数据类的模型或预测数据未来的趋势。分类和预测的应用十分广泛，例如,可以建立一个分类模型，对银行的贷款客户进行分类，以降低贷款的风险；也可以通过建立分类模型，对工厂的机器运转情况进行分类，用来预测机器故障的发生。

(4) 聚类分析

根据最大化类内相似性、最小化类间相似性的原则进行聚类，使得在同一个类中的对象具有很高的相似性，而与其它类中的对象很不相似。聚类形成的每个类可以看作一个对象类，由它可以导出规则。聚类也便于将观察到的内容组织成分层结构，把类似的事件组织在一起。

(5) 孤立点分析

数据库中可能包含一些数据对象，它们与数据的一般行为或模式不一致。这些数据对象就是孤立点。许多数据挖掘算法试图使孤立点的影响最小化，或者排除它们。但在一些应用中孤立点本身可能是非常重要的信息。例如在欺诈探测中，

孤立点可能预示着欺诈行为。

(6) 演变分析

数据演变分析描述行为随时间变化的规律和趋势，并对其建模。可以从股票交易数据中挖掘出整个股票市场和特定公司的股票演变规律，以帮助预测股票市场的未来走向，帮助对股票投资做出决策。

2.1.4 数据挖掘的应用和研究方向

目前，数据挖掘的应用主要集中在以下几个方面[5]：

（1）金融数据分析；

（2）商业零售数据分析；

（3）电信和网络数据分析；

（4）生物医学和 DNA 数据分析；

（5）天文，陆地和海洋地理等科学探测数据分析等。

数据挖掘的应用越来越广泛。广泛的应用促使越来越多的研究机构，公司和学术组织从事数据挖掘系统原型与产品的研制和开发。这些系统和工具一般可根据其应用领域分为通用单任务类型，通用多任务类型和面性专用领域类型。它们的挖掘功能和方法上的差别不仅体现在关键技术上，还体现在运行平台、处理的数据类型、与数据库或数据仓库的耦合关系、提供的数据挖掘查询语言和可视化工具、价格等方面。但是，无论是专门用于某一方面和领域的系统，还是紧密结合数据库和数据仓库的综合的系统，除了采用了传统的统计方法外，还采用基于人工智能的技术，包括决策树、规则归纳、神经网络、可视化、模糊建模等，这是数据挖掘工具的发展趋势。同时，数据挖掘工具的开发不仅要面向专业人员，而且要面向非专业人员以及决策人员，这也是数据挖掘系统的一个发展方向[5]。

数据、数据挖掘任务和数据挖掘方法的多样性给数据挖掘提出了许多挑战性课题。数据挖掘语言的设计，高效而有用的数据挖掘算法和系统的开发，交互和继承的数据挖掘环境的建立，以及应用数据挖掘技术解决大型应用问题，都是目前数据挖掘研究人员、系统和应用开发人员所面临的主要问题。目前，数据挖掘的研究方向主要有[5]：

（1）数据挖掘的应用研究；

（2）可伸缩的数据挖掘算法研究；

（3）数据挖掘与数据库、数据仓库和 Web 数据库系统的集成；

（4）数据挖掘语言的标准化研究；

（5）数据挖掘的可视化研究；

（6）对于复杂数据类型进行挖掘的新方法研究；

（7）数据挖掘中的隐私保护与信息安全。

2.2 聚类分析

数据挖掘中的聚类分析是传统聚类方法（特别是统计学中的聚类方法）的继承和发展，是数据挖掘中的一个重要研究热点。

聚类分析是研究数据间逻辑上或物理上的相互关系的技术，它通过一定的规则将数据集划分为在性质上相似的数据点构成的若干个类。聚类分析的结果不仅可以揭示数据间的内在联系与区别，同时也为进一步的数据分析与知识发现提供了重要的依据，如数据间的关联规则，分类模式以及数据的变化趋势等。作为统计学的重要研究内容之一，聚类分析具有坚实的理论基础并形成了系统的方法学体系，然而，基于统计学的聚类分析方法大多局限于理论上的分析并依赖于对数据分布特征的概率假设，较少考虑具体应用中的实际数据特征与差异。由于数据挖掘技术的迅速崛起，聚类分析得以在数据库技术领域获得长足的发展[5]。

2.2.1 聚类分析的定义

聚类（clustering）是一种常见的数据分析工具，简单地说，就是将物理或抽象对象的集合分组成为由类似的对象组成的多个类或类（cluster）的过程。由聚类所生成的类是对象的集合，这些对象与同一个类中的对象彼此相似，与其它类中的对象相异。在许多应用中，可以将一个类中的数据对象作为一个整体来对待[4]。

2.2.2 数据挖掘对聚类分析的要求

在数据挖掘领域，研究工作已经集中在大型数据库的有效和实际的聚类分析寻找适当的方法。活跃的研究主题集中在聚类方法的可伸缩性，方法对聚类复杂

形状和类型的数据的有效性，高维聚类分析技术，以及针对大型数据库中混合数据的聚类方法[5]。

具体地说，数据挖掘对聚类的特殊要求如下[5]：

•1，伸缩性：这里的可伸缩性是指算法要能够处理大数据量的数据库对象，比如处理上百万条纪录的数据库。这就要求算法的时间复杂度不能太高，最好是多项式时间的算法。

•2，处理不同字段类型的能力：即算法不仅要能处理数值性的字段，还要有处理其它类型字段的能力，例如：布尔型、枚举型、序数型以及混合型等。

•3，发现具有任意形状聚类的能力：很多聚类分析算法采用基于欧几里德距离的相似性度量方法。这一类算法发现的聚类通常是一些球状的、大小和密度相近的类。但可以想见，现实数据库中的聚类可以是任意的形状，甚至是具有分形维度的形状，故要求算法有发现任意形状聚类的能力。

•4，输入参数对领域知识的弱依赖性：很多聚类算法都要求用户输入一些参数，例如需要发现的聚类数，结果的支持度、置信度等。聚类分析的结果通常都对这些参数很敏感，另一方面，对于高维数据，这些参数又是相当难以确定的。这样就加重了用户使用这个工具的负担，使得分析的结果很难控制。一个好的聚类算法应该针对这个问题，给出一个好的解决方法。

•5，增加限制条件后的聚类分析能力：现实的应用中总会出现各种其它限制，我们希望聚类算法可以在考虑这些限制的情况下，仍旧有较好的表现。

•6，能够处理异常数据：现实数据库中常常包含有异常数据，或者数据不完整，缺乏某些字段的值，甚至是包含错误数据的现象。有一些聚类算法可能会对这些数据很敏感，从而导致错误的分析结果。

•7，处理高维数据的能力：一个数据库或者数据仓库都有很多的字段或者说维。一些分析算法再处理维数比较少的数据集时表现不错，例如两、三维的数据;人的理解能力也可以对两、三维数据的聚类分析结果的质量做出较好的判别。但对于高维数据就没有那么直观了。所以对于高维数据的聚类分析是很具有挑战性的，特别是考虑到在高维空间中，数据的分布是极其稀疏的，而且形状也可能是极其不规则的。

•8，结果对输入记录顺序的无关性：有些分析算法对纪录的输入顺序是敏

感的，也即，对同一个数据集，将它以不同的顺序输入到分析算法，得到的结果会不同，这是我们不希望的。

2.2.3 主要聚类方法及其研究进展评述

以下将讨论在数据挖掘领域中得到广泛应用的一些聚类算法及其改进方法。目前在文献中存在大量的聚类算法。算法的选择取决于数据的类型、聚类的目的和应用。聚类算法大体上可以分为五大类[4]。

1）划分方法

给定一个包含 n 个数据对象的数据集，一个划分方法构建数据的 k 个划分，每个划分表示一个类，并且 $k=n$。也就是说，它将数据划分为 k 个组，同时满足如下的要求：(i) 每个组至少包含一个对象；(ii) 每个对象必须属于一个组。给定要构建的划分的数目 k，划分方法首先创建一个初始划分。然后采用一种迭代的重定位技术，尝试通过对象在划分间移动来改进划分。一个好的划分的一般准则是：在同一个类中的对象之间尽可能"接近"或相关，在不同类中的对象之间尽可能"远离"或不同，即使下列准则函数最小，

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - m_i\|, \qquad (2\text{-}1)$$

式中的 E 是数据集中所有对象的平方误差的总和；m_i 是类 C_i 的平均值（或中心点），p 是数据空间中的数据对象（p 和 m_i 都是多维的)。为了达到全局最优，基于划分的聚类要求穷举所有可能的划分。实际上，绝大多数应用采用了以下两个比较流行的启发式方法：

• 1，K-means 算法

该方法首先由 MacQueen 提出[6]，在数据挖掘领域中得到了广泛的应用。在 K-means 算法中，每个类用该类中对象的平均值来表示（故此得名）。K-means 算法是解决聚类问题的一种经典方法。它的主要优点是算法简单、快速。然而这种方法对不同的 k 值可能会导致不同的聚类结果。其次，该方法不能发现非凸面的类，或大小差别很大的类。而且对"噪声"和孤立点很敏感，因为少量的该类数据能对平均值产生极大的影响。

• 2，K-medoids 算法

在 Kaufman 和 Roussseeuw 提出的 PAM（Partitioning around Medoid）和 CLARA（Clustering Large Applications）[7]算法中，每个类用接近该类中心的对象来表示，因此称之为 K-medoids 算法。K-medoids 算法可以看作是 K-means 算法的改进方法，因为中心点不像平均值那么容易被极端数据影响，所以当存在噪声和孤立点数据时，K-medoids 算法比 K-means 算法更强壮。

上述启发式方法对在小规模的数据库中发现球状类很适用。为了对大规模的数据集进行聚类，以及处理复杂的聚类，基于划分的方法有了很多改进算法。

将 K-means 算法与其它技术结合也可大大提高 K-means 算法的聚类能力。比如利用遗传算法较好地解决了全局最优(或近似最优)解的问题；利用窗口技术提高了 K-means 的聚类能力；有人将免疫规划与 K-means 聚类相结合，提出了基于免疫规划的 K-means 聚类方法；文献采用禁忌搜索技术聚类分类数据；文献也从不同角度改进了传统的 K-means 方法。

近年来，文献中出现了大量的模糊 K-means 聚类方法，也是对传统 K-means 聚类方法的有益扩展。

2）层次方法

层次的方法对给定的数据对象集合进行层次的分解。根据层次的分解如何形成，层次的方法可以分为凝聚和分裂两大类。凝聚的方法，也称为自底向上的方法，一开始将每个对象作为单独的一个类，然后相继地合并相近的类，直到所有的类合并为一个（层次的最上层），或者达到一个终止条件。分裂的方法，也称为自顶向下的方法，一开始将所有的对象置于一个类中。在迭代的每一步中，类被分裂为更小的类，直到每个类只包含一个对象为止，或者达到一个终止条件。

在凝聚或者分裂层次聚类方法中，通常以用户定义希望得到的类的数目作为结束条件。

在类的合并或分裂过程中，需要考察类间的距离。类间距离的度量广泛采用如下四种方法：最小距离，最大距离，平均值距离和平均距离。

基本的层次聚类方法是由 Kaufman 和 Rousseeuw 提出的凝聚方法 AGNES（agglomerative nesting）和分裂方法 DIANA（divisive analysis）。

层次聚类方法虽然简单，但经常会遇到合并点或分裂点选择的困难。这样的决定是非常关键，因为一旦一组对象被合并或者分裂，下一步的处理将在新生成

的类上进行。已做的处理不能被撤销，类之间也不能交换对象。如果在某一步没有很好地做出合并或分裂的决定，可能会导致低质量的聚类结果。而且，这种聚类方法不具有很好的可伸缩性。因此人们提出众多改进的层次聚类算法以改进层次聚类方法的性能。

3）基于密度的方法

绝大多数划分方法基于对象之间的距离进行聚类，这样的方法只能发现球状的类，而在发现任意形状的类上遇到困难。因此，出现了另一类基于密度的聚类方法，其主要思想是：只要邻近区域的密度（对象或数据点的数目）超过某个阈值，就继续聚类。也就是说，对给定类中的每个数据点，在一个给定范围的区域内必须至少包含某个数目的点。这样的方法可以过滤"噪声"数据，发现任意形状的类。但算法计算复杂度高，一般为 $O(n^2)$，而且对于密度分布不均的数据集，往往得不到满意的聚类结果。

多数基于密度的聚类方法对参数都很敏感，参数设置的细微不同可能导致差别很大的聚类结果。

4）基于网格的方法

基于网格的方法把对象空间量化为有限数目的单元，形成一个网格结构。所有的聚类操作都在这个网格结构（即量化空间）上进行。这种方法的主要优点是它的处理速度很快，其处理速度独立于数据对象的数目，只与量化空间中每一维的单元数目有关。但这种算法效率的提高是以聚类结果的精确性为代价的。

Wang 等提出的 STING（statistical information grid）和 STRING 是基于网格的多分辨率方法。该种方法效率高，而且网格结构有利于并行处理和增量更新，但其降低了聚类的质量和精确性。

5）基于模型的方法

基于模型的方法为每个类假定一个模型，寻找数据对给定模型的最佳拟合。基于模型的方法主要有两类：统计学方法和神经网络方法。

•1，统计学方法

常用方法包括 Fisher 提出的 COBWEB，Gennari 等提出的 CLASSIT，及 Cheeseman 等提出的 AutoClass, Pizzuti Clara 等提出的 P-AutoClass 等。

•2，神经网络方法

竞争学习（Competitive Learning）采用若干个单元的层次结构，它们以一种“胜者为王（winner-take-all）”的方式对系统当前处理的对象进行竞争。

学习矢量量化(Leaning Vector Quantization 简称 LVQ)是由 Kohonen 提出的。这是一种自适应数据聚类方法，它基于对具有期望类别信息数据的训练。尽管是一个有监督训练方法，然而 LVQ 采用了无监督数据聚类技术，对数据集进行预处理，可获得聚类中心。文献提出在特定的条件下更新获胜单元和第二单元(下一个最接近的向量)，以便更有效地利用训练数据。

自组织特征映射（Self-Organizing Feature Map 简称 SOFM），也由 Kohonen 提出。以其所具有的诸如拓朴结构保持、概率分布保持、无导师学习及可视化等特征，广泛应用于聚类分析之中。

除上述五大类方法以外，在各种文献中还存在着大量的聚类方法。如基于遗传算法的聚类方法；模糊聚类方法；处理高维数据的聚类方法；处理大规模数据的聚类方法；处理动态数据的聚类方法；以及将基本聚类方法与各种新技术相结合的聚类方法等。因此也出现了大量文献对各种方法进行了比较研究。

所有的聚类方法都具有各自的特点[4]。有些以方法简单、执行效率高见长（如 K-means）；有些对任意形状、大小的类识别能力强（如 CUBN）；有些能很好的过滤噪声数据（如 DBSCAN）。但这些方法都有各自的局限性。如 K-means 方法只能识别大小近似的球形类；CUBN、DBSCAN 的时间复杂度都为 O(n2)。另外，很多聚类方法对输入参数十分敏感，而且参数很难确定，这加重了用户的负担。因此，尽管已存在众多的聚类方法，人们仍在致力于研究聚类能力强、执行效率高、参数设置简单的聚类方法。

2.3Web 挖掘

2.3.1Web 挖掘

伴随着互联网上的“信息爆炸”，Web 挖掘从数据挖掘“脱颖而出”。但是，Web 挖掘与传统的数据挖掘相比有许多独特之处。Web 挖掘是指从大量，异质，分布的 Web 文档的集合中抽取感兴趣的，有用的模式和隐含信息[8]。

Web 挖掘包括以下四个子任务：信息检索，信息提取，预处理和概括，分析。

一般地，Web 挖掘可分为三类[8]：Web 内容挖掘(web content mining)、结构挖掘(web structure mining)和 Web 使用记录的挖掘(web usage mining)。

(1) Web 内容挖掘是从文档内容或其描述中抽取知识的过程。主要包括直接对 Web 页面文档内容以及搜索引擎的查询结果进行文本的总结、分类、聚类、关联分析等。Web 内容挖掘包括：Web 文本挖掘和针对多媒体数据等的挖掘，不过目前研究较多的还是文本挖掘。

(2) Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识。由于文档之间的互连，WWW 能够提供除文档内容之外的有用信息。利用这些信息，可以对页面进行排序，发现重要的页面。

(3) Web 使用记录挖掘的主要目标则是从 Web 的访问记录中抽取感兴趣的模式。WWW 中的每个服务器都保留了访问日志(web access log)，记录了关于用户访问和交互的信息。分析这些数据可以帮助理解用户的行为，从而改进站点的结构，或为用户提供个性化的服务。

按照文本挖掘的对象可把文本挖掘分为：基于单文档的数据挖掘和基于文档集的数据挖掘。基于单文档的数据挖掘对文档的分析并不涉及其它文档，其主要的挖掘技术有：文本摘要、信息提取(包括名字提取、短语提取、关系提取等)。基于文档集的数据挖掘是对大规模的文档数据进行模式抽取， 其主要的技术有：文本分类、文本聚类、个性化文本过滤、文档作者归属、因素分析等。从功能上 Web 文本挖掘主要是对 Web 上大量文档集合的内容进行总结、分类、聚类、关联分析以及利用 Web 文档进行趋势预测等。Web 文本挖掘中，文本的特征表示是挖掘工作的基础，文本的分类和聚类是最重要、最基本的挖掘功能[8]。

2.3.2Web 文本挖掘的定义

Web 文本挖掘[9]是指借鉴数据挖掘的基本思想和理论方法，从大量非结构化、异构的 Web 文档的集合 *D* 中发现有效的、新颖的、潜在可用的及最终可理解的知识 *K*(包括概念、模式、规则、规律、约束及可视化等形式)的非平凡过程。如果将 *D* 看作输人，将 *K* 看作输出的话，那么 Web 文本挖掘的过程就是从输人到输出的一个映射 *e*：*D-K*。

在这里，过程通常是指多阶段的一个过程，涉及数据预处理、学习与知识模式的生成、模型质量的评价及反复的修改求精；该过程要求是非平凡的，意思是要有一定程度的智能化、自主性(仅仅给出所有数据的总和不能算作是一个发现过程)。而以上所提及的有效性、新颖性、潜在有用性和最终可理解性综合在一起可称为兴趣性[9]。

2.3.3Web 文本挖掘的研究方法和关键技术

Web 文本挖掘的目的[9]是帮助人们更好地发现、组织、表示信息，提取知识；发现用户所需文档的模式，找出用户的浏览兴趣，进而能够实现自动为用户提供相关度较高的文档等个性化服务。研究方法是通过对大量 Web 文档的内容进行文本分类、聚类和摘要等。

文本分类[9]：按照已经给定的主题类别，为文档集合中的每个文档确定一个类别。这样，不但能够方便用户浏览文档，而且可以通过限制搜索范围使文档的查找更为容易。文本分类的算法有很多种，目前比较常用的有 TF*IDF 等方法。下文会对分类方法有具体的介绍。

文本聚类[9]：将文档集合分成 K 个类，使处于同类内的文档相似度尽可能大，不同类的文档之间相似度尽可能小。利用文本聚类技术可以将搜索引擎的检索结果划分为若干个类，用户只需要处理那些相关类内的文档，大大缩小了所需要浏览的文档数量。文本聚类算法主要分成两类：以 G-HAC 等算法为代表的层次凝聚法和以 K-means 等算法为代表的平面划分法。

文本摘要[9]：从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释、概括。以便用户不需要浏览全文就可以了解文档或文档集合的大概内容。文本摘要在有些场合十分有用，例如，搜索引擎在向用户返回查询结果时，通常

需要给出文档的摘要。

Web 文本挖掘中的关键技术有文本的特征表示、特征提取、权值计算和聚类分析等，这些具体的技术细节将在下文中涉及到时的在具体介绍。

第三章 Web 文档聚类系统的总体架构及各模块介绍

3.1 系统的总体架构图

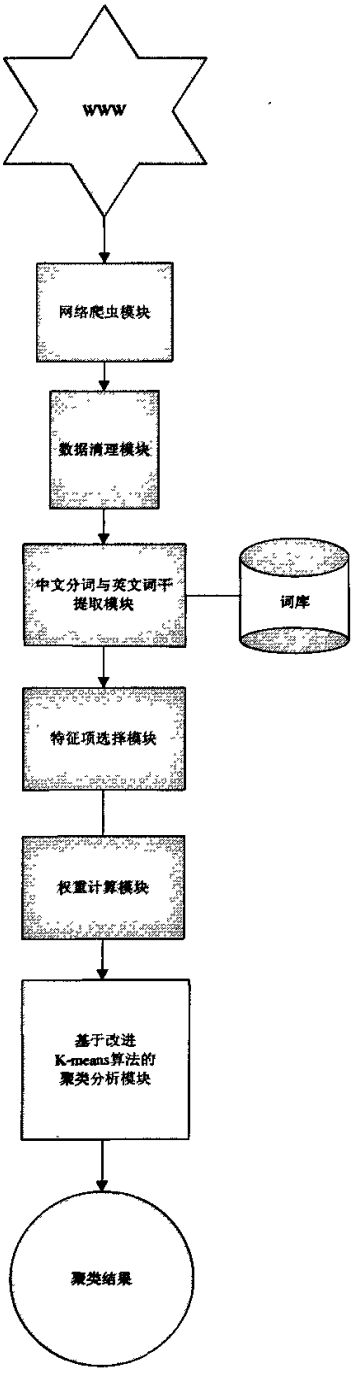


图 3-1 系统的总体架构图

图 3-1 给出了 Web 文档聚类系统的总体构架，从图中大致可以看出整个系统大致可以分为网络爬虫、数据清理、中文分词、特征提取、权值计算和聚类分析六个模块。在下面的章节中将分别予以详细介绍。

3.2 网络爬虫模块

网络爬虫系统是一个根据初始 URL 自动搜索 Internet 上网页的系统。首先要输入一些 URL，系统将这些 URL 放入一个队列。系统根据每一个 URL，自动到指定地址下载内容，并从这些内容之中解析新的 URL 地址；当得到新的地址之后，又重新将这些地址加入到队列之中。重复上述的行为，直到最终队列为空为止。将这些下载到的内容有序的保存到本地文件系统之中，并准备将其作为需要进行聚类的素材输入到聚类分析模块之中[10]。

3.2.1 网络爬虫模块的原理

本文将任何一个网页的 URL 看成是一个顶点 V[10]，如果此网页中含有指向其它网页的 URL U，那么我们将在有向图中加入一条弧(V,U)，如此类推，那么整个网络就可以看成为一个有向图(可能有环)[10]。图 3-2[10]概念性的给出了一个可能的网络拓扑图，在这个网络中：网页 S 拥有指向 $V1$、$V2$、$V3$ 的 URL 链接，$V1$ 中包含 $U1$ 的链接等等，同时还可能出现环路的可能，如：$V1 \rightarrow U1 \rightarrow T \rightarrow V1$。

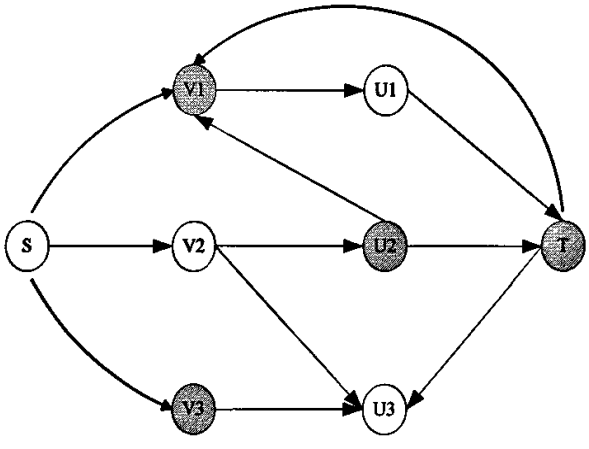


图 3-2 网络拓扑图

为了能够将网络上的所有网页抓取，首先必须得到所有网页的 URL，这就

相当于将上述的有向图作一个遍历。需要注意的是：当某网页已经抓取下来之后，我们不再需要重新获取[10]。由于本系统只需要能够将网页自动下载即可，采用如下简单通用算法[10]：

【算法 3-1】网络爬虫通用算法[10]：

```
1.  Set starting_url ← P0;
2.  enqueue(url_queue , starting_url);
3.  while (not empty(url_equeue)) do
    a)  set url  = dequeue(url_queue);
    b)  set page = crawl_page(url);
    c)  equeue(crawled_pages , (url, page));
    d)  set url_list ← extract_urls(page);
    e)  for (u∈url_list) do
        i.   enqueue(links , (url , u));
        ii.  if [u∉url_queue & [u , -]∉crawled_pages]
        iii. then enqueue(url_queue , u);
```

其中 url_queue 是用来保存即将访问的 URL，crawled_pages 是用来保存已经下载的 url 和 page，links 是用来保存 URL 之间关系的。

3.2.2 网络爬虫模块的研究与实现

上述算法描述起来是非常容易的，但在实际中由于以下两点原因[10]，系统不得不重新考虑其它的构思。

1，网页数目相当之大，导致系统运行速度急剧下降；

2，内存容量远远不够。

3.2.2.1 数据结构

下面是系统需要的三个数据结构[10]：

1，线性数组。图 3-3 是一个大的线性数组 url_table。在一个拥有成千上万数目的网络爬虫系统之中，首先需要这样一个数组，用来保存 URL 和对应文件标题的条目。这个数组指出已经下载下来的网页和已经保存下来的文件位置。由于这个数组非常之大，往往需要考虑虚存机制。在我们的系统之中，虚存管理由操作系统自动维护，系统不再介入。

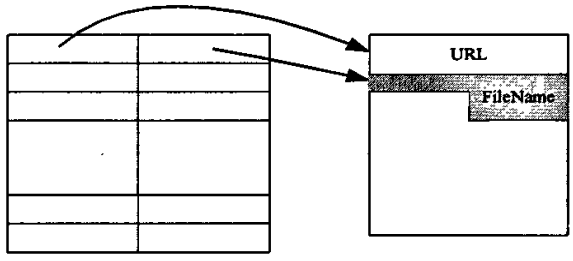


图 3-3 url_table 数据结构

2，散列表。如图 3-4 所示，用来快速的判断一个 URL 是否已经存在于系统之中。其中散列表右边的数字为 url_table 的序号，由于这个散列表是存在于内存之中的，因此必须尽量的减少其负荷量；左边框中的数字代表本桶项目的多少，主要用于持久化目的。

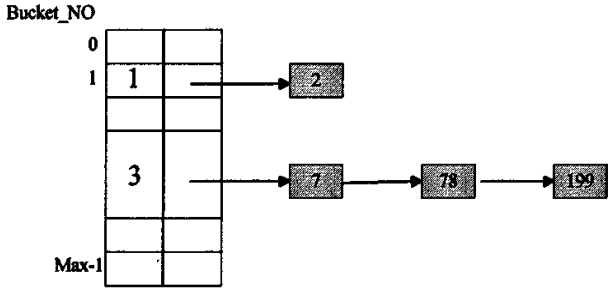


图 3-4 散列表

3，队列。如图 3-5 所示，用来组织整个爬虫系统的算法。系统在做遍历时不能利用深度搜索算法，因为这样很快就会导致堆栈溢出而导致系统崩溃。因此，为了很好的进行广度搜索算法，引入如下的列队。

图 3-5 队列

3.2.2.2 Hash 策略

系统采用了链式方法来解决碰撞问题[10]，因为网页数目的很难预计，所以系统避免了采用线性开址等系列 hash 表。对于任一网址的 URL，系统首先将其映射到 hash 表的桶号 i，然后判断此链中是否包含了此 URL。如果没有包含，就将 URL 对应于 url_table 中的序号放入到此桶号的冲突链表之中。

对于字符串 url 到桶号映射函数，系统如此设计[10]：

假设 URL ="$Z_0Z_1Z_2Z_3...Z_n$", hash 表的维数为 Max；

构造映射函数[10]：

$$F: url \rightarrow N,\ F(url) = (\sum_{i=0}^{n} Z_i * 256^i)\%Max, \qquad (3\text{-}1)$$

由于任意一个字符其值总是在 0-255 之间，因此系统选择了基数 256。更为合理的是，对于上述字符串 URL，系统可以采用如下算法[10]求解 *F* 的值：

```
F=0;
for(int i=n; i>=0; i++)
{
   F=(F<<8)+Zi;
}
F=F%Max;
```

因此，整个算法的时间复杂度就是 URL 的长度，非常合乎本系统的选择。

3.2.2.3 本地文件存储策略

本地文件存储是网络爬虫模块的重要一环，它有两个任务[10]：

1，保证文件名合乎本地文件的命名方式；

2，保证将所有的网页下载到本地之后能够按照原来网站访问方式在本地进行访问。

由于并不是所有的网页都具有标题，而且连接是按照 URL 的方式来组织的，因此系统采用了将 url 转化成文件名的方式。然而，URL 中往往包含像"？"，"/"这样不允许在文件名中出现的字符，因此产生了如下的转化规则[10]：

- (1) 对于"?"前面的 URL 部分将按照"/"的出现将文件名转化成本地目录名。例如：对于 URL="http://chinaren.com/news/index.asp?path=/a/b/c" 将建立目录 ./chinaren.com/news/ ；
- (2) 将"?"转换成"+"；
- (3) 将"?"后面的"/"转换成为"_"；

- (4) 将规则(1)中的剩余串经过规则(2),(3)转化之后作为文件名保存到本地；
- (5) 将文件内容中所有的 URL 连接按照上述转换方式进行修改并且保存。

按照这种策略，系统可以根据任意一个给定的 URL 很容易定位到本地的文件；而且浏览起来和从网上浏览方式完全没有什么改变[10]。

3.2.2.4 持久化机制

为了能够保证系统在重新启动之后继续工作，系统需要持久化几个数据结构的内容[10]。

1，URL 队列：为了让系统在再次启动之后不重新从初始连接开始，系统将 URL 队列中的内容保存到硬盘，当下次启动的时候将其加载到队列中作为本次运行的初始 URL 继续运行。

2，散列表：为了不重复的将同一 URL 的内容从网站上下载，系统将散列表的内容进行保存。对散列表的保存将按照桶号逐次保存，在保存本桶号数据时，先保存本桶号的项目数，如果为 0 那么下次加载的时候就不往本桶中填充数据。否则，将依次填充本桶数据。

3，url_table：主要需要维护其 URL 和序号之间的对应关系即可。在具体的系统实现之中，直接顺序保存 URL 即可，因为文件名称可以按照 3.2.2.3 的策略得到。

3.2.2.5 设计实现

图 3-6 给出了网络爬虫的主要类[10]。其中 Queue 就是用来控制操作的 URL 队列；Pager 是专门用来从网上下载指定 URL 地址内容的类，它还外带了从文件中解析新的 URL 连接功能；UrlTable 即是 3.2.2.1 中包含数据结构 url_table 的类；Hash_Url 是用来处理 Hash 表 PagerFilter 用来过滤掉某些不想处理的网页。

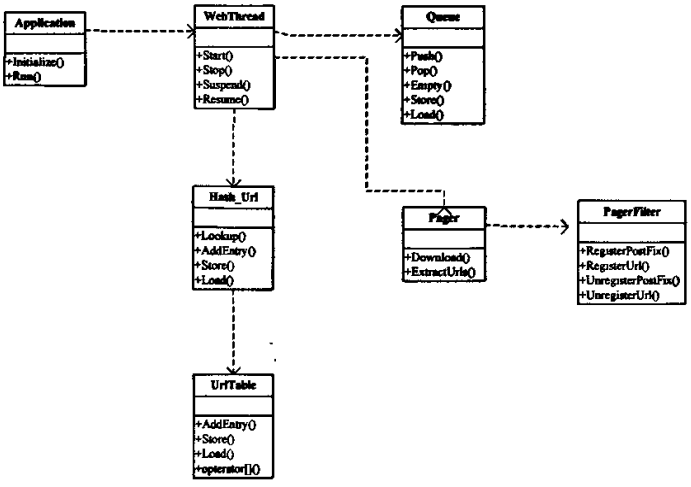


图 3-6 网络爬虫实现类图

3.3 数据清理与分词模块

3.3.1 数据清理

系统在这一模块解析网络爬虫模块得到的网页，摘除 HTML 的标记，在处理过的 HTML 页面基础上保存它的内容文本，超链接等信息。在得到 Web 信息各自相应的文档后，以标点符号为边界把文档且分成多个较短的字符串，并去掉其中多余的空格，变不规范格式为标准格式，为后面的分词做准备[8]。

在网页参加系统运算前，都要对其文本内容进行预处理，中文分词和词干提取是其中一个很重要的环节。英文是采用空格分开的语言，其分词相当的容易；但是系统仍然需要对其单复数以及时态做相应的处理，即所说的词干提取。而中文句子中每个词条之间没有固定的分隔符，为了对其进行聚类，首先就需要对其进行分词，即是在词条之间加上合适的分隔符，如空格或者其他标识符等[8]。

3.3.2 中文分词技术

本文采用的是层叠隐马模型的汉语词法分析[11]，该方法主要有 N-最短路径粗分[12]，未登录词的识别，排除歧义和词性标注等过程。本文将按照此步骤求解和设计的角度展开，具体原理参见[11,12]，本文不再赘述。

为了将文档用一系列的特征来描画，很自然的就是用词语来充当特征的角

色。但是，汉语是以字为基本的书写单位，词语之间没有明显的区分标记。因此，中文词语分词技术是整个信息处理的基础。目前一些主流的分词方法有最大匹配、最小切分等方法[10]。

3.3.2.1 中文分词模块

由于本系统不需要对词语做出词性标注，因此该子系统中去掉了词性标注过程。在中文分词子系统[10]中（如图 3-8[10]），N-最短路径粗切分快速的产生了 *N* 个最好的粗分结果，粗分结果集能够尽量的覆盖歧义切分。未登录词是指[10]在词典中没有进行收录的词语，例如人名、地名和机构名等等。基于类的隐马分词是[10]在未登录词识别之后进行的一次切分，在该切分中未登录词和普通词语一样参加竞争。二元切分词图[10]是个关键的中间数据结构，它在未登录词的识别、排歧和分词过程中起了举足轻重的作用。

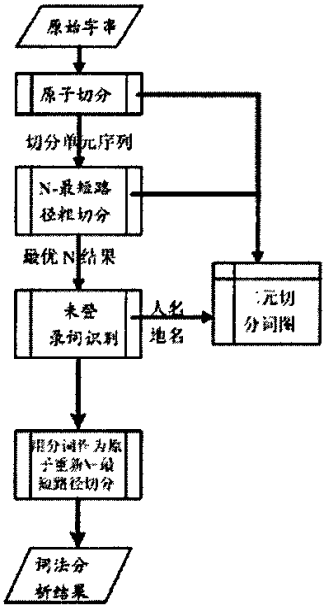


图 3-8 中文分词模块框架

3.3.2.2 N-最短路径粗切分

对于任何一个字符串，系统将其划分成单个的原子[10]。每个原子为单个汉字，标点或者由单字节、字符、数字组成的非汉字串。例如："2006.9，English 软件的自由源码开始发布"对应得分词原子序列为"2006.9/，/English/的/自/由/源/码/

开/始/发/布/”。

在得到原子序列后，系统根据每个原子建立有穷自动机，并且若状态 *A* 到状态 *B* 之间构成的词语存在于字典之中的话，那么就存在状态 *A* 到状态 *B* 的转换[10]。我们以“他说的确实在理”为例，图 3-9[10]给出了系统建立的自动机：

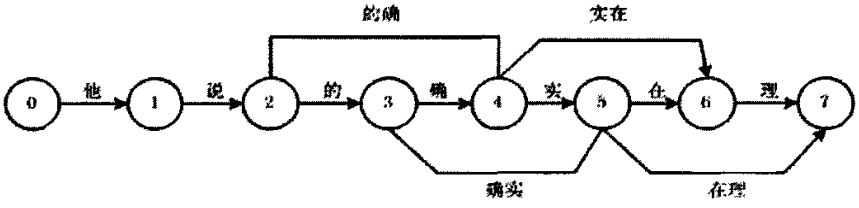


图 3-9 原始切分图

由于“的确”，“实在”，“确实”，“实在”都能够从字词中找到，因此存在图 3-9 中的状态转换[10]。

假设[10]整个汉字串为 *C*，分词结果为 *W*，那么系统需要寻找 *max*：*P(W|C)* 作为 *W* 的解。又：*P(W|C) = P(C|W)P(W)/P(C)*；其中，*P(C)*是汉字串的概率，它是一个常数。从分词结果 *W* 到字符串 *C* 的恢复也只有一种结果，因此 *P(C|W)*=1；因此，我们的目标就是确定 *P(W)*最大的 *N* 种切分集合。

假设[10]$W=w_0w_1w_2...w_m$ 是字符串 *C* 的一种切分，那么 $P(W)=w_0\prod_{i=1}^{m}P(W_i|W_{i-1}...W_0)$。为了处理方便，我们假定 $P(W_i|W_{i-1}...W_0)=P(W_i|W_{i-1})$，即第 *i* 个词语的划分只和第 *i*-1 个词语有关系。对其取负对数：$-\log(P(W))=-\log(w_0)+\sum_{i=1}^{m}-\log(p(W_i|W_{i-1}))$；

为了更加便于处理：系统将图 3-9[10]原始切分图转换成图 3-10[10]的二元切分图[10]。

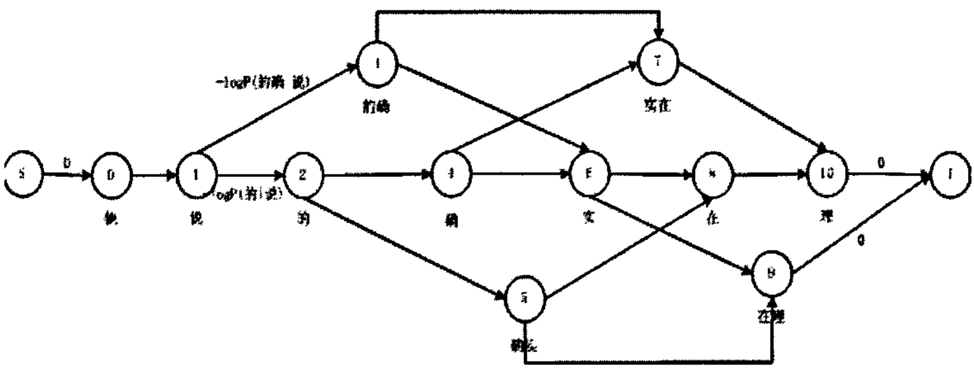


图 3-10 “他说的确实在理”的二元切分图

其中，除了初始状态 S 和终结状态 T 之外，其他的每个状态都代表着原始切分图的一条弧，即一个可能的单词。并且任何从初始状态到终结状态的一条路径都对应着对原汉字符的一种切分。在此二元切分图中，与初始状态或终结状态相连接的所有弧其权值均为 0，而其他状态之间的均为 $-logP(W_i|W_{i-1})$，因此，原问题的解就是从此二元切分图中找出从初始状态 S 到终结状态的 N 条最短路径[10]。

为了求解这个问题，系统在每个状态节点上保存一个队列[10]。这个队列中每项保存的内容为：从初始状态到当前状态的距离，当前距离所属的类别，到当前状态路径上的前驱状态节点以及前驱状态节点所属的类别。对距离所属的类别作如下解释：从初始状态 S 到达某状态节点 i 有很多个路径，每个路径都有一个路径长度，其中路径长度相同的路径就属于一个“距离所属的类别”。并且由于初始状态节点没有前驱节点，所以他不会保存任何队列信息。如图 3-11[10]所示(由于只有唯一一个到 T 结尾的转移，因此省略了节点 T 的队列信息，但在实际实现中，依然是存在状态 T 的信息队列的)，对于每个状态都有 1 个队列信息。如节点 4：它有 3 条从初始状态 S 到 4 的路径，其距离分别为 2，2，3；距离类别为 1，2；前驱节点有(2,1)、(3,1)、(3,2)；例如，为了获得一条距离为 2 的路径，那么首先找到项(2,1,(3,1)),由于其前驱为 3 并且是种类为 1 的条目；因此，在 3 中选择一个距离类别为 1 的项目，发现有（1，1，(1，1)）,于是又直接找到节点 1 中，并最终得到路径 S->1->3->4[10]；

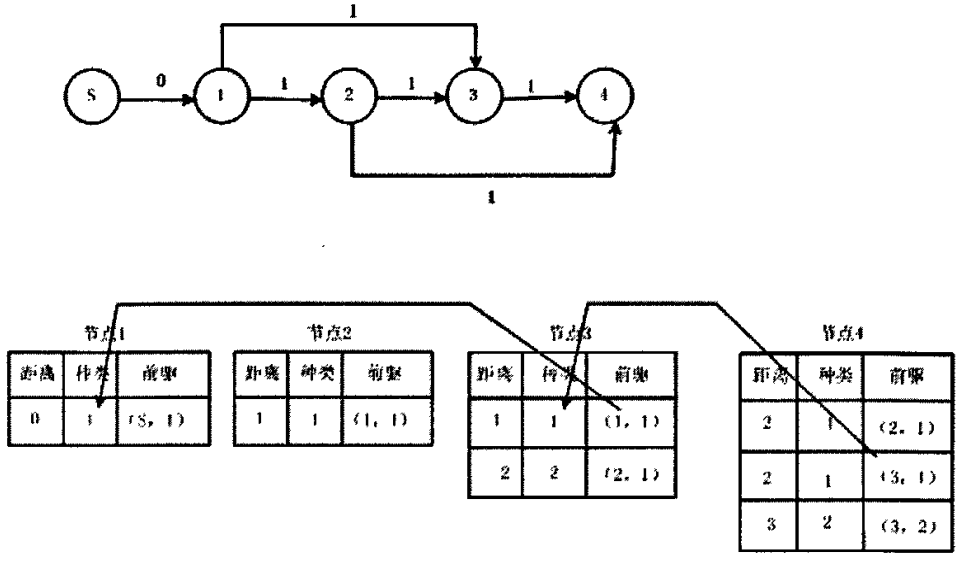


图 3-11 *N*-最短路径求解示意图

为了从二分图中得到从初始状态 *S* 到 *T* 的 *N* 条最短路径，系统只需要对节点 *T* 的所有节点按照距离排序，然后从中选出 *N* 个项目依次按照上述方法找到原状态节点 *S* 即可得解[10]。

3.3.2.3 未登录词的识别

未登录词[10]是指在核心字典中没有收录的词，由于这类词的存在大大的干扰了正常情况下的分词结果。例如，在汉字串“克林顿对内塔尼亚胡说”中，“内塔尼亚胡”是一个词典中没有收录的译名，实际切分的时候，“对”与“内”，“胡” 与“说”往往会粘在一起，最终导致错误的切分结果：“克林顿/对内/塔尼亚/胡说/”。怎样正确的将未登录词识别出来是本小结的主题。在进入讨论正题前，让我们来看看角色的概念。表 3-1 为人名识别的角色表[10]。

角色[10]就是词语在一个语境中所盼演的一种身份。对于一个给定的初始划分结果 $W=(w_0w_1w_2...w_n)$，在角色范畴内，假定 $R=(r_0r_1r_2...r_n)$为词语序列对应的角色序列，我们取概率最大的角色序列作为词语序列最终的角色。我们将词语看成是观测值，将角色看成是状态值，根据马尔科夫链可以得到[10]：

$$P(R)=\prod_{i=1}^{m}p(r_i\,|\,r_{i-1})p(w_i\,|\,r_i)\,, \qquad (3\text{-}2)$$

两边取负对数，得到

$$-\log P(R) = \sum_{i=1}^{m}(-\log p(r_i \mid r_{i-1}) \; - \; p(w_i \mid r_i)), \qquad (3\text{-}3)$$

通过后文描述的算法[10]，系统得到一个最大概率角色序列，并在此基础上通过某特定人名模板的匹配实现特定类型的未登录词识别。打个比方，如果通过后文算法得到词语序列 *W* 得到的角色序列是 *CDEAEF*，那么 *CDE* 所对应的词语就被组合到一起，因为中文双名的名字组合就是 *CDE*(*C*：中国人名的姓，*D*：双名的首字，*E*：双名的末字)[10]。

表 3-1 人名识别的角色表

角色	意义	示例
A	人名的上文	又/*来到*/于/洪/洋/的/家
B	人名的下文	新华社/记者/黄/文/*摄*
C	中国人名的姓	*张*/华/平/先生；*欧阳*/修
D	双名的首字	张/*华*/平/先生
E	双名的末字	张/华/*平*/先生
F	单名	张/*造*
G	人名的前缀	*老*/刘、*小*/李
H	人名的后缀	王/总、刘/老、肖/氏
L	译名的首部	*蒙*/帕/蒂/·/梅/拉/费
M	译名的中部	蒙/*帕*/*蒂*/·/*梅*/*拉*/费
N	译名的末部	蒙/帕/蒂/·/梅/拉/*费*
O	日本人名末部	小泉/纯/一/*郎*
X	连接词	邵/钧/林/*和*/稽/道/青/说
Z	其它	*人民*/*深切*/缅怀/邓/小/平

求解最大角色序列[10]：如图 3-12 所示，为了便于计算假设从 $r_{i-1,k}$ 到 $r_{i,j}$ 的弧长为 $(-\log p(r_i \mid r_{i-1}) \; - \; p(w_i \mid r_i))$，那么问题将转化为求 *S* 节点到终结点 T 的最短路径。假设 $D[i][j]$ 代表从 S 到第 i 个词语的第 j 种角色的最短路径，那么 $D[i][j] = \min_{k\in[1,n]}\{D[i-1][k] + (-\log p(r_{ij} \mid r_{i-1,k}) \; - \log p(w_i \mid r_{ij}))\}$，其中 n 是单词 W_{i-1} 存在的角色数。为此，本文采用动态规划算法来求解从 *S* 节点到 *T* 节点的最短路

径。为了记录 $D[i][j]$时的 k，系统引入路径记录矩阵 BestPrev[m][n],其中 m 是单词个数，n 是一个单词最多的角色个数[10]。

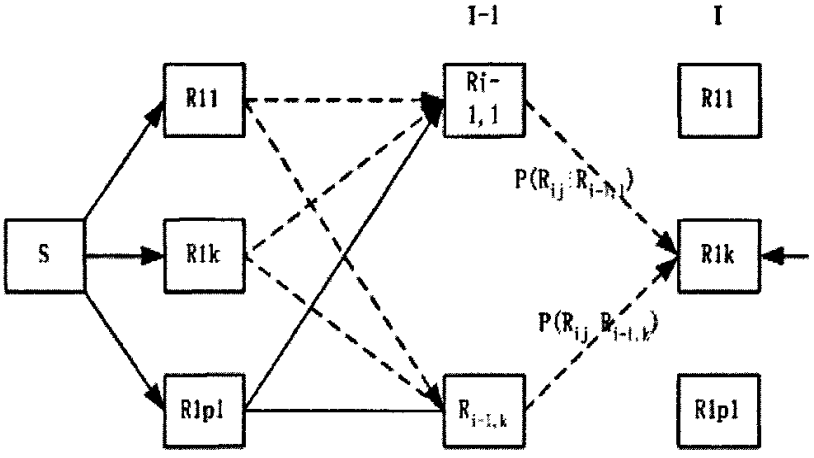


图 3-12 求最大序列示意图

【算法 3-2】[10]：求取从 S 到 T 最短路径时，填充矩阵 BestPrev

```
ShortestPath()
{
    1. For (int i=0; i<Words[0].nPos; i++)     //第一个单词的所有前驱为-1
        a) BestPrev[0][i]=-1;
    2. For (int i=1; i<nWords; i++)
        a) For (int j=0; j<Words[i].nPos; j++)     //对于每一个角色 R[i][j]
            i.  Temp = maxno.    //temp 相当于前文中的 D[i][j]
            ii. For (int k=0; k<Words[i-1].nPos; k++)
                1. If (D[i-1][k] + (-log p(r_ij|r_i-1,k) - logp(w_i | r_ij)) < temp)
                    a) Temp = D[i-1][k] + (-log p(r_ij|r_i-1,k) - logp(w_i | r_ij));
                    b) BestPrev[i][j] = k;
                2. D[i][j] = temp;
}
```

【算法 3-3】[10]：获取最佳角色，并将其保存在栈 PosStack 中

```
GetBestPos()
{
    Int k = BestPrev[nWords][ 0];//从终结点开始
    While (k!=-1) //第一个单词
    {
        PosStack.Push(Words[nWords-1].Pos[k]);   //压入角色
        K = BestPrev[nWords-1][k];
        nWords--;
    }
}
```

至此，一个给定单词划分的最佳角色就已经决定[10]。系统会拿这个角色串和特定未登录词的模式进行匹配，如果匹配就将其组合在一起作为一个新词识别出来[10]。

新未登录词的权重[10]：为了让新识别出来的未登录词和普通词一起参与下一轮的竞争，我们需要计算出其竞争的权重。具体计算方法是 $P(w|C) = \prod_{j=0}^{k} p(w_{p+j}|r_{p+j})$；其中 k 是识别出来的未登录词的长度，r_{p+j} 是上述求取的最优的角色值。

3.3.2.4 中文分词小结

系统在进行如图 3-13[10]的运算之后，需对整个 N 条路径进行排序，将结果最好的作为我们最终的结果。此外，系统会使用到核心词典，未登录词词典和角色频率统计词典。这些词典都是北京语言计算所所提供的语料库，是对《人民日报》进行语料分析形成的[13]。

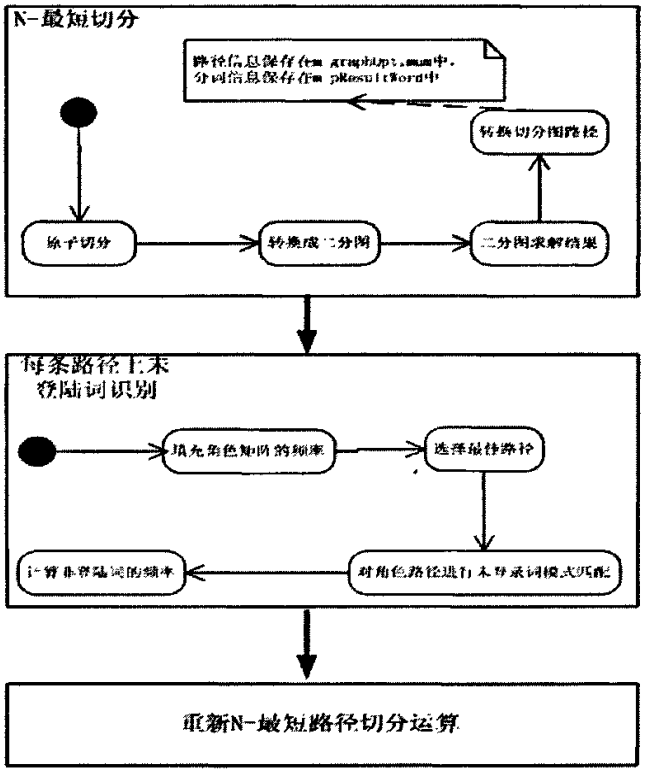


图 3-13 中文分词过程

在本文的系统中将采用北京语言计算所所提供的语料库作为中文分词的依据[10]。并合理修改了中科院汉语自动分词系统 ICTCLAS 源代码[11]，用以处理中文网页[10]。

3.3.3 英文 stemming 技术

在英文中存在大量的时态、语态和单复数形式。这些形式的多样化导致了英文词语数量的急剧膨胀。如何将由于时态、语态和单复数引起的词语还原到词根是本节的主题。形态分析常常采用基于自动机的规则方法，即将词形变化的规律总结成规则，然后通过自动机的方法对词形进行转换。转换的过程一般需要使用词典。但和中文相比，英文各个词语之间有空格分开，这种自然的特性导致不需要再对英文进行分词处理[10]。

本文采用的 Stemmer 是由 Martin Porter 提出的 Porter Stemmer[14]。这个 Stemmer 仅仅采用了一组规则，不需要词典，而且效率也非常的不错，本文将在此作一个简单的介绍[10]。

Porter Stemmer 算法[10]：

在具体介绍之前，我们需要定义一些概念[10]。

1，*m* 计数

假设 *C* 代表一系列辅音字母组成的字串，*V* 代表一系列元音字母组成的字串。因此，一个词语将由下列四种形式之一构成。

CVCV ... C

CVCV ... V

VCVC ... C

VCVC ... V

我们将其表示成为[*C*](*VC*){*m*}[*V*]。其中，[]表示内容可选，{*m*}表示前面内容的重复 *m* 次。如下：

m=0　TR, EE, TREE, Y, BY.

m=1　TROUBLE, OATS, TREES, IVY.

m=2　TROUBLES, PRIVATE, OATEN, ORRERY.

2，(condition) *S*1 -> *S*2

这个表达式表示一个如果一个词语的后缀为 *S*1,并且 *S*1 前面的词干满足给定的条件 condition,那么就用 *S2* 替换掉 *S1*。上述的条件 condition 往往通过前面的 *m* 来表示。例如，有表达式(*m* > 1) EMENT ->*null*，那么词语 REPLACEMENT 被替换成 REPLAC，其中 *m*=2。

这个表达式部分也可以表示成如下形式：

- **S* – 单词以 S 结束(S 是一个字母或者是一个字符串)；
- **v** - 词干中含有一个元音字符；
- **d* - 词干以两个相同的辅音字母结束；
- **o* - 词干以 cvc 的形式结束,并且第二个 *c* 不是 *W*，*X* 和 *Y*(*c* 是一个辅音字符，*v* 是一个元音字符)；
- condition 可以是用“and”，“or”和“not”对前面条件的组合。

Porter Stemmer 算法[14]实际上是对英文后缀的一些列替换工作，具体介绍如下：

(1)Step 1a 替换单复数；

SSES -> SS	caresses -> caress
IES -> I	ponies -> poni
	ties -> ti
SS -> SS	caress -> caress
S ->	cats -> cat

(2)Step 1b 替换分词形式，如(-ed, -ing)；

(m>0) EED -> EE	feed -> feed
	agreed -> agree
*(*v*) ED ->*	plastered -> plaster
	bled -> bled
*(*v*) ING ->*	motoring -> motor
	sing -> sing

如果上述第二条或者第三条转换成功的话，那么继续进行下述转换：

AT -> ATE	conflat(ed) -> conflate
BL -> BLE	troubl(ed) -> trouble
IZ -> IZE	siz(ed) -> size
*(*d and not (*L or *S or *Z))-> single letter*	
	hopp(ing) -> hop
	tann(ed) -> tan
	fall(ing) -> fall
	hiss(ing) -> hiss
	fizz(ed) -> fizz
*(m=1 and *o) -> E*	fail(ing) -> fail
	fil(ing) -> file

(3)Step 1c 将前面词干中含有元音的最后一个 y 改称 i

*(*v*) Y -> I*	happy -> happi
	sky -> sky

(4)Step 2 对一些常见的词根进行处理

(m>0) ATIONAL -> ATE	relational -> relate

(m>0) TIONAL	-> TION	conditional	-> condition
		rational	-> rational
(*m*>0) ENCI	-> ENCE	valenci	-> valence
(*m*>0) ANCI	-> ANCE	hesitanci	-> hesitance
(*m*>0) IZER	-> IZE	digitizer	-> digitize
(*m*>0) ABLI	-> ABLE	conformabli	-> conformable
(*m*>0) ALLI	-> AL	radicalli	-> radical
(*m*>0) ENTLI	-> ENT	differentli	-> different
(*m*>0) ELI	-> E	vileli	-> vile
(*m*>0) OUSLI	-> OUS	analogousli	-> analogous
(*m*>0) IZATION	-> IZE	vietnamization	-> vietnamize
(*m*>0) ATION	-> ATE	predication	-> predicate
(*m*>0) ATOR	-> ATE	operator	-> operate
(*m*>0) ALISM	-> AL	feudalism	-> feudal
(*m*>0) IVENESS	-> IVE	decisiveness	-> decisive
(*m*>0) FULNESS	-> FUL	hopefulness	-> hopeful
(*m*>0) OUSNESS	-> OUS	callousness	-> callous
(*m*>0) ALITI	-> AL	formaliti	-> formal
(*m*>0) IVITI	-> IVE	sensitiviti	-> sensitive
(*m*>0) BILITI	-> BLE	sensibiliti	-> sensible

(5)Step 3 做一些常见形容词和名词之间的后缀转换

(*m*>0) *ICATE*	-> *IC*	triplicate	-> triplic
(*m*>0) *ATIVE*	->	formative	-> form
(*m*>0)*ALIZE*	-> *AL*	formalize	-> formal
(*m*>0) *ICITI*	-> *IC*	electriciti	-> electric
(*m*>0) *ICAL*	-> *IC*	electrical	-> electric
(*m*>0)*FUL*	->	hopeful	-> hope
(*m*>0)*NESS*	->	goodness	-> good

(6)Step 4 作一些名词、形容词到动词的转换

(*m*>1) *AL* -> revival -> reviv

(*m*>1) *ANCE* -> allowance -> allow

(*m*>1) *ENCE* -> inference -> infer

(*m*>1) *ER* -> airliner -> airlin

(*m*>1) *IC* -> gyroscopic -> gyroscop

(*m*>1) *ABLE* -> adjustable -> adjust

(*m*>1) *IBLE* -> defensible -> defens

(*m*>1) *ANT* -> irritant -> irrit

(*m*>1) *EMENT* -> replacement -> replac

(*m*>1) *MENT* -> adjustment -> adjust

(*m*>1) *ENT* -> dependent -> depend

(*m*>1 and (**S* or **T*)) ION -> adoption -> adopt

(*m*>1) *OU* -> homologou -> homolog

(*m*>1) *ISM* -> communism -> commun

(*m*>1) *ATE* -> activate -> activ

(*m*>1) *ITI* -> angulariti -> angular

(*m*>1) *OUS* -> homologous -> homolog

(*m*>1) *IVE* -> effective -> effect

(*m*>1) *IZE* -> bowdlerize -> bowdler

(7)Step 5 继续做出一些小的修整

(*m*>1) *E* -> probate -> probat

rate -> rate

(*m*=1 and not *o) *E* -> cease -> ceas

(*m* > 1 and *d and *L) -> *single letter*

controll -> control

roll -> roll

这个算法在许多信息检索系统中得到广泛的应用，其不仅不需要词典而且速

度较快，被公认为是一种非常好的算法。在此算法中系统用 m 来衡量一个词语的长短。如果一个词语太短，那么一般不进行后缀替换工作。至此，整个算法已经介绍完毕[10]。

3.4 特征项的选择模块

特征项选择和提取是整个 Web 文档聚类系统的基础，它将系统从概念空间映射到可运算空间，从而使整个系统实现成为可能。

3.4.1 向量空间模型(VSM: vector space model)

G. Salton 等人于 60 年代末提出了向量空间模型 VSM (vector space model)的概念，即使用向量表示文本或页面[1]。

向量空间模型的基本概念可以描述如下[1]：

(1)文档：指一般的文本或文本的片段（段落、句群或句子），一般指一篇文章。尽管文档可以是多媒体对象，但在我们的讨论中假设为文本对象，并且对文本和文档不加以区别。

(2)项（特征项）：文本的内容由一些特征项来表达，一般由文本所含有的基本语言单位（字、词、词组或短语等）来表示，即文本可以表示为 $Document = D(t_1,t_2..t_n)$，其中 t_i 表示各个项。换句话说，由这些项形成了一个向量空间，每个项表示一个维度。

(3)项的权重：在文本中，每个特征项都被赋予一个权重 W，以表示这个特征项在该文本中的重要程度。权重一般都是以特征项的频率为基础进行计算的。

(4) 相似度度量：两个文本 d_1 和 d_2 之间的相关程度常常用它们的相似度 $Sim(d_1,d_2)$来度量。在向量空间模型下，我们可以借助向量之间的某种距离来表示文本间的相似度。

(5)向量空间模型（VSM）：给定一个自然语言文本，由于在文本中既可以重复出现又应该有先后次序的关系，分析起来仍有一定的难度。为了简化分析，可以暂不考虑在文本中的先后次序并要求互异（即没有重复）。在舍弃了各个特征项之间的顺序信息之后，一个文本就表示成一个向量，也就是特征项空间中的一个点；而一个文本集就可以表示成一个矩阵，也就是特征项空间中的一些

点的集合。

3.4.2 特征项的选择

文本挖掘的特征属性是灵活的、多变的，这与数据挖掘使用固定的属性特征是不同。抽象概念难于表示、难于形式化，文本特征往往是高维的。根据 Dunja Mladenic 和 Marko Grobelnik 在著名的搜索引擎 YAHOO 上的实验结果表明，对全体文档集进行特征表示时，其维数将高达 69,280-255,602 维。而另一方面文档的许多信息又是高冗余的，所以文本特征的提取(缩减)是相当重要的，这往往决定了文本挖掘的效率[3]。

对目标表示中词条 t 的选取被称为特征提取。主要有两大类方法：独立评估方法和综合评估方法，前者的基本思想是对特征集中的每个特征进行独立的评估，让每个特征都获得一个权值，然后按权值大小排序，根据权值或预定的特征数目选取最佳特征子集作为特征提取的结果。后者则是从高维的、彼此间不独立的原始特征集中找出较少的描述这些特征的综合指标，且这些综合指标之间相互独立，然后又用得到的综合指标对特征集进行特征选择[3]。

向量空间模型（VSM）表达效果的优劣直接依赖于特征项的选取，以及权重的计算，在此简单地介绍一下[3]。

特征项的选择有以下三个原则[3]：

（1） 应当选取那些包含语义信息较多的，对文本的表示能力较强的语言单位作为特征项；

（2） 这种选取的过程本身应当比较容易实现，其时间和空间开销都不应过大。

（3） 文本在这些特征项上的分布应当有比较明显的统计规律性；

词特征：词特征对文本的表示能力比较好，词汇能够比较完整地表达语义信息。然而，并不是所有词都适合作为特征项，研究表明，高频词和低频词对文本的表示左右均小于中频词。因为高频词在所有文章中都有相近的较高频率；低频词在文本中出现次数少，不适合采用统计方法来处理；而中频词和文本表达的主题比较相关，表示能力最强。

字特征：字对文本的表示能力相对于词特征比较差，不能完整地独立地表达

语义信息。使用字特征的特征抽取过程比较简单，而且由于常用的汉字数目很少，因此抽取过程的时间和空间开销都不会太大。

对于文本中常常出现一些没有实在意义的虚词、助词等等，比如“的”，“地”，“得”，这些词出现次数很多，然而对于有无实在的意义。常用的方法是建立停用词表，统计词频时过滤掉这些词，或者采用属性标注的方法，直接过滤掉所有的虚词[1]。

3.5 权重计算模块

通过特征提取的处理以后，用抽取的词作为向量的维数来表示文本，最初的向量表示完全是 0、1 形式，即，如果文本中出现了该词，那么文本向量的该维为 1，否则为 0。这种方法无法体现这个词在文本中的作用程度，所以逐渐 0、1 被更精确的权重替代[1]。

向量 $d=(w_1,w_2,...,w_m)$ 表示文档 d 的特征词条及相应权重。其中：m 为文档集中词条的数目，$w_i(i=1,...,m)$ 表示词条 t_i 在文档 d 中的权重。特征权重 w_i 的计算通常采用经典的 TF * IDF[10]算法，并进行规格化处理[18]：

$$w_i=\frac{TF\times\log_2(N/DF_i)}{\sqrt{\sum_{i=1}^{m}[TF\times\log_2(N/DF_i)]^2}}\ , \qquad (3\text{-}4)$$

其中：TF 表示该词条 t_i 在文档 d 中的频数，DF_i 表示文档集中包含词条 t_i 的文档数，N 表示文档集中的文档数。

3.6 基于改进 K-means 聚类分析模块的简介

经过前面的网络爬虫、数据清理、中文分词、特征提取、权值计算等模块的处理，系统已将最初由网络爬虫下载下来的网页转化成了可进行运算聚类运算的向量。

基于改进的 K-means 聚类分析模块就是将可进行聚类运算的向量进行聚类分析。具体内容我们在下一章详细介绍。

第四章 K-means算法和改进的K-means聚类算法

4.1 原始的K-means算法

K-means算法首先由J. MacQueen在1967年提出[6]，这是一种应用非常广泛的聚类算法，广泛应用在数据挖掘领域中。在K-means算法中，每个类用该类中对象的平均值来表示，故此得名K-means算法。

K-means算法的输入为聚类个数 k，以及包含 n 个数据对象的数据库。输出为满足评价函数最小的 k 个聚类。

K-means算法流程[6]如下：

（1） 从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

（2） 根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的相似度；并根据最大相似度（最小距离）重新对相应对象进行划分；

（3） 重新计算每个（有变化）聚类的均值（中心对象）；

（4） 循环（2）到（3）直到每个聚类不再发生变化为止。

K-means 算法的工作过程说明如下。首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心。而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类。然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）。不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

K-means算法的主要优点有：

（1）算法简单；

（2）处理速度快；

（3）可伸缩性好；

（4）效率高；

（5）适合于处理大数据集和大文档集；

（6）这个算法试图找出使评价函数值最小的 k 个类。当结果类是密集的，而类与类之间区别明显时，它的聚类效果较好；

（7）中止条件明确；

（8）适应动态数据。

然而，这种算法也存在一些主要的缺点和不足，比如：

（1）受初始聚类中心选择的影响较大。比如，对不同的 k 值可能会导致不同的聚类结果；如果初始聚类中心选择得不好，聚类结果可能会陷入局部最优解，而得不到较好的聚类效果。

（2）对“噪声”和孤立点很敏感，因为少量的该类数据能对平均值产生极大的影响。

目前针对 K-means 算法的一些主要的缺点和不足，已经提出了很多改进算法。一般都是基于以下几个方面进行的：

（1）初始聚类中心的选择；

（2）相似度的计算；

（3）聚类平均值的计算。

本文采用 K-means 聚类算法主要基于以下原因：

（1）文档聚类普遍采用的算法是 K-means 算法[15,16,17]。

（2）K-means 聚类算法适合于处理大文档集，对处理大数据集，该算法是相对可伸缩的和高效的。并且算法简单，易实现，效率高。

（3）虽然存在一些缺点和不足，但是已经提出了很多有针对性的改进算法。

在本章下面的三个小节中，本文针对 K-means 聚类算法的主要缺点和不足，对 K-means 算法中的关键环节：相似度计算公式，初始聚类中心的选择和新聚类中心的计算方法进行了改进。

首先，对K-means聚类算法中，使用的变量和名词作以说明：

k：聚类的个数；

n：需要进行聚类的对象（文档）的总数；

m：文档集中词条的个数；

t_i（i=1...m）：提取出来特征词条

w_i（i=1...m）：词条t_i在文档d中的权重；

对象=文档（本文的聚类对象就是Web文档）；

c_i（i=1...k）：聚类划分出类别；

s_i（i=1...k）：k个聚类的类中心；

在之前介绍的K-means算法流程中，我们可以看到有三点非常重要：

1，相似度的计算；

2，初始聚类中心的选择；

3，新聚类中心的计算。

在接下来的三个小节中，本文将针对以上三点进行改进。

4.2 相似性度量的改进

在进行相似性度量的改进之前，首先要对权重评价函数进行改进。因为相似性的度量主要是对各个向量的权重值进行计算。

向量空间模型(VSM)是K-means算法中文档的表示模型，其中的词条权重评价函数用TF*IDF表示。然而实际上这种表示方法没有考虑对于该词条在文档中出现的位置及不同位置对文档内容的决定程度不同这一情况，只体现了该词条是否出现以及出现多少次的信息[18]。

向量空间模型(VSM)将 Web 文档分解为由词条特征构成的向量，利用特征词条及其权重表示文档信息[18]。向量$d=(w_1,w_2,...,w_m)$表示文档 d 的特征词条及相应权重。其中：m 为文档集中词条的数目，$w_i(i=1,...,m)$表示词条 t_i在文档 d 中的权重。特征权重 w_i的计算通常采用经典的 TF * IDF[10]算法，并进行规格化处理[18]：

$$w_i=\frac{TF\times\log_2(N/DF_i)}{\sqrt{\sum_{i=1}^{m}[TF\times\log_2(N/DF_i)]^2}}\ , \qquad (4\text{-}1)$$

其中：TF 表示该词条 t_i在文档 d 中的频数，DF_i表示文档集中包含词条 t_i的文档数，N表示文档集中的文档数。从公式可以看出，这种特征权重的计算方法是把文档当作一组无序词条，词条特征权重只是体现了该词条是否出现以及出现次数多少的信息，而对于词条在文档中的不同位置对文档内容的决定程度不同这一问题却未加考虑。

Web 文档有其自身的特性：可扩展标识语言（XML）已经成为 Web 上新一代数据内容描述标准，因此 Web 上的文档聚类应体现 XML 文档的特性。XML 文档中的基本单位是元素(element)。元素由起始标签、元素的文本内容和结束标

签组成。它的语法格式为：<标签>文本内容</标签> 。因此本文采用加入了词条隶属度的权重评价函数[18]。

基于 XML 的 Web 文档中，用户把要描述的数据对象放在起始标签和结束标签之间，无论文本的内容多长或者多么复杂，XML 都可以通过元素的嵌套进行处理。不同标签下，同一个词条也可能有不同含义。由此可见，XML 文档中不同位置的词条对文档内容的决定程度会有很大的不同[18]。

通常，一个文档的标题、关键词、摘要以及段首和段尾出现的词条对整个文档内容有很大的决定作用[18]。在 XML 文档中，通过标签可以得出词条对文档内容的决定程度，但很难对这种决定程度进行准确的定义。因此，本文利用模糊集理论，根据 XML 文档特性计算词条从属关系系数，并且将其量化为介于 0 和 1 之间的隶属度，加入到原有权重评价函数，从而表明 XML 文档具有该词条特征的程度。为了简化计算，词条在文档中出现的位置主要分为标题、摘要、关键词、段首尾、特殊标识处和正文几个部分。其相应权重为 σ_i，在[0,1]之间取值，用 l_i 表示词条在相应位置出现的次数。加入了词条隶属度的权重评价函数[18]为：

$$w_i' = w_i \times \frac{\sum l_i \sigma_i}{\sum l_i}, \qquad (4\text{-}2)$$

其中 w_i 是原权重评价函数。如公式所示；w'_i 取值在[0,1]之间，表示该词条特征 w_i 隶属于文档的程度或对文档内容的贡献程度。由此一个文档的特征向量可以用 $d = (w'_1, w'_2, ..., w'_m)$ 表示。由于改进后的权重评价函数 w'_i 比原权重评价函数 w_i 增加了 $\frac{\sum l_i \cdot \sigma_i}{\sum l_i}$ 部分，所以使计算复杂度略有增加。

改进了权重函数评价函数之后，接下来我们对相似性度量进行改进。

聚类中必须对文本进行相似度度量。不同的相似度度量方法会出现不同的聚类结果。在聚类分析中，要区分三个相似度度量（或者距离度量）[1]：

1，文本与文本之间的相似度度量；

2，文本与文本类之间的相似度度量；

3，文本类与文本类之间的相似度度量。

不同的聚类算法应用了这三种类型的度量方法中的某种或者几种。比如凝聚式层次聚类算法(HAC: hierarchical agglomerative clustering)中使用了(1)、(2)、(3)

三种。这三种距离中不同的定义方式会对结果产生不同的影响，衍生出算法的不同变种。（相似度和距离是两个相对的概念，聚类要求相似度大的聚到一起往往也就是要求距离小的聚到一起，在聚类分析中这两个概念经常一起使用，但其意义是刚好相反的。）[1]

K-means算法使用第1种基于距离的相似性度量，然而文档的特征向量一般超过万维，有时可达到数十万维，这种高维度使得这种度量方法不再有效。

利用向量空间模型处理Web文档时，由于文档的繁杂性，表示文档的特征向量可以达到数万维，甚至更多。通过预处理阶段停用词和无用高频词的过滤后，特征向量的维数虽然显著减少，但剩余的维数仍然很多[18]。

高维的特征向量使得聚类算法的处理时间大大增加，同时对算法的准确性产生不利影响，并且这些特征对于聚类来说大多是无用的，例如聚类算法（STC：suffix tree clustering)将特征向量的维数减少到几十维仍然能够准确聚类[18]。这主要是因为，对于非结构化的文档，体现其类别特点的特征词有很多，当进行某一方面的聚类时，与此无关的特征词就成了噪音。从这一点来说，文中前面改进的权重评价函数 w'_i 体现了特征词对文档内容的贡献程度，从而突出了与聚类相关的特征词，降低了无关特征词的干扰。另一方面，过多的特征词使得特定的特征词出现的频率较低，容易被噪音所淹没[18]。

已经有实验[19]验证了向量余弦距离比欧氏距离方法更适合于文本相似度的计算，所以本文采用向量余弦距离计算相似度。

采用的是余弦函数计算公式为[18]：

$$sim(d_i,d_j)=\frac{w_{i1}\times w_{j1}+\ldots+w_{im}\times w_{jm}}{\sqrt{\sum_{k=1}^{m}w^2_{ik}}\times\sqrt{\sum_{k=1}^{m}w^2_{jk}}},\qquad(4\text{-}3)$$

其中 $(w_{i1},w_{i2},\ldots,w_{im})$ 为 d_i 文档的特征向量，$(w_{j1},w_{j2},\ldots,w_{jm})$ 为文档 d_j 的特征向量。

特征向量过高的维度使得通常基于距离的相似性度量不再有效，而且由于权重评价函数的值在[0,1]之间，利用上式计算相似性使得计算复杂，容易引起较大的计算误差。因此本文采用一种新的计算文档相似性的方法[18]。

利用向量空间模型对文档进行聚类只能根据文档的二种信息[18]：

(1) 文档的长度；

(2) 文档中每个特征词出现的频率。

由于文档长度与文档所属的类别之间的关系不大，因此可以把所有的文档长度进行归一化处理，从而使文档向量具有统一的特征维数 m。这样每一个利用 VSM 的文档可以看成一个离散无记忆的信息源，使用信息论中的汉明距离可以描述二个信息源之间的距离[18]。例如，二个 m 位的码字：$x=(x_1,x_2,...,x_m)$，$y=(y_1,y_2,...,y_m)$，它们之间的汉明距离为：$D(x,y)=x_1\oplus y_1+...+x_m\oplus y_m$。其中 x_i 和 y_i 的值为 0 或 1，$\oplus$ 表示模 2 运算。汉明距离 $D(x,y)$ 的值表示二个码字在相同位置上不同符号的个数之和，从而反映出二个码字之间的差异，为码字之间的相似性度量提供了依据[18]。

在文档聚类中运用这一概念，将文档特征向量 $d=(w'_1,w'_2,...,w'_m)$ 用一个 n 位的码字来表示[18]。当 $n=m$ 时，文本特征向量的每一位用 0 或 1 表示，0 表示文档在此位置上没有特征词，1 表示此位置上有相应的特征词。同样，这种方法不能反映特征词对文档内容的贡献程度。因此本文采用 $n=4m$，即特征向量的每一维用码字的四位表示。根据已经归一化的 w'_i，在[0, 1]中利用四位码字表示，即将[0,1]区间分为 $2^4=16$ 份。码字值的大小表示相应特征词对文档内容的贡献程度，0000 表示此特征词对文档的内容贡献很小，1111 表示此特征词对文档内容贡献很大。其他码字同理[18]。

根据以上的分析，定义文档 $d_i=(w'_{i1},w'_{i2},...,w'_{im})$ 和 $d_j=(w'_{j1},w'_{j2},...,w'_{jm})$ 的相似性计算公式[18]为：

$$sim(d_i,d_j)=1-\frac{\sum_{k=1}^{m}\alpha_k}{(2^4-1)\times m}, \qquad (4\text{-}4)$$

其中：m 为特征向量维数，α_k 为二个文档对应特征词条的四位码字的十进制数值差的绝对值。可以利用简单的示例验证公式的合理性。当二个文档完全相似时，$sim(d_i,d_j)$ 的值等于 1，而二个文档完全不同时它的值为 0。这种方法不仅反应了文档之间的差异，而且定量地描述了这种差异性，从而为文档的聚类提供了依据[18]。由于这种相似性的计算使用的是整数，所以计算速度得到一定的提

高。

4.3 初始聚类中心选择的改进

如4.1中所述，K-means算法的聚类结果受初始聚类中心的选择影响较大，如果初始聚类中心选取不当，可能会产生局部最优解，而不是全局最优解。从不同的初始聚类中心出发会得到不同的聚类结果且准确率也不一样。通常K-means算法是从n个数据对象中随机选择k个对象作为初始聚类中心，这样就使得产生的聚类结果具有很大不确定性。如何选择初始聚类中心点成了影响最后聚类结果的重要因素[20]。

本文采用了改进的选择初始聚类中心的方法[20]，尽量使最初的初始聚类中心在空间分布上与数据实际的分布相一致。

在K-means算法中，选择不同的初始聚类中心会产生不同的聚类结果且有不同的准确率。此处改进的目的是如何找到与数据在空间分布上相一致的初始聚类中心。对数据进行划分，最根本的目的是使得一个聚类中的对象是相似的，而不同聚类中的对象是不相似的，即相似度大的对象聚在一起。如果能够寻找到k个初始聚类中心，它们分别代表了相似度较大的数据集合，那么就找到了与数据在空间分布上相一致的初始聚类中心[20]。

假设有一个2维数据集，包含有10个样本。假设要把它们划分为两类，按照上面的思想寻找初始聚类中心。a、b之间的相似度最大，那么选择a、b构成一个样本集A_1，并将它们从总的集合U中删除。U中与A_1相邻最近的点是c，这样便将c加入集合A_1并将它从U中删除。如果规定每个样本集中样本最大个数为4，则A_1中将会再添加样本d。然后在U中再找出相互之间相似度最大的两个点g、h构成A_2，并将它们从U中删除。U中与A_2相邻最近的点是i，这样便将i加入A_2并将它从U中删除，同样j也会并入A_2。最后，将这两个样本集分别进行算术平均，形成两个初始聚类中心。这样得到的初始聚类中心与实际样本的分布更加相符，从而可以得到更好的划分效果[20]。

如前所述，已经有实验[19]验证了向量余弦距离比欧氏距离方法更适合于文本相似度的计算。本文采用的余弦函数计算公式[18]为：

$$sim(d_i,d_j)=\frac{w_{i1}\times w_{j1}+\ldots+w_{im}\times w_{jm}}{\sqrt{\sum_{k=1}^{m}w^2_{ik}}\times\sqrt{\sum_{k=1}^{m}w^2_{jk}}},\qquad(4\text{-}5)$$

其中：$(w_{i1},w_{i2},\ldots,w_{im})$为$d_i$文档的特征向量，$(w_{j1},w_{j2},\ldots,w_{jm})$为文档$d_j$的特征向量。一个文档与一个文档集的相似度定义为这个文档点与这个文档集中所有文档点当中最大的相似度值。则一个文档点d_i和一个文档集合V之间的距离定义如下：

$$sim(d_i,V)=\min(sim(d_i,d_j),d_j\in V),\qquad(4\text{-}6)$$

假设文档集U有n个文档，将其聚为k类，m的初值为1。

改进算法[20]描述如下：

（1）计算任意两个文档的相似度$sim(d_i,d_j)$，找到集合U中相似度最大的两个文档，形成集合$A_m=(1\le m\le k)$，从集合U中删除这两个点；

（2）在U中找到与集合A_m相似度最大的文档，将其加入集合A_m并从集合U中删除；

（3）重复第（2）步直到集合中的文档个数大于等于$\alpha\cdot\frac{n}{k}(0<\alpha\le 1)$；

（4）如果$m<k,m++$，再从集合U中找到相似度最大的两个文档，形成新的集合$A_m=(1\le m\le k)$并从集合U中删除这两个文档，返回第（2）步执行；

（5）将最终形成的k个集合中的文档分别进行算术平均，从而形成k个初始聚类中心；

（6）从这k个初始聚类中心出发，应用K-means聚类算法形成最终聚类。

α的取值因实验数据不同而有所不同。α的取值过小则可能使几个初始聚类中心点在同一区域得到，α的取值过大则可能使初始聚类中心点偏离密集区域。从实验的情况来看取0.75时效果是比较好的[20]。

4.4 新聚类中心计算方法的改进

在用 K-means 算法进行聚类实验时，我们发现实验结果的稳定性方面存在问题，实验结果偶尔出现较大的偏差。通过分析表明，导致这种偏差产生的原因在于数据的分散性。因为在实际应用中，文档集通常是分散的，含有各种孤立点。

K-means 算法对于文档集中孤立点很敏感，少量的这类文档将对聚类结果产生较大的影响。尽管同样基于划分的 K-medoids 聚类算法对存在孤立点的文档集能得到较好的聚类结果，但其效率很低，执行代价很高[21]。为此，本文采用了 K-means 算法的一种改进算法[21]，能高效率地处理文档集中存在孤立点的情况。

本文采用将聚类均值点与聚类中心相分离的思想[21]。在原始 K-means 算法中，每一轮直接用类中所有对象的均值点作为该类的聚类中心，而在改进算法中，在进行第 k 轮聚类中心的计算时，采用类中那些与第 *k-1* 轮聚类中心相似度较大的数据，计算它们的均值点(几何中心点)作为第 *k* 轮聚类的中心。将聚类均值点（类中所有数据的几何中心点）作为新的聚类中心进行新一轮聚类计算，此时新的聚类中心将偏离真正的数据密集区[21]。

具体计算方法[21]如下：

(l)对于第 *k*-1 轮聚类获得的类 $c_{i(k-1)}$，计算类中数据与该类聚类中心 $s_{i(k-1)}$相似度最小的相似度 $sim\_min_{i(k-1)}$；

(2)选择类 $c_{i(k-1)}$中与聚类中心 $s_{i(k-1)}$相似度大于$1-\beta*(1-sim_\min_{i(k-1)})$的数据，其中 β 为 0 到 1 之间的常数，记该数据集合为 $cn_{i(k-1)}$；

(3)计算 $cn_{i(k-1)}$中数据的均值点，作为第 *k* 轮聚类的聚类中心。

对处理大文档集，该算法跟改进前 K-means 算法一样是相对可伸缩的和高效的，其时间复杂度为侧 O（nkt），其中，n 为所有对象数目，k 为类的数目，t 为算法迭代次数，通常 $k<<n$，$t<<n$[21]。

第五章　基于改进 K-means 算法的 Web 文档聚类系统的实验及聚类效果评价

5.1 聚类效果的评价方法

目前有关文本聚类分析结果的评价机制比较混乱，尚无得到普遍认可的科学评价机制及与之相关的深入研究，并且缺乏对目前大量存在的聚类算法进行深入分析与比较研究。在一般的关于文本聚类分析的论文中，众多的指标混杂，在定义上有些意义相同但名称却不同，比如类的准确率与类的纯度的定义。各个指标与其他指标相比有哪些特性和适用场合也没有得到很好的论述[8]。

聚类分析结果的评价方法可以分为外部和内部两种，简而言之，有数据集先验知识的是外部评价方法，没有先验知识的是内部评价方法[22]。

本文采用一种常用的外部评价方法 F-measure[23]。其组合了信息检索中查准率(precision)与查全率(recall)的思想。一个聚类 j(即一个类)及与此相关的分类 i(即原始的一类)的 precision=p 与 recall=r 定义为：

$$p(i,j)=\frac{N_{i,j}}{N_i}, \tag{5-1}$$

$$r(i,j)=\frac{N_{ij}}{N_j}, \tag{5-2}$$

其中 N_{ij} 是在聚类 j 中分类 i 的数目，N_j 是聚类 j 中所有对象的数目，N_i 是分类 i 中所有对象的数目。则分类 i 的 F-measure 定义为：

$$F(i)=\frac{2pr}{p+r}, \tag{5-3}$$

分类 i 对应的聚类可能有多个，哪个聚类的 F-measure 值高，就认为该聚类代表分类 i 的映射，分类 i 的 F-measure。就取该最大值。对一个聚类结果 k。它的总 F-measure 值由每个分类的 F-measure 加权平均得到：

$$F_k=\frac{\sum_i(N_i \cdot F(i))}{\sum_i N_i}, \tag{5-4}$$

其中 N_i 仍然为分类 i 中所有对象的数目。本文测试聚类算法的性能时，全部采

用 F-measure 值。

5.2 基于改进 K-means 算法的 Web 文档聚类系统的实验及聚类效果评价

本文通过网络爬虫下载了一批 Web 文档，并对这些文档进行了人工分类。分类结果如表 5-1 所示。

表 5-1 实验数据

数据集	文档个数	初始类别	类别名称
D1	163	8	旅游，通信，政治，影视，金融，交通，能源，医疗
D2	258	5	通信，政治，影视，金融，交通
D3	164	3	通信，交通，能源
D4	167	8	旅游，通信，政治，影视，金融，交通，能源，农业
D5	225	7	政治，影视，金融，交通，能源，农业，保险
D6	176	6	能源，农业，保险，政治，影视，金融
D7	269	8	金融，交通，能源，农业，旅游，通信，政治，影视
D8	158	8	影视，金融，交通，能源，农业，医疗，旅游，通信

表 5-2 未改进的 K-means 算法的 F-measure 值

类别	未改进的 K-means 算法的 F-measure 值								
	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	平均值
D1	0.618	0.676	0.732	0.603	0.633	0.672	0.667	0.653	0.657
D2	0.782	0.736	0.532	0.609	0.612	0.562	0.562	0.593	0.623
D3	0.658	0.675	0.712	0.623	0.676	0.672	0.766	0.578	0.670
D4	0.768	0.771	0.752	0.503	0.643	0.572	0.662	0.503	0.646
D5	0.608	0.772	0.662	0.605	0.636	0.672	0.664	0.553	0.647
D6	0.733	0.766	0.695	0.663	0.618	0.572	0.569	0.803	0.677
D7	0.623	0.696	0.682	0.507	0.538	0.532	0.761	0.703	0.630
D8	0.758	0.736	0.637	0.793	0.738	0.579	0.862	0.603	0.713

为了检验算法的准确程度，先用未改进的 K-means 算法对数据进行聚类。

随机地运行了 8 次，每次都是随机的选取初始聚类中心。8 次聚类的 F-measure 值和平均值如表 5-2 所示。

再用改进的 K-means 算法对数据进行聚类。算法改进后的 F-measure 值与未改进的 F-measure 平均值的对比如表 5-3 所示。

表 5-3 K-means 算法改进前后的 F-measure 值对比

类别	D1	D2	D3	D4	D5	D6	D7	D8
改进前（均值）	0.657	0.623	0.670	0.646	0.647	0.677	0.630	0.713
改进后	0.783	0.736	0.808	0.801	0.798	0.813	0.783	0.738

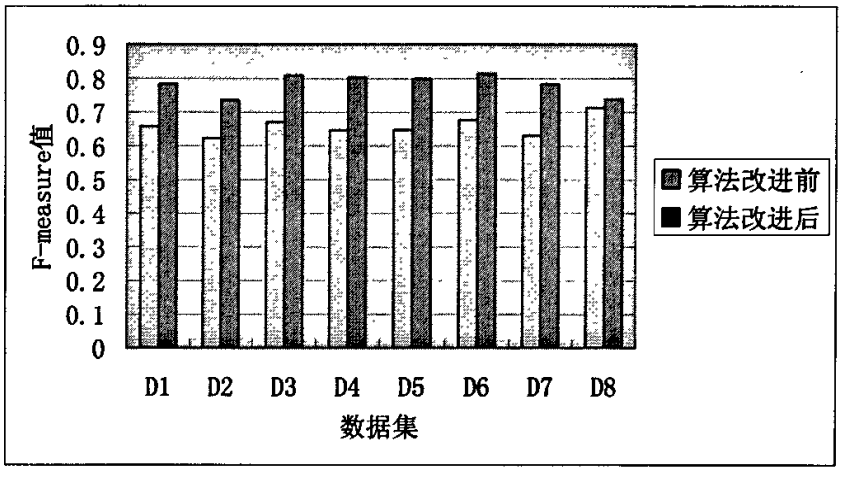


图 5-1 K-means 算法改进前后 F-measure 值比较柱状图

表 5-3　F-measure 值的分布情况

F-measure 区间	未改进算法的 F-measure 值落入此区间的次数	改进算法的 F-measure 值落入此区间的次数
[0.45，0.55）	3	0
[0.55，0.65）	9	0
[0.65，0.75）	10	8
[0.75，0.85）	8	18
[0.85，0.95）	0	4
[0.95，1.00]	0	0

为了更直观的比较 K-means 算法改进前后的结果，根据表 5-3 绘出 K-means 算法改进前后 F-measure 值比较柱状图（如图 5-1）。

为了验证改进后聚类算法结果的稳定性，本文采用了多组数据分别利用改进前后的两种算法进行对比实验，实验结果中F-measure值的分布情况如表5-3所示。

从表5-3种可以看出，采用未改进的K-means算法得到的聚类结果稳定性不好，F-measure值的跨度区间较大，比较分散；而采用改进的K-means算法得到的聚类结果稳定性较好，F-measure值比较集中，而且F-measure平均值较高。

5.3 实验小结

通过针对K-means算法的主要缺点和不足，对K-means算法中的关键环节如相似度计算公式，初始聚类中心的选择和新聚类中心的计算方法进行了改进。并且使用F-measure评价方法对K-means算法整体改进后的聚类效果进行评价，通过实验性能对比说明了改进后的算法在准确性和稳定性方面都有所提高。

第六章 结束语

6.1 论文总结

本文研究了一种基于改进 K-means 算法的 Web 文档聚类系统，开发出了一套由网络爬虫、数据清理、中文分词、特征提取、权重计算和聚类分析等模块组成的 Web 文档聚类系统。同时，针对 K-means 算法的主要缺点和不足，本文对 K-means 算法中的关键环节如相似度计算公式，初始聚类中心的选择和新聚类中心的计算方法进行了改进。并且使用 F-measure 评价方法对 K-means 算法整体改进后的聚类效果进行评价，通过实验性能对比说明了改进算法的优越性。

文章对数据挖掘、聚类分析和 Web 挖掘进行了概述和总结，介绍了整个系统的架构。并对网络爬虫、中文分词、英文词干提取、特征提取，权重计算和聚类分析等模块进行了深入的研究。最后，通过开发的由网络爬虫、数据清理、中文分词、特征提取、权重计算和聚类分析等模块组成的 Web 文档聚类系统进行了对比实验，验证了基于改进 K-means 算法的 Web 文档聚类系统在准确性和稳定性方面都有所提高。

在文章最后，对论文进行了总结，介绍了本文的主要工作内容，并对进一步的研究和需要完善的问题提出了看法。

本文主要做了下述工作：

（1） 本文对文本挖掘中的文本表示方法、特征提取、权值计算进行了系统的研究，并对网络爬虫、分词、聚类等过程进行了比较详细的阐述。

（2）开发了一整套由网络爬虫、数据清理、中文分词、特征提取、权值计算和聚类分析等模块组成的系统。

（3）对 K-means 算法进行了有特色的改进。区别于其它论文中只对算法的某一方面进行改进，本文综合了大量参考资料，针对 K-means 算法的主要缺点和不足，对 K-means 算法中的关键环节：相似度计算公式，初始聚类中心的选择和新聚类中心计算方法进行了改进。

（4）使用 F-measure 评价方法对 K-means 算法改进前后的系统进行评价，通过 F-measure 值对比表、对比柱状图和 F-measure 分布情况说明了改进算法的在准确性和稳定性方面都有所提高。

（5）论文最后除对研究工作进行了总结外，还对今后的研究方向进行了展望。

6.2 下一步研究的方向

本论文的研究工作只是所涉及领域很少的一部分，对有些问题还只是进行了一些有意义的尝试与探讨。因此今后将在现有研究基础上继续作深入的研究，主要包括：

1，对于传统的聚类分析算法的效果进行更多的实验分析、比较，加深对文本上的聚类分析处理特点的认识，指导工程实践；

2，在改进算法，提高准确性和稳定性的同时，对运算效率的提高进行优化；

3，用其他聚类算法实现，进行对比分析。总结出不同算法的优点和不足。

4，聚类结果可视化表示，将聚类结果更形象地展示出来。

5，应用领域的扩展。将改进后的K-means算法应用于其它热点领域。

参考文献

[1] 李雄飞，李军，数据挖掘与知识发现。高等教育出版社，2003

[2] Fayyad U M, Piatetsky-Shapiro G, Smyth P. Advance in Knowledge Discovery and Data Mining. Cambridge MA: AAAI/MIT Press.1996

[3] 王丽坤，王宏，陆玉昌，文本挖掘及其关键技术与方法，计算机科学，2002，Vol29N0.12，12-13

[4] 王莉，数据挖掘中聚类方法的研究：[博士学位论文]，天津：天津大学，2003

[5] 赵恒，数据挖掘中聚类若干问题研究：[博士学位论文]，西安：西安电子科技大学,1998

[6] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, 1:281-297, 1967

[7] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York; John Wiley & Sons, 1990

[8] 王涛，孙河山，Web挖掘技术在搜索引擎中的应用，信息系统，2002，25（4）：296-298

[9] 李颖，阎保平，Web文本挖掘在互联网信息统计中的研究与设计，微电子学与计算机，2005年第22卷第1期

[10] 鲁鹏俊，一种基于SVM的网页分类系统的设计与实现：[硕士学位论文]，上海 ：复旦大学，2006

[11] 张华平，刘群，基于N-最短路径方法的中文词语粗分模型. 中文信息学报，16卷5期，Sep. 2002.

[12] 刘群，张华平， 基于层叠隐马模型的汉语词法分析. 计算机研究与发展, 41 卷 8 期, 2004.

[13] http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm

[14] *The Porter Stemming Algorithm*, http://www.tartarus.org/martin/PorterStemmer/

[15] Dubes R C Jain A K. Algorithms for Clustering Data. Prentice Hall, 1988

[16] Kaufman L, Roussecuw P 1. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, 1990

[17] Fasulo D. An Analysis of Recent Work on Clustering Algorithms Technical Report UW-CSE-0I-03-02, University of Washington, 1999

[18] 王子兴，冯志勇，Web文档聚类中k-means算法的改进，微型机与应用，2004 Vol.23 No.4 P.50-52

[19]高茂庭，王正欧，基于LSA降维的RPCL文本聚类算法，计算机工程与应用，2006 Vol.42 No.23 P.138-140

[20] 袁方，孟增辉，于戈，对k-means算法聚类算法的改进，计算机工程与应，2004 Vol.36 P.177-178

[21] 万小军，杨建武，陈晓欧，文档聚类中 k-means 算法的一种改进算法，计算机工程，2003 Vol.29 No.2 P.102-103

[22] M. Halkidi, M. Vazirgiannis, Y. Batistakis. Quality Scheme Assessment in the Clustering Process. Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. 2000.265-276

[23] H. Ayad, Kamel M. Topic discovery from text using aggregation of different clustering methods. Cohen R, Spencer Bed. Advances in artificial intelligence: 15th conference of the Canadian society for computational studies of intelligence. Calgary, 2002.161-175

攻读硕士期间参与的科研项目和撰写的论文

科研项目情况：

1，网络信息安全理论研究和原型开发

2，互联网环境的信息内容安全的技术与管理

撰写的论文情况：

1，王钦平，张世永，钟亦平，傅维明，“基于 Web 的分布式网络流量监测系统，”计算机工程，Vol. 31, PP. 158-160, Dec. 2005。

2，王钦平，张世永，钟亦平，傅维明，“一种改进 TTCP 性能的新方法，”计算机应用与软件。

致谢

首先我要衷心地感谢我的导师张世永教授和钟亦平教授！感谢他们在这三年的研究生学习生活中，对我孜孜不倦的指导，而且为我提供了良好的学习和实验环境。尤其是在撰写论文的过程中，两位老师一直给予我鼓励和支持，并多次从百忙之中抽出宝贵的时间给予认真而细致的指导，提出了很多宝贵的意见。

我也要感谢傅维明老师，费舒红老师，吕智慧老师以及全体网络中心所有的老师和同学们，在我学习，找工作和撰写论文的过程中他们都给了我很大的支持和帮助。

我还要特别感谢侯亚飞师兄！在整个毕业论文的撰写过程中，侯师兄给了我很多的帮助和指导。从师兄身上我学到的不仅是写论文，做学术的方法，更重要的是一种严谨，踏实，认真负责的态度！

同时，我要感谢我的父母，亲人和朋友。无论我走到哪里，无论我做什么，都离不开父母和亲人的支持，牵挂和爱。特别是在那些关键和重要的时刻，我深深地感受到他们的心是与我相连的！

研究生的三年即将过去，此刻的我真正体会到了“这个世界上唯一不变的就是变化。”学习是没有止境的。我已做好了终身学习的准备，去迎接新一轮的挑战。

谨以此文献给所有关心过我的老师、亲人和朋友，衷心地祝愿你们：好人一生平安！

基于改进K-means算法的Web文档聚类系统的研究与实现

作者： 王钦平
学位授予单位： 复旦大学

——附加文档一篇——
工程概况

刘家湾北段市政工程总长度545m；道路设计红线宽度主线30m,一副路面；车行道16m;绿化带2*4m；人行道2*3m。

刘家湾北段市政工程设计内容包括：道路、雨水、污水、给水、照明、弱点管道、标志标线工程。

技术指标：

1、道路性质：城市主干道（2级）

2、设计行车速度：40km/h

3、使用年限：15年

4、车行道、人行道设计坡度：2%

主要设计依据：

1、咸阳市住房和城乡建设局：“刘家湾北段市政工程”设计委托书。

2、咸阳市城乡规划建筑设计院：“咸阳市彬县泾河区建设规划图”。

3、《城乡道路设计规范》（CJJ37-90）

4、《城市道路和建筑物无障碍设计规范》（JGJ50-2001、J114-2001）

5、《公路沥青路面设计规范》（JTGD50-2006）

6、《公路沥青路面施工技术规范》（JTGD40-2004）

7、《公路路基设计规范》（JTGD30-2004）

8、《公路路基施工技术规范》（JTGF10-2006）

9、国家其它有关设计规范及标准。

第一章　　项目经理部组成

1.1　工程项目管理模式

项目经理部由公司总部授权管理，按照企业项目管理模式GB、T19001-
ISO91001标准模式建立的质量保证体系来运作，质量管理为中心环节，以专业管理和计算机管理相结合的科学化管理体制。

项目经理部按照我公司颁布的《项目管理手册》、《质量保证手册》、《项目技术管理手册》、《项目质量管理手册》、《项目安全管理手册》、《项目成本管理手册》执行。

1.2　工程项目管理的主要目标

质量目标：合格

工期目标：我公司投标自报工期为180日历天，满足建设单位的要求。在满足合同工期和工程质量的前提下，尽量加快施工进度，使工程提前交付使用，实现我公司

的承诺。

成本目标：科学管理，精密组织，在“人、机、料、法、环”五个影响工程造价的因素方面加强管理和监控，杜绝返工等质量事故，提高工程一次验收合格率，从而降低工程的成本。

安全目标：确保不发生重大伤亡事故，杜绝死亡事故，轻伤事故频率控制在5‰以内。

文明施工目标：以创建“陕西省文明工地”，作为文明施工的标准。

消防目标：严格按照施工组织设计进行管理，消除现场所有的消防隐患，达到各级消防主管部门和公司上级主管部门的验收标准。

环保目标：创“绿色、无污染施工工地”。

教育培训目标：选派年轻有发展潜力的技术人员参加公司举办的各类培训学习，与工程施工有关的先进技术可以单独联系在外学习。

协作目标：积极配合甲方、监理、设计院和其他相关单位的工作和监督检查，圆满完成工程项目的施工，给公司创形象，为业主增光彩。

竣工回访和质量保修计划：根据我公司对业主的承诺，每年夏季对用户进行回访。根据第80号建设部令，《房屋建筑工程质量保修办法》的有关规定进行保修。房

屋地基基础和主体结构工程，保修年限为设计规定的该工程的合理使用年限，屋面防水、卫生间防水以及外墙面和房间的防渗漏、保候修期为5年，供热系统保修期为二个采暖期，电气系统、给排水设备安装保修期为二年；装修工程保修年限为二年。其它保修期限由建设单位与我方以合同形式约定。

1.3　项目组织机构及主要管理人员名单

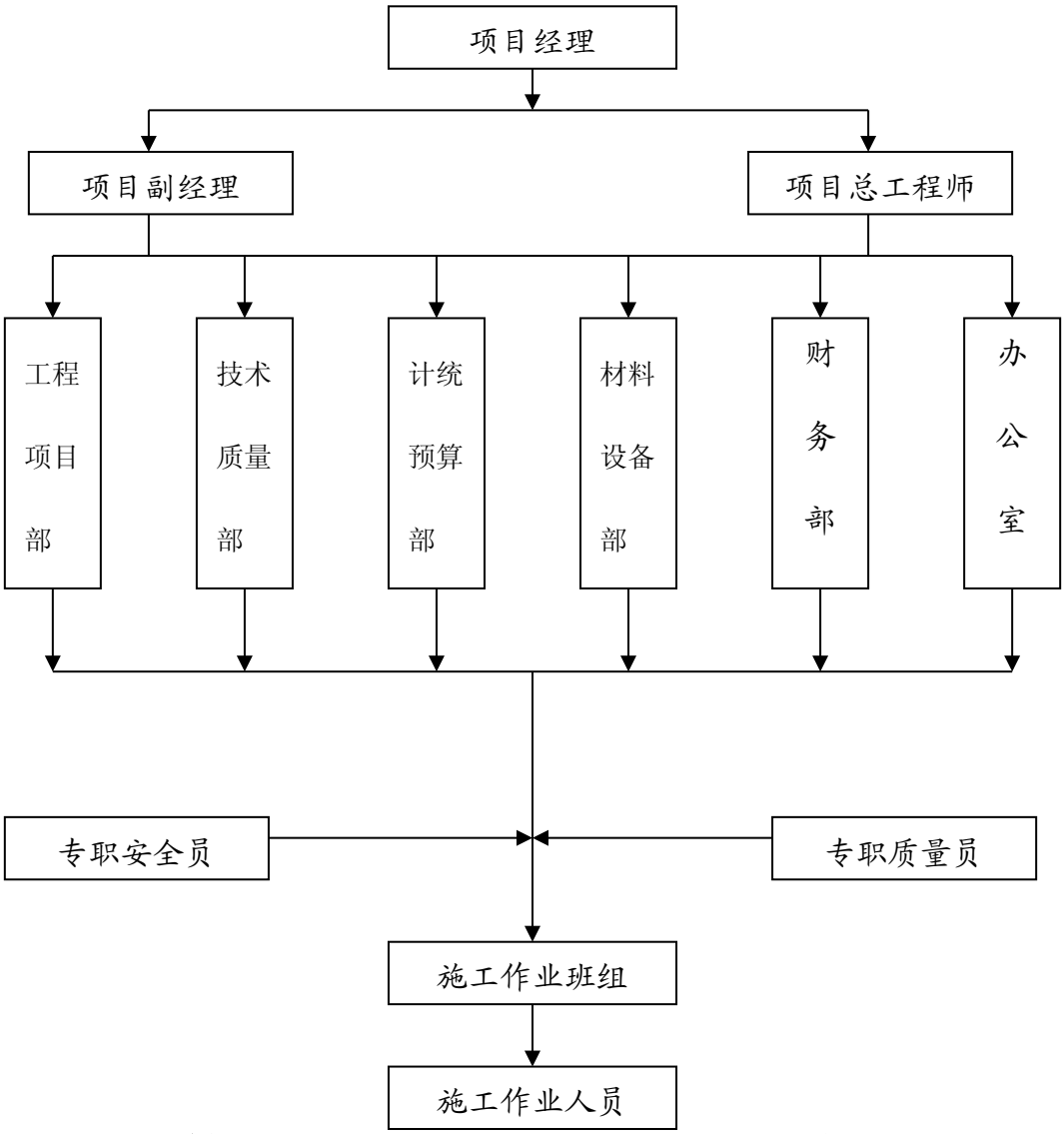
本工程工期紧、任务重，本着“建造满意工程，提供优质服务，一切为了业主与用户”的原则，我公司将选配具有多年施工经验的管理人员及技术干部，组成一个高效精干，开拓务实、富有活力的项目经理部，在现场全权代表我单位行使管理职能，履行合同的各项权力和义务，确保该工程如期、高效、优质、安全建成。项目部下设工程技术室、财务室、物资设备室、预核算室及综合办公室，各部室配备专业技术人员，负责现场施工组织及质量、安全、技术、进度计划以及文明施工等各项管理工作，监督检查各分部、分项工程施工。

项目组织机构及项目主要管理人员名单见附表。

职务	姓名	性别	职称
项目经理	高志春	男	工程师
材料员	尚文琴	男	助工
施工员	吴小茹	男	助工

安全员	雷明录	男	助工
质检员	范雄飞	男	助工
专职安全员	郝卫星	男	技术员

项目组织机构图



第二章　施工部署及总平面布置

本工程质量标准要求高、计划施工总工期180日历天，期间经历雨期，工期较紧张。为了保证基础、面层、照明均尽可能有充裕的时间施工，优质高效地完成本工程，须充分考虑到各方面的影响因素，从任务划分、人力、资源、时间、空间的充分利用与合理配置上，科学部署，严密组织工程流水施工。

2.1 施工部署原则

为了保证基础、面层、排水、交通设施装饰均可能有充裕时间施工，保证如期完成施工任务，全面考虑各方面的影响因素，充分酝酿任务、人力、资源、时间、空间的总体布局。

2.2 总体施工顺序和部署原则

总体施工顺序：基层工程、面层工程、排水工程、照明工程、竣工验收。

2.3

工程组织分段流水作业，采用机动翻斗车运输；采取“样板法”施工，严格按照“三样板一顺序”法施工。

2.4 季节性施工的考虑

根据施工进度的安排，需历经雨季施工，编制较为详细的季节性施工措施，以加强质量控制与管理。

2.5 交叉施工的考虑

为了贯彻空间占满时间连续，均衡协调有节奏，力所

能及留有余地的原则，保证工程总进度计划完成，需要采用基层、面层、排水和指示牌的交叉施工。

2.6　机械设备的投入

根据施工工程量和现场实际条件投入机械设备。灰土搅拌采用一台双卧强制式JS500型搅拌机，一台HPD电子配料机计量，配二台沥青泵车，机动翻斗车，进行现场材料运输。

2.7　施工流水段划分

2.7.1 基层工程施工流水段划分

本工程基层不划分施工段。

2.7.2 排水施工流水段划分：

该工程施工划分为二个流水段流水施工。

2.8　现场施工条件

2.8.1由业主负责提供现场施工的用电，现场施工用电总柜从变压器引入。

2.8.2

业主在现场有供水点，作为现场临水的水源，水源的水管要求为Φ100，以满足施工用水要求。

2.8.3

进场前现场内有障碍物已消除，具备施工条件。

2.8.4　现场内地面全部硬化处理，做到无黄土外露。

2.9　施工现场平面布置

详见现场平面布置图

2.10 施工扰民问题

本工程要认真考虑尽量减少施工扰民问题，并严格按照有关规定执行。

2.11 临时用电、用水设计

1.临电设计：

综合以上11组用电设备的计算负荷，取周期系数KP=KQ=0.8，则PJ=0.8*(53+27+.52.5+30.87+8.32+46.08+25.12+12+17.6+9.45+50.5)=0.8*332.44=265.95Kw

QJ=0.8*(54.09+27.54+46.2+23.15+11.07+61.27+25.62+12.24+13.2+7.09)=0.8×281.47=225.18KVA

SJ= =318KVA

现场供电为三级配电，即总箱、二级箱、三级箱(手提箱及插座箱)，两级空开、三级漏电保护。配电方式以放射和串接结合，采用三相五线制，供电选重型橡套电缆。考虑到现场实际情况及工程各个施工阶段的用电负荷情况，配电箱平面位置见施工平面布置图。

2、临水设计

a、临水计算

根据本工程现场实际布置特点，现场临时用水管道采用聚乙烯PE管。现场临时用水包括：施工用水Q1，现场

生活用水Q2，现场消防用水Q3。其中计算如下：

Q1=K1q1N1K1/8/3.6/300=1.15(16000+4000)÷8÷3.6÷300=

2.7L/s)

Q2：现场施工高峰期施工人员按200人计算。

Q2=q2N2K2/8/3600=200×20×1.5÷8÷3600=0.21L/s

消防用水Q3：本现场物料堆放齐全，因此现场消防器材布置相当重要。根据现场施工临水水量规定，当施工现场占地不大于1ha（公顷）时，q3取15L/s。

考虑管道漏水系数1.1，则q3=1.1×15=16.5L/s。

现场施工用水干管管径D为：

D=SQR（4Q/π/v/100)=SQR（4×16.5÷3.14÷2÷100)=0.1=100mm。

b、临时用水布置说明

临时用水管管材采用镀锌钢管。埋地敷设，埋设深度为80厘米，以免被现场过往车辆损坏，冬季还可起到保温作用。用水由闸阀控制。供现场消防。现场设临时用水值班人员两名，负责现场用水的巡视维修。

增强职工的环保意识，节约用水，严禁水管出现跑、冒、滴、漏象，临时用水及消防平面布置见附图。

第三章　施工进度计划及措施

3.1 工期目标

本工程我们经仔细研究，并结合我单位实际，在保证安全、质量的前提下用180日历天完成全部施工任务，并争取提前竣工。

3.2 工期保证措施

本工程按我单位较成熟的项目法管理体制，建立规范化的项目体系，实行项目经理负责制，成立本工程项目经理部，实行项目法施工，对本工程行使计划、组织、指挥、协调、施工、监督六项基本职能，配备有力的管理层，选择能打仗、并有大型建筑施工阶段组成作业层，承担本工程施工任务。

3.2.1 制度保证

3.2.1.1 建立生产调度例会制度

每周至少召开一次工程调度例会，检查上一次例会以来的工程计划执行情况，
布置下一次例会前的工作安排。找出拖延施工进度的原因，并及时采取有效措施保证计划完成。

3.2.1.2　动态控制施工进度的制度

采用施工进度总计划与月、周计划相结合的各级网络管理体系，进行施工进度计划的控制与管理，采用网络技术进行动态管理。在施工中抓主导工序，找关键矛盾，组织流水交叉，安排合理的施工顺序。做好劳动力调配和协调工作，通过施工网络节点控制目标的实际现保证各控制点工期目标的实现，从而进一步通过各控制点工期目标的实际，确保总工期控制进度计划的实现。

3.2.1.3　提高劳动工效，落实劳动承包制度

通过逐级签定劳动承包合同，约束劳动双方的行为。即项目部要求作业队在一定时间完成某项工作，必须及时提供充足的设施料，状态良好的设备以及必要的技术指导。谁违反合同就处罚谁，改变只约束作业队的“单方合同”。并实行合同价与工期指标挂钩，组织劳动竞赛等活动，充分调动作业层的积极性和主动性。确保劳动工效提高20-30%。

3.2.2　劳动力保证

依据本工程的特点，与劳务队伍签订施工人员保证合

同，同时与公司其他劳务分包商签订补充劳务合同，作为劳动力储备。一旦出现劳务队人员不能满足施工要求，立刻限期补充，如不能及时补充，项目部将执行与其他劳务分包商签订的补充劳务合同，以此补充劳动力原劳务队进行严厉处罚。

3.2.3　　施工技术保证

3.2.3.1

加强技术管理工作，精心组织施工，合理安排好施工程序和流水作业，加快施工进度，缩短施工周期。

3.2.3.2

科学地制订施工进度网络计划，强化计划管理，加强日进度计划控制、旬进度计划检查和月进度计划考核，以日进度促进旬进度，以旬进度保证月进度，以月进度确保总工期的实现。

3.2.3.3

认真进行图纸预审和参加图纸会审，与设计单位加强联系和沟通，抓好设计变更的落实工作。

3.2.3.4

在整个工程施工阶段合理组织立体交叉作业，充分利用场地、空间，加快施工进度。

3.2.3.5

充分利用技术、新工艺等科技手段，加快施工进度。

3.2.3.6

科学地制订季节性施工方案，合理安排冬雨季施工期内的工作内容，采取切实可行的有效措施确保产品质量，使工程持续和均衡进行，促进工程进度。

3.2.3.7

积极作好各种影响施工进度因素的预防工作，如停水、停电、风、雨天等，采取各种积极有效的措施和手段，如配备发电机、

蓄水箱、防雨布等，把不利因素降到最低。

3.2.4 机械保证

3.2.4.1

充分利用机械化程度高的有利条件，配备适宜的施工机械，减轻劳动强度，提高工作效率。

3.2.4.2

加强施工机械、设备和设施料的配备、维修工作，充分保证施工进度的需要。

3.2.5 资金保证

3.2.5.1

本工程项目资金全部专款专用，确保施工中各项费用开支。

3.2.5.2

施工期内如建设单位资金一时发生缺口，内部银行及时

给予适当解决，满足工程进度需要。

3.2.6　　物资保证

3.2.6.1

机械物资保障部积极协助项目部做好各种物资的供应工作。

3.2.6.2

项目部保障部按施工预算和工程进度及时编制物资用量计划并组织采购和进场。及时对进场物资进行验收和质量验证，保证合格物资投入施工。

3.2.7　　后勤工作保证

3.2.7.1

做好职工思想教育工作，关心群众生活，提高食堂饭菜质量，夏季做好防署降温工作，及时供应茶水、饮料和绿豆汤、冬季做好职工宿舍的保温取暖工作。

3.2.7.2

搞好现场文明施工，做好工地宣传和开展各种娱乐活动同，创造良好的的工作和生活环境，增强职工的凝聚力，形成一个团结、 紧张、奋发向上的工作局面。

3.2.7.3

开展劳动竞赛，建立奖励制度，精神鼓励与物资奖励相结合，激励施工管理人员和操作工人的生产劳动积极性。

3.3　施工总进度计划图

施工总进度控制计划详见施工进度网络图。

第四章　施工方案

根据工程概况，平面布置，施工组织机构，在充足劳动资源的基础上，将招标文件要求的实质性文字说明叙述如下：

（一）设备人员动员周期和人员、材料运到施工现场的方法

1、设备人员动员周期：一旦接到中标通知书，立即进行人员设备的动员和调遣工作。3天内派项目经理部的主要负责人进驻工地，与监理、业主接洽，详细勘察、了解沿线及场地情况，安排落实施工用地，着手水、电、路三通准备工作，对料场、预制场及机械停置场进行硬化。

按照施工工序的先后，组织相应机械同期进驻工地，做到“三通一平”，保证线位恢复、原材料试验、砼配合比设计、人员进场同步进行。

设备人员动员周期5天，其中主要工程机械和人员动员

周期控制在3天之内。

2、设备、人员、材料运到施工现场的方法：大型机械设备如压路机、发电机组设备采用平板车运输，自卸汽车、吊车、洒水车等适合于公路行驶的设备，直接开往工地。其它小型机具及施工设备采用汽车运输。

主要材料如钢材、水泥等直接送货到场，碎石、砂等地材从当地料场采购，自卸汽车运到现场。汽油、柴油等其它材料在就地购买送往工地。

（二）施工准备工作

1、临时设施

（1）临时房屋：经理部拟设在道路旁，施工队部分考虑租住当地住房。

（2）临时用电：充分利用沿线已有电力资源，架设拌和场专用线路，用电率考虑70%。配备发电机组3套，工程自备用电率考虑30%。

（3）临时通讯：因沿线通讯设施较好，可采用安装电话解决通讯问题。施工中配备对讲机加强联系。

（三）各分项工程的施工顺序及施工要点

1、路基填方

①施工顺序

填土：测量放样 — （场地、地表）清理— 翻松— 填前碾压— 检查验收— 纵横向取料— 分层摊铺— 推土机整平— 洒水— 碾压成型— 检查验收压实度— 合格进入下层施工。

②施工要点

a、路基施工设计困难，挖掘机、推土机考虑大吨位履带式设备，压实设备应考虑双轮驱动振动压实设备，并作好土工实验。

b、取料、弃土场地因运输困难，应根据设计文件就近选择。

c、应翻松30cm，进行填前碾压。

d、严格控制层厚，填土不大于30cm。

e、做好临时排水设施，为保证压实度填筑宽度应两侧多填50cm。

2、路基挖方

测量放样—　　路基开挖—　　路床整理—大于120马力推土机稳压—振动压力机碾压—　边坡防护—路基封闭 —检查验收。

3、涵洞

（1）施工顺序：

施工放样—挖基—　基底处理—混凝土基础　—安管—管缝处理—混凝土管座—防水沥青—　涵背填料及夯实—涵顶填土。

（2）施工要点：

①基底夯实，墙后回填料密实，避免不均匀沉降。

②混凝土强度达到75%以上方可进行路基填土。

③涵管安装时要位置要准确，不得碰掉棱角。

（四）主要工程项目的施工方案

1、路基土石方工程施工方案

本标段全长775.136m,该段路基土方施工安排一个作业队，按顺序进行施工。

（1）基准点、导线点、水准点复测

测量人员组织对设计单位、监理单位所交基准点、导线点、水准点进行全面复测，整理出复测结果并报监理工程师审批，以获取第一资料。

（2）对全线原地面纵、横断面复测

复测完成后，根据监理工程师审批结果，开展对原地面的纵、横断面进行测量，并将测量成果报监理工程师审批复核。

（3）场地清理

①测量放样，确定清理范围。对清理范围内的清理工程量、挖除、砍伐树木数量，造表统计，请监理工程师验收认可后，再进行清理，挖除及砍伐。

②对清理的草皮、垃圾、腐殖土现场堆积，然后按照招标文件要求或监理工程师的指令，运离施工现场。对砍伐的树木按招标文件要求或监理工程师指令堆放整齐，验收数量后报业主单位。

③对不同的地段，采用不同的清理方式，对能够进行机械施工的，采用机械清理，对陡坡、局部小坑机械不能清理的采用人工清理，清理结果，要达到招标文件和监理工程师要求为至。

④填土碾压。清理工作完成后，通知监理工程师进行验收

（4）路基填方施工

路基填方将按以下方法来完成。

①首先根据招标文件及监理工程师要求，由技术部牵头，中心试验室、安全质量部配合，完成土质分析试验报告，确定土质的最佳含水量及击实报告。

②修试验路段。由技术部牵头，工程部、安全质检部、中心试验室、路基土方施工队配合，按监理工程师要求完成一段试验路修筑，确定路基施工中的各种技术参数，试验路修筑完成后，总结成报告，报监理工程师审批，再进行路基填方的大面积施工。

③零填挖地段。对于零填挖地估的施工，对于含水量接近最佳含水量的地段，可直接翻松整平，然后进行碾压；对于含水量偏大的，可采用换土的方法进行碾压，总之，对零填挖地段，可视具体情况确定其施工方案。

④填方路基，本合同段土方工程以路基填方为主，为了搞好此项工作，我经理部要求：技术部、工程部、安全质检部、中心试验室全力配合路基土方施工队，严把

质量关，共同搞好路基填方工作。

A、选择好填方路基的土源

土源质量的好坏，直接决定着路基的质量，我们将按照招标文件所指明的土场，逐一进行试验，以确定最佳土源。保证路基施工的原材料没有质量问题。

B、确定施工人员最佳组合

施工人员组合的好坏，是关系到路基填方质量的一个重要方面，因此我们考虑在路基土方施工队人员基础上，委派工程部、质检部、中心试验室各一人到路基土方施工队加强管理，及时发现问题，解决问题，使问题在萌芽状态得以消除。

C、机械配制

机械是做好路基填方工作的主要工具，为了搞好路基填方工作，保证路基压实效果和质量，我们将以我单位最新、最好的设备予以投入。对用于填方路基需要的压路机、推土机、平地机、运输车辆，我单位将全力支持，在保证质量的前提下，按合同工期按时完成。

D、填土施工方法

严格按路基试验路段总结的各项技术参数及监理工程师要求进行施工，并按以下步骤逐段逐层进行。

a、准确放样中线、边线、测量每层标高。

b、按照换算结果，运输车辆排开卸土，避免造成卸土过密。

c、推土机打开卸土并初平。

d、测量松铺厚度，推土机精平。

e、光轮压路机初压。

f、振动压路机终压。

压路机碾压时按照先两边后中间，先轻后重，先慢后快的原则进行。

g、试验人员按质量评定标准进行试验自检,并将自检

结果报监理工程师审批。

h、邀请监理工程师抽检，抽检合格再进行下层施工，否则，进行补压，直到监理工程师认为合格，再进行下一层施工。

i、资料整理并保存。

⑤土方路基填筑：

路堤填料分层平行摊铺，且每层的压实厚度不宜大于20cm。在填筑前，应对填料进行各项试验，以确定填料的最佳含水量和压实度，并在路基填筑前，应根据监理工程师的指示现场确定铺筑长度不小于100m路幅全宽的试验路段，以确定设备的选择、工序、压实遍数、行进速度以及能够压实的有效厚度。

各层的铺筑和整平使用平地机在每层压实前进行整平作业，以保证均匀的压实度。土方压实采用拖式振动羊角碾压（激振力在40T以上），辅助平滚压路机一台（激振力20T）。

首先，用于路堤填筑的材料不应含有杂物或其他有害物质，路基填筑前应取土试验，其强度和粒径应符合设计要求，当达不到要求时，可采取掺石灰固化材料处理，掺入量通过试验确定。

路基填筑前，填方材料应按JTJ051-93标准方法进行颗粒分析、液限和塑限、有机质含量、CBR和击实试验，进行击实试验时，用重型击实法确定土的最大干密度和最佳含水量。

路基填筑时，当填土高度低于1.0m时对原地表清理与掘除之后的土质基底应将表面翻松深30cm；当填土高度大于1.0m时，对于土质基底应将原地面整平压实到无轮迹时方可填筑。

填筑时，应均匀的把填料摊铺在整个宽度上，在压实前应先整平，并作2-

4%的横坡，碾压时，振动压路机前后两次轮迹须重叠40-50cm，前后相邻西两区段纵向重叠1.2-1.5m，并应达到无漏压、无死角，确保碾压均匀。压实后的土方压实度应不小于95%，按JTJ051-93重型击实法进行检验，直至达到规范要求。

在路基填筑同时，我们将作好整个路基排水工作，做到路基表面无积水，排水畅通。

（5）路基挖方施工

①路基挖方施工步骤

本合同段路基挖方工程量较小，但为了搞好路基挖方施工，我们准备采用以下步骤进行。

A、准确放出公路中线，并每50m精确测出地面横断面线，根据招标文件要求计算出挖方开挖边界线，以避免造成多挖或少挖。

B、将上述放线结果报监理工程师批准后方可进行开挖，对于开挖高度及工程量较小的，可采用全断面同时开挖方式进行。

C、做好挖方中利用调配计划，对于不能利用的按招标文件及监理工程师要求进行废弃，并要求平整复耕。

D、刷边坡，要求挖方在进行的同时，人工配合机械刷边坡，做到一次过手，决不返工。

E、做好开挖中的防排水设施。开挖线以外，人工修筑挡墙，高度不小于0.3m，宽度不小于0.5m。开挖路槽内两侧设排水沟，并在适易地段设积水坑，以便将降雨汇积一处，天放晴后及时排走。

F、当标高挖至路床标高时，先用推土机初平，再用平地机精平，如果含水量合适，可直接将路床标高下30cm翻松用压路机分两层碾压达到路床设计标高，如果含水量偏大，可适当翻松晾至含水量稍大于最佳含水量再进行碾压。

G、压实度检测。碾压成型后，要对路槽压实度进行检测，合格后方能交验。首先由试验人员自检，并将自检结果报监理工程师，然后由监理工程师抽检。原则上压实必须一次合格，不能返工。

②路基土方开挖：

A、施工工序：施工放样—挖掘机就位挖装—自卸车运输—路槽翻松压实—人工修整边坡。

B、工序要点：开挖过程中，应注意不松动边坡，避免破坏边坡稳定性。场内排水应畅通。。

3、石灰土底基层的施工方案和施工方法

经过认真分析和研究，对石灰土底基层决定采用路拌法施工，成立一个底基层施工队，负责底基层的施工。

A、材料要求

土、灰除满足规范要求外，施工中控制点为：

（1） 石灰应符合III级以上标准，石灰在使用前10天充分消解，并过筛（10MM筛孔）；

（2）消石灰存放时间宜控制在2个月以内；

（3）一个作业段内采用土质相同的土（击实标准和灰剂量相同）， 以便对压实度进行准确控制。

B、准备下承层

（1）石灰土施工前，应对路槽进行严格验收，验收内容除包括压实度、弯沉、宽度、标高、横坡度、平整度等项目外，还必须进行碾压检

验，即在各项指标都合格的路槽上，用18-21T压路机连续碾压2遍，碾压过程中，若发现土过干、表层松散，应适当洒水继续碾压；如土过湿、发生翻浆、软弹现象，应采用挖开晾晒、换土、外掺剂等措施处理。路基必须达到表面平整、坚实，没有松散和软弱点，边沿顺直，路肩平整、整齐。

（2）按要求设置路面施工控制桩。

C、备土、铺土

用于石灰土的土必须符合规范要求，不含树皮、草根等杂物。备土前要用土培好路肩，路肩应同结构层等厚。

备、铺土分两种方法：

（1）用汽车直接堆方备土

按照每平米的松土用量及每车的运土量，用石灰粉标出每车的卸土位置（划出方格），直接整齐地卸土于路槽上。但须注意备土时纵向必须成行，每车的运土量要基本准确，同一作业段内土质基本均匀一致。该方法有利于机械化施工，但备土数量不易准确控制。铺土时，先用推土机大致推土，然后放样用平地机整平，清余补缺

，保证厚度一致，表面平整。

（2）码条备土

用拖拉机等小型机械备土可采用此方法。

按照每延米的松土用量，分两条成梯形状均匀地码条于路槽上，用卡尺逐段验收备土数量。

备土时应在备土位置用石灰粉标出两条标线（码条的边沿位置），保证备土顺直，码条应均匀、数量准确。铺土时可直接用平地机均匀地将土铺开，保证表面平整、厚度一致。此备土法数量控制准确、摊土方便。

D、备灰、铺灰

备灰前，用压路机对铺开的松土碾压1~2遍，保证备灰时不产生大的车辙，严禁重车在作业段内调头。

备灰前应根据灰剂量、不同含水量情况下的石灰松方干容重及石灰土最大干容重计算每延米的石灰用量。

根据计算出的每延米石灰的松方用量，分两条成梯形状均匀地码条备灰，并用卡尺逐段验收数量，不准用汽车直接大堆备灰。备灰前应事先在灰条位置标出两条灰

线，以确保灰条顺直。铺灰前应在灰土的边沿打出标线，然后将石灰均匀地铺撒在标线范围内，铺灰应用人工撒铺。

E、拌和

采用灰土拌和机拌和，铁犁作为附助设备配合拌和。

（1） 土的含水量小，应首先用铁犁翻拌一遍，使石灰置于中、下层，然后洒水补充水份，并用铁犁继续翻拌，使水份分布均匀。考虑拌和、整平过程中的水份损失，含水量适当大些（根据气候及拌和整平时间长短确定），土的含水量过大，用铁犁进行翻拌凉晒。

（2） 水份合适后，用平地机粗平一遍，然后用灰土拌和机拌和第一遍。拌和时要指派专人跟机进行挖验，每间隔5~10米挖验一处，检查拌和是否到底。对于拌和不到底的段落，及时提醒拌和机司机返回重新拌和。

（3） 桥头两端在备土时应留出2米空间，将土摊入附近，拌和时先横向拌和两个单程，再进行纵向拌和，以确保桥头处灰土拌和均匀。

第二遍拌和前，宜用平地机粗平一遍，然后进行第二遍拌

和。若土的塑指高，土块不易拌碎，应增加拌和遍数，并注意下一次拌和前要对已拌和过的灰土进行粗平和压实，然后拌和，以达到拌和均匀，满足规范要求为准。压实的密度愈大，对土块的破碎效果愈好，采用此法可达到事半功倍的目的，否则既使再多增加拌和遍数也收效甚微。拌和时拌和机各行程间的搭接宽度不小于10cm。对于桥头处拌和同样采用先横向拌和2个单程，再进行纵向拌和。

F、整平

用平地机，结合少量人工整平。

（1）灰土拌和符合要求后，用平地机粗平一遍，消除拌和产生的土坎、波浪、沟槽等，使表面大致平整。

（2）用震动压路机或轮胎压路机稳压1~2遍。

（3）利用控制桩用水平仪或挂线放样，石灰粉作出标记，样点分布密度视平地机司机水平确定。

（4） 平地机由外侧起向内侧进行刮平。

（5）重复（3）~（4）步骤直至标高和平整度满足要求为止。灰土接头、桥头、边沿等平地机无法正常作业的地方，应由人工完成清理、平整工作。

（6）整平时多余的灰土不准废弃于边坡上。

（7）要点提示

- 最后一遍整平前，宜用洒水车喷洒一遍水，以补充表层水份，有利于表层碾压成型。
- 最后一遍整平时平地机应“带土”作业。
- 切忌薄层找补。
- 备土、备灰要适当考虑富余量，整平时宁刮勿补。

G、碾压

碾压采用振动式压路机和18~21三轮静态压路机联合完成。

（1） 整平完成后，首先用振动压路机由路边沿起向路中心碾压（超高段自内侧向外层碾压），有超高段落由内侧起向外侧碾压，碾压采用大摆轴法，即全轮错位，搭接15~20厘米，用此法震压6~8遍，下层压实度满足要求后，改用三轮压路机低速1/2错轮碾压2~3遍，消除轮迹，达到表面平整、光洁、边沿顺直。路肩要同路面一起碾压。

（2） 要点提示

- 碾压必须连续完成，中途不得停顿。
- 压路机应足量，以减少碾压成型时间，合理配备

为振动压路机1~2台，三轮压路机2~3台。

- 碾压过程中应行走顺直，低速行驶。
- 桥头处10米范围内横向碾压。

H、检验

（1）

试验员应盯在施工现场，完成碾压遍数后，立即取样检验压实度（要及时拿出试验结果），压实不足要立即补压，直到满足压实要求为止。

（2）

成型后的两日内完成平整度、标高、横坡度、宽度、厚度检验，检验不合格要求采取措施预以处理。

（3） 要点提示

- 翻浆、轮迹明显、表面松散、起皮严重、土块超标等有外观缺陷的不准验收，应彻底处理。
- 标高不合适的，高出部分用平地机刮除，低下的部分不准贴补，标高合格率不低于85%，实行左中右三条线控制标高。
- 压实度、强度必须全部满足要求，否则应返工处理。

I、接头处理

碾压完毕的石灰土的端头应立即将拌和不均，或标高误差大,或平整度不好的部分挂线重直切除，保持接头处顺直、整齐。

下一作业段与之衔接处，铺土及拌和应空出2米，待整平时再按松铺厚度整平。

桥头处亦按上述方法处理，铺土及拌和应空出2米，先横拌2遍再纵拌，待整平时再按松铺厚度整平。

J、养生

不能及时覆盖上层结构层的灰土，养生期不少于7天，采用洒水养生法，养生期间要保持灰土表面经常湿润。养生期内应封闭交通，除洒水车外禁止一切车辆通行。有条件的、对7天强度确有把握的，灰土完成后经验收合格，即可进行下道工序施工，可缩短养生期；但一旦发现灰土强度不合格，则需返工处理。

4、石灰土碎石基层的施工方案和施工方法

石灰土碎石计划用场拌法施工，由基层施工队负责全标段的基层施工。

A、材料要求

碎石（最大料径40mm）、土、灰质量应满足规范要求，施工中控制点为：

（1）石灰应符合III级以上标准，石灰在使用前10天充分消解，并过筛（10mm筛孔）；

（2）消石灰存放时间宜控制在2个月以内；

（3）一个作业段内采用土质相同的土（击实标准和灰剂量相同），　　以便对压实度进行准确控制。

（4）碎石压碎值不大于30%。

B、计算材料数量

根据结构层的厚度、宽度（按设计图纸）及预定的干密度，计算各段的干集料数量。

C、拌和

石灰土碎石在中心站用多种机械进行集中拌和，集中拌和时，必须做到：

1．土块要粉碎，最大尺寸不应大于15mm，粒料的尺寸要符合要求；

2．配料要准确，严格按照设计配合比施工（石灰：土：碎石＝5：15：80）；

3．含水量要略大于最佳值（约1%），使混合料运到现场摊铺后碾压时的含水量能接近最佳值；

4．拌和要均匀。

D、运输和摊铺集料

1．在摊铺段两侧先培土，以控制结构层的宽度和厚度。

2．用自卸翻斗车运输集料。装车时，应控制每车料的数量基本相同 。

3．卸料距离应严格控制，打石灰方格控制材料用量，必须由专人指挥卸料，避免铺料过多或不够。

4．卸料和摊铺时，通常由远而近全断面摊铺尽量不留纵缝。

5．提前通过试验确定集料的松铺系数。机械摊铺石灰土碎石，松铺系数约为1.2~1.4。

6．检验松铺材料层的厚度，视其是否符合预计要求。必要时，应进行减料或补料工作。

E、整型与碾压

用平地机结合人工整平。

（1）混合料卸料后，用平地机粗平一遍，使表面大致平整。

（2）用震动压路机稳压1~2遍。

（3）利用控制桩用水准仪测量，石灰粉作出标记，样点分布密度视平地机司机水平确定。

（4）平地机由外侧起向内侧进行刮平。

（5）重复（3）~（4）步骤直至标高和平整度满足要求为止。灰土接头、桥头、边沿等平地机无法正常作业的地方，应由人工完成清理、平整工作。

（6）整平时多余的混合料不准废弃于边坡上。

（7）要点提示

- 最后一遍整平前，宜用洒水车喷洒一遍水，以补充表层水份，有利于表层碾压成型。
- 切忌薄层找补。

（8）在最佳含水量的范围内，用12t以上的三轮压路机、振动压路机进行碾压，由两侧向中间，直到过到规定的压实度。严禁压路机在已完成的或正在碾压的基层上调头或急刹车。

F、接缝的处理

横缝：压实后末端做成斜坡（可为1：2），在第二天开始摊铺新混合料之前，应将留下的末端斜坡挖除，挖成一横向（与道面中心线垂直）垂直向下的断面，便可继续向前摊铺。

纵缝：尽量避免纵缝，在不能避免纵缝情况下（如较大站坪的石灰稳定土施工），纵缝必须垂直相接，严禁斜接，并尽可能减少纵缝的数量。

G、养生

在养生期间应保持一定的湿度，不应过湿或忽干忽湿。养生期不少于7d，可采用洒水（分散水流）或采用不透水薄膜。

5、沥青碎石面层的施工方案和施工方法

A、材料要求

（1）沥青采用石油沥青A-100；

（2）碎石最大料径19mm，压碎值不大于30%。

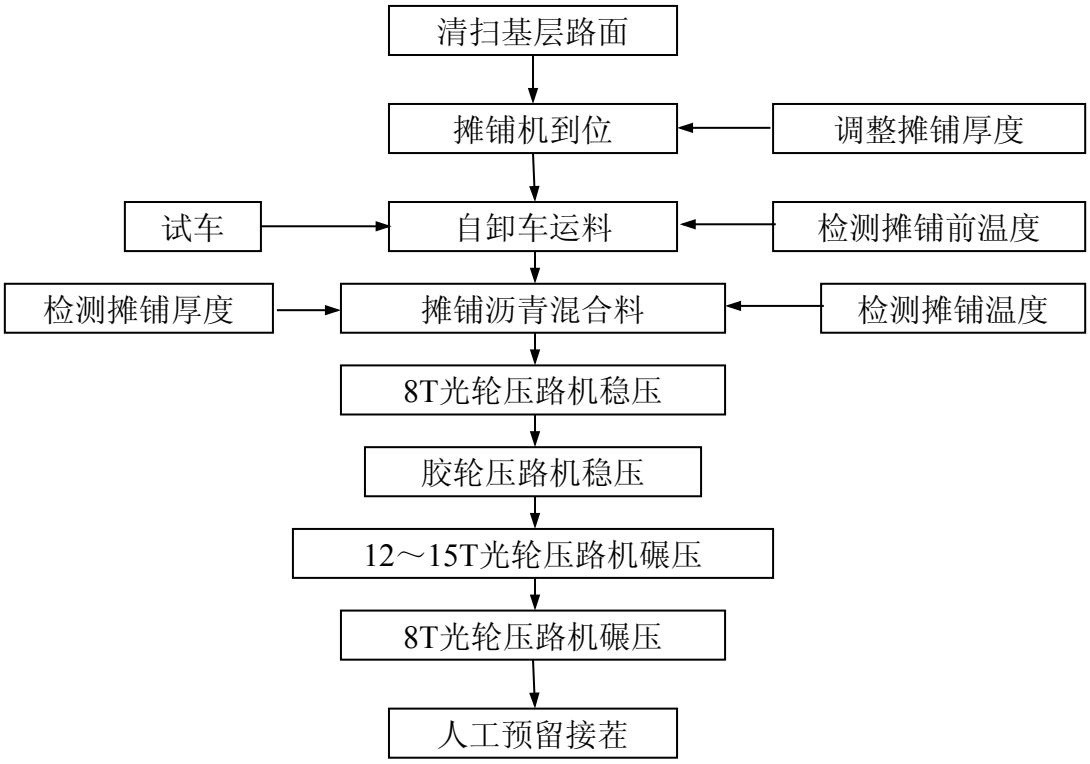
B施工程序：

测量放样→清扫基层顶面→自卸车运卸混合料→摊铺机摊铺→8T光轮压路机稳压→胶轮压路机稳压→12～1

5T光轮压路机碾压→8T光轮压路碾压→开放交通。

C施工工艺:

沥青混凝土面层施工工艺框图



D施工方法:

(1)测量放样，放出中线、两边边线，拉控制高程钢丝线。

(2)在沥青面层施工前，应将路面基层清扫干净，对有坑槽、不平整的路段应先修补和整平。

(3)用自卸车将混合料运至施工现场，摊铺机摊铺施工。边摊铺边用刮板整平，刮平时应轻重一致，往返刮2～3

次到达平整即可，不得反复撒料反复刮平引起粗集料离析。碾压原则：由低到高，由边到中，由慢到快。

(4)混合料出厂温度宜在140℃左右，摊铺温度控制在130℃左右，施工过程中要随时检测摊铺温度，以保证沥青混合料有良好的流值。摊铺过程中，如发现摊铺数量不足，有空白、缺边等应在现场技术人员的指导下，用人工补足，有积聚现象应予刮除。分两幅摊铺时，应保证接茬搭接良好，纵向搭接宽度10～15cm。半幅施工时，路中一侧宜事先设置挡板或采用切刀切齐。铺另半幅前必须将缝边缘清扫干净，并涂洒少量粘层沥青。摊铺时应重叠在已铺层上5～10cm，摊铺后用人工将摊铺在前半幅上面的混合料铲走。先在已压实路面上行走，碾压新铺层10～15cm，然后压实新铺部分，再伸过已压实路面5～10cm，充分将接缝压实紧密。表层的纵缝应顺直，且宜留在车道区画线位置上。

(5)摊铺不得中途停顿，摊铺好的沥青混合料应紧接碾压，如因故不能及时碾压或遇雨时，应停止摊铺，并对卸下的沥青混合料覆盖保温。

(6)适度的碾压对于沥青下封层路面的施工极为重要，碾压不足和过度碾压都会影响路面质量，因此在施工中，应根据矿料的等级、沥青材料的标号、施工气温等因素确定各次碾压所使用的压路机质量和碾压遍数。

E施工注意事项：

(1)施工前对各种材料进行调查试验，经选择确定的材料在施工过程中应保持稳定，不得随意变更。

(2)施工前对各种机具应作全面检查，并经调试证明处于性能良好状态,机械数量足够,施工能力配套，重要机械宜有备用设备。

(3)沥青加热温度及沥青混合料施工温度应符合施工技术规范的规定，并根据沥青品种、标号、粘度、气候条件及铺筑层的厚度选择。

(4)沥青混合料必须在沥青混合厂有采用拌和机械拌制，拌和厂的设置应符合国家有关环保、消防、安全等规定。

(5)拌和厂设置在空旷、干燥、运输条件良好的地方。沥青应分品种分标号密闭储存。各种矿料应分别堆放，不得混杂。矿料填料不得受潮。集料宜设置防雨顶棚。

(6)热拌沥青混合料可采用间歇式拌和机或连续式拌和机拌制。各类拌和机均应有防止矿粉飞扬散失的密封性能及除尘设备，并有检测拌和温度的装置。当工程材料不能连续供料或质量不稳定时，不得采用连续式拌和机。

(7)拌和厂拌的沥青混合料应均匀一致、无花白料、无结团成块或严重的粗细料分离现象，不符合要求时不得

使用，并应及时调整。

(8)热沥青混合料应采用较大吨位的自卸汽车运输、车厢应清扫干净。为防止沥青与车厢板粘结，车厢侧板和底板右涂一层薄柴油与水的混合液，但不得有余液积聚在车厢底部。

(9)运料车应用蓬布覆盖，用以保温、防雨、防污染。

(10)沥青混合料运输车的运量应较拌和能力或摊铺速度有所富余，施工过程中摊铺前方应有运料车在等候卸料。

(11)沥青混合料运至摊铺地点后应凭运料单接收，并检查拌和质量。已遭雨淋湿的混合料不得铺筑在道路上。

(12)摊铺后紧接着碾压，缩短碾压长度。摊铺过程中应随时检查摊铺层厚度及路拱、横坡，对不符合要求时应根据铺筑情况及时进行调整。

(13)沥青混合料必须均匀、连续不间断地摊铺。摊铺过程中不得随意变换速度或中途停止。摊铺好的沥青混合料应紧接碾压，如因故不能及时碾压或遇雨时，应停止摊铺，并对卸下的沥青混合料覆盖保温。

第五章　质量、安全保证措施

一、确保工程质量的措施

（1）具体质量目标：本标工程质量验收合格率100%，优良率不低于90%。

（2）质量控制和创优规划：质量管理领导小组是整个工程质量管理的最高领导，由项目经理、总工程师、工程部长、质检部长、中心试验室主任、测量组长组成，制定整个合同段工程质量创优规划、方针、措施。各施工队分别设现场质量管理小组、工长、技术主管和工班长等有关人员参加。质检组和试验室专职抓现场质量管理。施工队一级的质量管理在项目经理部质量管理小组领导下，制订本队施工区段的创优措施，质量实施计划，并在现场落实。施工队所属各班组根据自己的创优任务，拟定项目工程具体的分项实施计划，由工序检查小组和QC小组落实责任到人，严格要求，确保全员、全方位、全过程的质量控制。

（3）强化质量意识，健全规章制度

①坚持质量意识教育，落实质量岗位责任制，用工作质量保证工程质量。

②开工报告制度：

a、组织工程部和工程队有关人员编制施工组织设计，经项目经理、主管副经理或总工程师、安全员、材料员、质检工程师等签字。

b、施工组织设计必须在工程实施前10天报监理工程师。内容包括《公路监理规范》要求的开工报告的全部内容，并保证各种测试数据的真实性和各种生产力资源的足够到位。经审批，并且按审批意见进行修改完善后，方可进行施工。

③技术复核，隐蔽工程验收制度：

a、明确复核内容、部位、复核人员及复核办法。

b、技术复核结果应填写《分部分项工程技术复核记录》，作为施工技术资料归档。

c、凡分项工程的施工结果被下道工序施工所覆盖，均应进行隐蔽工程验收。隐蔽验收的结果必须填写《隐蔽工程验收记录》。

d、施工过程中应积极配合监理工程师搞好平行试验、测量验收和现场检测等工作，各工序、位置都复核，做到万无一失。

④、技术、质量交底制度：技术、质量的交底工作是施工全过程基础管理中一项不可缺少的重要工作内容,交底必须采用书面签证确认形式,具体可分以下几方面：

a、项目总工程师组织项目部全体人员对图纸进行认真学习，并同设计代表联系进行设计交底。

b、施工组织设计编制完毕并送业主和总监审批确认后，主管工程师编制操作规程，总工程师组织有关人员认真学习施工方案，并进行技术、质量、安全书面交底，列出关键部分工程和施工要点。

c、本着谁负责施工谁负责质量、安全工作的原则，各分管分项工程负责人在安排施工任务的同时，必须对施

工班组强进行书面技术质量、安全交底，必须做到交底不明确不上岗，不签证不上岗。

⑤、三级验收及分部分项质量评定制度

a、分项工程施工过程中，各分管负责人必须督促班组做好自检工作，确保当天问题当天整改完毕。

b、分项工程施工完毕后，由总工负责及时组织有关人员进行分项工程质量评定工作，并填写分项工程质量评定表交施工队长确认最终评定由项目经理部质检工程师检定。

c、项目经理部每旬组织一次全标段的质量检查，每月进行一次评定，并作为奖惩和内部结算的依据。

d、质检部对每个项目进行经常性和不定期的抽样检查，发现问题以书面形式发出限期整改指令单，施工队负责指定期限内将整改情况以书面形式反馈到工程部。

⑥现场材料质量管理制度：

a、严格控制外加工、采购材料的质量。各种地方材料、外购材料要求质量证明资料齐全，到现场后必须由质检部、中心试验室和机材部有关人员进行抽样检查，如发现问题立即与供货商联系，进行退货或换取合格材料。

b、搞好原材料二次复检取样、送样工作。

⑦计量器具管理制度：

a、工程部和中心试验室负责所有计量器材的鉴定和管理工作。

b、现场计量管理器具必须确定专人保管、专人使用。他人不得随意动用，以免造成人为的损坏。

c、全站仪、压力机等按规范要求校对的计量器具要定期进行校对、鉴定；严禁使用应校对未经校对过的量具。

⑧工程质量奖罚制度

a、遵循“谁施工、谁负责”的原则，对各施工队、班组进行全面质量管理和追踪管理。

b、凡各施工队、班组在施工过程中违反操作规程，不按图施工，屡教不改或发生质量问题，项目部有权对其进行处罚，处罚形式为整改停工，罚款直至清出本工地。

c、凡各施工队、班组在施工过程中，按图施工，质量优良且达到优质，项目部对其进行奖励，奖励形式为表扬、表彰、奖金。

d、项目经理部在实施奖罚时，以平常检查、抽查、业主大检查、监理工程师评价等方面的评定资料为依据。

（4）分部项工程质量控制：

a、严格按照业主要求、设计图纸、技术规范、监理指令施工，分项工程开工前制定书面技术要求，操作细则，施工过程中定程度、定期、定项目检查，发现问题及时整改，对完成的工程及时验收检查，作到“预防为主、过程控制、产品验收”。

b、对涵背回填等质量问题较多的项目作为专项进行质量控制，要加强技术难点和外观质量的控制，确保工程总体质量目标的实现。

c、有针对性地组织专家组咨询，达到重要项目、关键工序、控制难点施工质量稳定提高。

雨季的施工安排直接影响到整个工程是否能按期完工，现就有关雨季的施工安排叙述如下：

1、雨季施工前的准备：

雨季施工前应做以下列准备工作：

①对选择的雨季施工地段进行详细的现场调查研究，编制实施性的雨季施工组织计划。

②修好施工便道、硬化施工场地，保证晴雨畅通，雨季施工场面处的人、机械都设置防雨棚。施工电路、电

器不受雨淋、不受潮，能安全使用。

③住地、仓库、车辆机具停放场地，生产设施设在最高洪水位以上地点，并应与泥石流、沟槽冲积堆保持一定的安全距离。

④修建临时排水设施,保证雨季作业的场地不被洪水淹没并能及时排除地面水。

⑤贮备足够的工程材料和生活物资。

2、雨季施工：

①路堤填筑：

场地处理：在填筑路堤前，应在填方坡脚以外挖掘排水沟，保持场地不积水。如果原地面松软，还应采取换填等措施进行处理。

填料选择：在路堤填筑时，应选用透水性较好的碎石土、卵石土、砂砾石碎渣和砂性土作为填料。利用挖方土作填方时，应随挖随填及时压实。含水量过大无法晾干的筑路材料不得用作雨季施工填料。

填筑方法：若是土路堤应分层填筑。每一层的表面，应做成2%~4%的排水横坡。当天填筑的土层应当天完成压实。防止表面积水和渗水，将路基浸软。如需借土填筑时，取土坑距离填方坡脚不宜小于3m。

路床排水：土路堤填筑完成后，为防止路床积水，应在路肩处每隔5——
10m挖一道横向排水沟，将雨水排出路床。

②路堑开挖

场地处理：路堑开挖前在路堑边坡顶2m以外修筑截水沟，并做好防漏处理。截水沟应接通出水口。

土方开挖方法：雨季开挖路堑宜分层开挖，每挖一层应设置排水纵横坡。挖方边坡不宜一次挖到设计位置，应沿坡面留30cm厚。待雨季过后再整修到设计坡度。以挖作填的挖方应随挖、随运、随填。开挖路堑至路床设

计标高以上30—
50cm时应停止开挖，并在两侧挖排水沟。待雨季过后再挖到路床设计标高后压实。

弃土堆，雨季施工开挖路堑的弃土要远离路堑边坡顶堆放。弃土堆高度一般不应大于3m。弃土堆坡脚到路堑边坡顶的距离一般不应小于3m，深路堑或松软地带应保持5m以上。弃土堆应摊开整平，严禁把弃土堆放在路堑边坡顶上。

总之，雨季是施工抢进度的关键阶段，我单位将采取一切行之有效的方法，做到既保证工程质量，又保证工程进度。

安全预控措施

1

各专业工长在每道工序前，作好书面的质量安全技术交底，将工程施工中技术难点，质量标准及可能发生的质量问题讲清楚，施工班组长认定签字，并同时向班组全体人员传达，认真执行。

2

施工过程中，质检员跟班检测，实行现场施工全过程的质量监督，发现问题及时处理解决，把质量事故消灭在萌芽状态。实行质量监督人员质量一票否决制。

3

组织高素质的施工队伍，对特殊工种人员，例如：搅拌、养护、机械操作等主要工种执行持证上岗制来完善和强化质量管理工作。

4

加强技术管理，作到轴线、标高控制准确，组织专业测量小组用激光经纬仪、水平仪、钢尺等对柱的垂直度、柱网定位轴线和柱顶标高及预埋铁件的位置等进行施测，将测量结果报送项目技术负责人复核，报请监理工程师签字认可后，方能进行下道工序施工。

5

施工前工程技术人员必须熟悉设计图纸及有关规范，由施工员作质量技术交底，施工班组加强自检、层层把关，确保工程施工质量达到验收规范规定。

6

材料、半成品验收应符合验收规范，并按有关规定进行试(复)验，对不符合标准的材料一律不使用。

7

各工序必须按照我国《建筑施工技术操作规程》及招标文件中的有关施工技术操作规范进行施工，并接受建设主管部门、监理单位、甲方等的检查、监督。

计量检测控制措施

1

国家规定强制检定的计量检测器具必须100%按时送检，并要按时抽检。计量过程中，必须使用检定合格的计量检测器具，超过检定周期及经检验不合格的计量检测器

具均不得使用。

2

材料部门及时对水泥、钢材、砂、砖等进场消耗进行计量检测，管理好大中型材料消耗定额，做好原始记录，并对检测数据负责。

3

认真把好材料进场关，做好石料、沥青、砂等原料的进场检验，所有进场材料(成品及半成品)必须有出厂合格证，并按规定取样复检，复检合格后方可用于工程上。

4

专职质检人员应按施工顺序、质量评定标准及时做好质量检测，其量值应在规范允许的范围内。

5

贯彻以自检为基础和自检、互检、交接栓的“三检”制，层层把关，及时搞好质量等级的评定和质量验收工作，定期召开质量会议，公布各专业施工队已完工的工程质量情况，建立奖优罚劣制度。

6

搅拌站台要在运行前，须进行系统测试，制定搅拌工艺计量检测网络图。

7

施工现场设立全自动标准养护室，并做好配合比优化设

计工作。

8

试验人员每季度对实验仪器进行一次检验、维护和保养，无证人员不得使用仪器。

9

施工中严格执行隐蔽工程检查验收制，所有隐蔽工程必须按规定，经有关职能部门验收合格，报请监理工程师签字认可，方能进入下道工序的施工。

10

现场测量每季度要对所有测量仪器特别是经纬仪、水准仪进行检验校正，必须使用合格仪器。

11

计量器具必须妥善保管，非计量人员不得任意拆卸、改造、检修计量器具，认真作好计量器具的采购、入库、检定、降级、报废、保管、封存、发放等管理工作。

安全目标

1　杜绝因工亡人事故；

2　杜绝一切重伤事故；

3　轻伤率控制在5‰以内。

7 安全保证体系

贯彻“安全第一，预防为主”的方针，制定切实可行的安全保证体系，安全保证体系详见后附图。在明确对

项目经理安全责任的前提下，

按照“管生产必须管安全”和“谁主管谁负责”的原则，规定各种人员的安全责任，奖罚措施，并逐级签订安全保责任状，使安全生产贯彻落实于施工全过程。

8　　安全生产管理制度

安全生产技术交底

1

执行安全生产交底制度。施工作业前，由工长向施工班组作书面的安全交底，施工班组长签字，并及时向全体操作人员交底。

2　执行施工前安全检查制度

各班组在施工前对所施工的部位，进行安全检查，发现隐患，经有关人员处理解决后，方可进行施工操作。

3　安全教育制度

加强对施工人员的安全意识教育，提高自我防护意识，进场前对职工进行安全生产教育，以后定期不定期地进行安全生产教育，加强安全生产、文明施工的意识。

4　建立安全生产责任制

定期组织安全生产大检查，并建立安全生产评比制度，根据安全生产责任制的规定，进行评比，对安全生产优良的班组和个人给予奖励，对不注意安全生产的班组和个人给予批评，甚至处罚。

主要预报及控制措施

1

进入现场的所有人员必须带安全帽，高空作业必须系安全带，施工现场设置安全警告牌。

2

所有机电设备实行专人负责操作，持证上岗，非专业人员不得动用电器设备，供电设备要遮盖防雨，经常检修，所有移动设备均须设置漏电保护器。

3

覆带式机械负重操作时，履带下方必须垫平垫实，推土旋转半径内，严禁站人，且禁止人员通过。

4

现场施工用电严格遵照《施工现场临时用电安全技术规范》的有关规定及要求进行布置及架设，定期对闸刀开关、插座及漏电保护器的灵敏度进行常规的安全检查。用电按三相五线制架设，现场用电线路及电器安设由持证电工安装，无证人员不得操作。现场的所有机械、设备、电器须安设漏电保护器，并在每班前由持证电工进行检查。

5

加强对施工人员的防火安全教育，及对现场消防器材的管理，消防器材配备齐全，安放位置符合消防要求，定

期检查，更换灭火器的药品，保证消防器材随时处于良好状态。

6随时取得预防资料，根据气象预报，提前作好防风防雨措施，切实按措施严格执行实施，确保本工程安全控制的实现。

安全生产技术措施。

1

严禁在施工作业面上互相抛丢材料、工具等物质及向下抛丢杂物、高处作业人员衣着要简便，不准穿高跟鞋、硬底鞋、拖鞋或赤脚上班，严禁酒后作业。

2

认真贯彻落实安全帽、安全网、安全带的安全“三宝”使用要求和、通道口、预留洞口的安全防护。

3实施“施工生产安全否决权”，对于影响施工安全的违章指挥及违章作业，施工人员有权进行抵制，专职安全员有权停工，并限期进行整改。整改后经专职安全员检查合格后，方准继续进行施工。

4安排施工任务的同时，必须进行安全交底，按照安全操作规程及各项安全规定要求进行施工。安全交底应有书面资料并有交接人的签字。

5

现场内的各种防护设施和安全标语等，均不准擅自拆除

或移动，需要拆动时必须经安全负责人批准。

6严寒期间施工作业要做好防冻防滑工作。

7

施工机械设备的使用必须严格遵守《建筑机械使用安全技术规程》的有关规定，各种电器设备必须要有防雨措施，传动部位要有防护罩，并有良好的保护接地和接零，开关箱内部和顶部装钉防火板，一机一闸，闸刀熔丝不得用其它金属代替。

8

夜间施工作业配有足够的照明设施，上下施工联系采用多种类型的通讯设施，并保持通讯设施的正常使用。

9

各种气瓶要有防震胶圈和防爆防晒措施，要有明显标志，挂灭火器具。对生产必须用的气割与电焊有专人负责监护。

10

对参加施工的全体人员，做好安全教育，使各级管理人员、生产工人牢固树立“安全第一”的思想。

消防报卫措施

1

建立生产岗位防火责任制，把消防工作做到“五同时”，同计划、同布置、同检查、同总结、同评比。

2

严格执行现场用火制度，随着季度、气候工程进度的变化，因地制宜做到“四有”。

(1)有施工消防安全交底；

(2)有用火审批制度；

(3)有看火员和兼职消防组织；

(4)有消防器材和 救火措施。

3

对设置的消防水泵、消防给水管道、消防水箱和消火栓等设施，不得任意改装或挪作他用，在施工中如有冲突不得擅自变动。

4进入施工现场不得抽烟，对易燃材料要集中管理，做标记，小型工具房内不得存放汽油、煤油等易燃料。

5

电气焊工经常检查使用工具是否漏气、漏电，施焊中除清理周围易燃物外还要设专人负责看火。

安全奖罚措施

为保证施工安全，根据我单位安全奖罚办法，结合本工程特点， 特制定以下安全奖罚办法：

1

奖励：全年杜绝一切重伤及以上事故，轻伤率在5‰以下，授于安全工作先进单位称号，并发给单位安全奖10000

元，奖励主管领导1000元。

2 惩罚：发生职工因工重伤3人(含)以上，一次死亡1-2人的或造成经济损失在10-30万元的事故，给予事故直接责任人开除留用查看或开除处分，并按事故直接经济损失的3-5%进行罚款，构成犯罪的，依法追究刑事责任；对经理、书记、总工程师给予降级或撤职处分，并按事故直接经济损失的2~3%进行罚款。

第六章　　使用新技术、新工艺的可行性

本工程拟推广应用的新技术、新工艺、新材料、计算机应用和管理技术等。

计算机应用及管理技术

本项目拟配备电脑3台，项目施工后，从施工技术交底、验工计价、施工预决算及技术资料整编到项目对外文件往来及项目进度计划管理、人员管理等方面均采用电脑管理。

第七章　主要材料构配计划

根据现场施工管理体系，形成以主任工程师把关，生产组制订各分项工程所需材料、构配件计划用量，报技术组、项目经理审查。

开工前材料、构配件准备工作计划

7.1　熟悉、审查设计资料，准备图纸会审。

7.2　委托进行材料、试块(件)试验。

7.3　统计委托加工木制品和其它半成品。

序号	材料名称	单位	数量
1	标准砖	千块	33.75
2	镀锌钢管DN100	M	28.84
3	水泥	KG	25416.09
4	黄土	M3	11022.979
5	水	M3	3599.95
6	水泥花砖5*25*25cm	块	12215.52
7	水泥花砖6*12*24cm	块	98015.96
8	中粗砂	M3	470.36

9	PVC100	M	2800
10	混凝土路缘石	M	3200
11	乳化沥青	Kg	18300
12	中粒式沥青混凝土	T	969.9

第八章　主要机械设备供应计划

本项目工程主要机械设备的配置，以满足总工期保证施工进度为前提，以作业面的需要及便于调配为原则，充分发挥施工机械设备的交替，同时考虑一定的富余量，具体配备情况详见下表：（见后附表）

机械设备供应计划

序号	机械、设备名称	规格、型号	单位	计划数量	备注
1	推土机	QTZ-5513	台	2	
2	铲车	HBT60	台	2	
3	装载机	ZB-21	台	3	
4	压路机	SCD200/200	台	2	
5	沥青搅拌机	JS-350	台	1	
6	蛙式打夯机	HW-25	台	4	

7	冲击式打夯机	HC-70	台	2	
8	柴油发电机	50KW	台	1	
9	气焊工具		台	3	
10	电动试压泵	4BA-8A	台	2	
11	倒链	5t	个	2	
12	接地电阻测试仪	ZC-8	个	1	

第九章　　劳动力安排

依据本工程的特点，结合我单位的实际情况，施工人员精选我单位技术熟练，经验丰富的技术工人上场，配属一部分合同工及临时工作为补充，详见主要劳动力配备计划表。

主要劳动力配备计划表

序号	工种	高峰人数
1	土工	30
2	拌合工	40
3	沥青工	10

4	普工	30
5	电工	10
6	机械工	2
7	管道安装工	6

第十章　文明施工措施

我单位在施工中将尽最大限度维护原来的地貌地形，保持原来的生态环境，在施工中，从以下几方面加强文明施工管理：

确保文明施工的技术组织措施

项目经理部必须树立“外理市场，内抓现场，以创建文明工地强化现场管理保护”的思想。使全体职工认识到，现场是企业的窗口，效益的源泉。施工过程中必须以现场管理为重心，推行创建文明工地活动，使管理职能落实到现场，管理人才流向现场，技术进步渗透到现

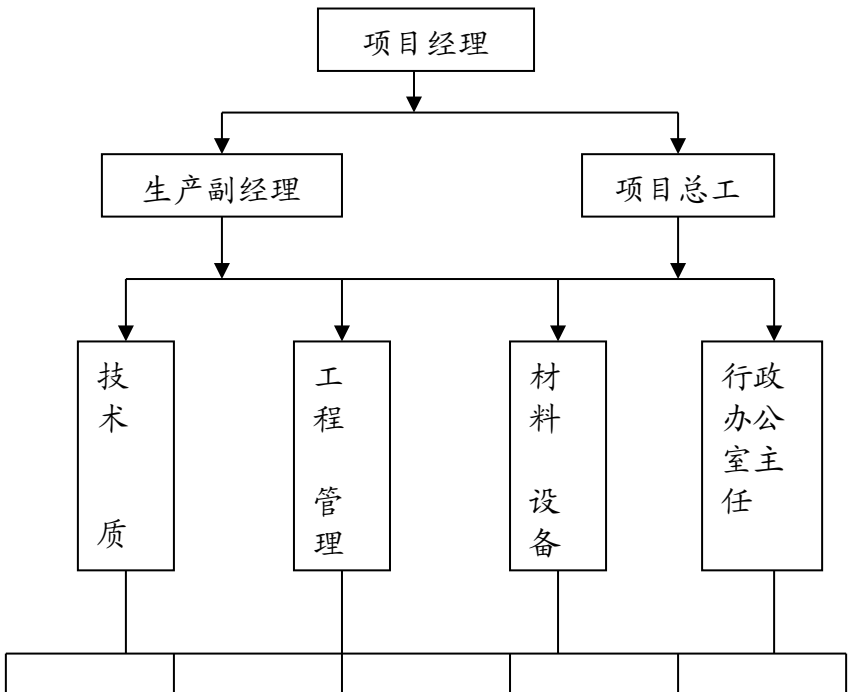
场，严抓进度、安全生产管理，促进施工现场管理水平稳步提高。

制定文明工地目标，建立文明施工管理体系，认真地落实创建文明工地责任，发扬我公司在本地区创立省、市级文明工地建设的先进经验和创建文明工地的优良传统，把要工种创建文明工地活动推向更高水平。

第一节　文明施工目标

本工程文明工地施工总目标：确保市级文明工地，争创省级文明工地。

第二节　　文明施工管理体系



组建以项目经理为组长的创建文明工地领导小组，项目部副经理和项目技术总负责为副组长，设立创建文明工地办公室主任，负责处理、协调创建文明工地的日常事务工作。

第三节　文明工地施工措施

一、施工现场管理规范

工地四周围墙稳固，统一刷白色外墙涂料，工地大门制作安装规范，大门砖柱水泥砂浆抹灰后刷白色外墙涂料。

大门内一侧布置尺寸统一的“六牌一图”和“项目经理部主要管理人员责任制”标牌，标牌内附照片、姓名和职务。设立固定的阅报栏并落实专人每天更换报纸。

场区内临时工地道路全部用砼硬化并做好排水沟，大门口设置冲洗台，运输车辆外运必须冲洗后方可放行，

防止污染城市道路。

材料堆场用砼硬化并找好坡度，防止雨水积水，施工现场材料物资按平面布置图堆置，砂子等材料堆置应成方见垛，机砖和空心砖应码垛见方，钢材和设施料应放整齐有序。

施工层应做到操作中随用随清，工完后人走场清。建筑垃圾和生活垃圾分开定点堆放，并设专人负责及时清运出场外。

施工现场内的中、小型机械设备均搭设防雨棚，机械的标识、编号清楚醒目，操作规程和操作负责人标牌制作规范，悬挂位置合理。

制定措施控制施工噪音大的机械设备，以避免干扰四周单位和居民休息。

消防管理制度健全，责任落实，措施到位，安全警示牌醒目，有针对性的制定防火，防爆和防毒预防措施。

施工现场内严禁用明火和吸烟。

二、施工安全达标

安全管理工作到位，经常组织安全生产竞赛活动，认真执行安全交底，加强安全日巡查和安全检查。有针对性的制定安全预防措施，排除安全隐患。

加强职工安全培训和安全教育，特种作业岗位必须持证上岗。

施工用电“三宝”和“四口”以及机械设备安全是重点检查的内容，检查应及时产做好记录。

三、工程质量创优

施工现场有微机管理，提高施工现场的现代化管理水平，配备必要的办公设备。

加强现场计量管理，砌筑砂浆拌制全部用电子磅秤控制。

实现质量目标，创建精品工程。

四、办公室设施整洁

办公室等临边设施搭建规范，整齐，室内安装暖气及吊扇，室内通风、采光、照明、卫生间要符合规定。

制定办公室“公约”和管理制度。

现场建水冲厕所，由专人负责保持卫生，安装排气扇。

大门口内外以及场区内地面的卫生安排2—3人，负责每天上班前，中午和下午班后打扫干净，并负责保持。

围墙边、宿舍前空闲地面，栽种花草搞好绿化。

五、良好的文明氛围

制定文明职工公约并自觉遵守，由行政管理组长负责组织经常性的娱乐活动，开展争做文明职工的竞赛评比活动，改善职工的精神面貌。

建立职工花名册，做到人数清，情况明。

现场设卫生治疗室，卫生室内一般药品急救药品量足，品种齐全，有1名专职卫生员。

现场开辟黑板报专栏，专人负责搞好黑板报，利用黑板报进行法制宣传和安全文明施工教育，使职工知法，懂法和提高职工创建文明工地的积极性。

由项目部治安员负责现场门卫安全值班管理，负责组织夜间现场巡逻，制定现场保卫制度和建立门卫登记制度并认真落实。

六、补充实施措施

工地设钢制大门双开，进大门右侧设门卫室，墙上挂铝合金框以保持大门内外的整洁。

大门出入口设置高压洗车台，凡出门车辆必须冲洗轮胎，方可出门，以保持大门内外的整洁。

在橱窗背面的内墙面，用水泥砂浆粉刷1.2×1.8米一块做黑板宣传栏。

进大门右侧的门卫室和围墙临近处设温饮水茶桶一只，茶桶边设有消毒水盆和引水杯，施工条件成熟时，在工地西部再设同样的茶水处一座，均由门卫用电水壶烧开水供应。

采用二层外廊式活动房为工地现场办公室，做鲜明规范的CI图案，并规范地安装办公室、会计室门牌和铁合金室内岗位责任制。

办公室内设一专人管理的保健箱，公司的医生定期到工地指导工作。

凡进入施工现场的施工人员必须佩戴公司统一制定的岗位证上岗。

按施工组织设计，用电要规范的架空或埋地，进三相五线制线路进入工地二级上级配电箱。

按施工组织设计的施工平面布置图位置整齐堆放施工材料、构件、料具，并分规格、名称、品种插上标示牌。

按施工组织设计，施工现场的主要道路和施工地坪要硬化处理，排水通畅。如排水需要经过硬化道路或硬化地面处，可在硬化前预先埋设过水管道或先做暗沟，以保证路面或地面无积水。

开工前向业主或有关部门了解场内有无城市管线埋设通过，如有，施工时必须采取切实有效的保护措施，施工完后，在竣工图中不应把所施工范围内的所有新、老管网位置标入图内，加以警示，边缘维修。

占用人行道前办理占道手续，施工期间及时清扫，不得污染路面。

施工场内的排水系统需设置必要的沉淀池，产定时清理，保持通畅。

近招标单位规定，工程主体施工一定使用商品砼，以

保证施工质量。

工地外架一律设密目式安全网全封闭，通道外设双层防砸棚，必须牢固、安全、可靠，三有施工废物，必须及时清理。

楼上工完料清的建筑垃圾，不得向下抛撒，砖头、钢筋头、砼头、木头及生活垃圾应装车由提升机上运下，并分类堆放，插上标牌，及时清运处理，以保持工地整洁。

工地要建立义务消防队，组织培训，掌握工地消防知识。防火的重点部位，如宿舍、食堂、配电室、木工场、仓库等应设置必备的灭火器材。工地动用明火，必须要有动火审批手续和防火措施，实施时要的专人看护，并不准在工地内燃烧产生有害气体的废弃物。

不在工地建设宿舍、食堂生活设施，所有工人均住在场外，对工人宿舍委派专人管理。宿舍内整齐安设双层铁架床，通风良好，宿舍外排水通畅、有绿化。

食堂内墙面全部刷白，厨房、灶台等地上2米高范围内贴白瓷砖。食堂工作人员持健康证，穿工作服上班，墙上挂“食堂管理制度”和食堂卫生标准”牌，操作必须符合《食品卫生法》。食堂内外经常冲洗，整洁卫生，设防蝇罩，并定期灭蚊、灭鼠。

建筑工地食宿处设厕所和澡堂，厕所内墙面贴白色瓷

砖，做简易平顶，要求有良好的通风，澡堂设沐浴，墙面1.8米以下及地面贴白瓷砖，设专人管理。

环境保护技术组织措施

（1）重视环境工作：为确保文明施工，促进施工顺利进行，我单位将采取以下环境保护措施。完善施工组织设计时，把环境工作作为施工组织要求组成部分，并认真贯彻执行施工的全过程。

（2）加强环保教育：组织职工学习环保知识，加强环保意识，使大家认识到环境保护的重要性和必须性。

（3）贯彻环保法规：认真贯彻各级政府的有关水土保护、环境保护方针、政策和法令，结合设计文件和工程地点，搞好环境保护。

（4）强化环保管理：定期进行环境检查，及时处理违章事宜，主动联系环保机构，请示汇报环保工作，做到文明施工。

（5）美化施工现场：场地废料、土石废方处理，应按设计要求或按监理工程师指定地点处理，防止水土流失，保持排水通道畅通，工地干净卫生。施工中还应尽量减少对周围绿化环境的影响和破坏。

（6）消除施工污染：施工废水、生活污水不得污染水源、耕地、农田、灌溉渠道。工地垃圾及时运往指定地点深埋。

3、承诺：

如果我单位有幸中标，我们将严格按本投标书所提供的人员设备迅速进场组织施工，通过和业主、监理的密切配合，在按合同工期完成任务的同时，保证优良质量目标的实现，力争本工程质量达到国内同期同类工程的先进水平。不论是我单位还是业主方的因素可能导致工期滞后和工期紧张时，将在2天内从单位调遣技术力量，在3天内从单位调遣补充劳力资源，确保目标的实现，发扬完成抢险或突击任务的精神，以回报建设单位的信任。