电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER DISSERTATION

论文题目 基于自然语言处理的自动文摘系统

学科专业 计算机应用
指导教师 杨国纬 教授
作者姓名 张 峰
班 学 号 200320604065

分类号_____

UDC[注1]_____

学 位 论 文

基于自然语言处理的自动文摘系统

（题名和副题名）

张 峰

（作者姓名）

指导教师姓名 杨国纬 教 授

电子科技大学 成 都

（职务、职称、学位、单位名称及地址）

申请专业学位级别 硕士 专业名称 计算机应用

论文提交日期 2006.1 论文答辩日期 200.6.3

学位授予单位和日期 电子科技大学

答辩委员会主席 卢显良 [illegible]

评阅人 卢显良 [illegible]

2006 年 1 月 5 日

注 1：注明《国际十进分类法 UDC》的类号。

独 创 性 声 明

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

签名：张峰　日期：2006年 1月6日

关于论文使用授权的说明

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

（保密的学位论文在解密后应遵守此规定）

签名：张峰　导师签名：杨国桢

日期：2006年 1月 6 日

摘　要

在本文中，首先介绍了自然语言处理的基础概念体系，给出了自然语言处理的定义及其研究和处理的方法和过程，接着便介绍国内外关于自动文摘系统等方面的研究方向和发展动态，并指出了自动文摘系统研究的某些不足。然后重点说明了文摘和自动文摘系统的基本概念体系，并针对目前几种主要的自动文摘系统形式化模型和方法：基于统计的机械文摘、基于理解的文摘、基于概念依存的文本结构分析方法和信息抽取的文本摘要等模型和方法进行了比较和分析，对它们的优点和缺点进行了讨论，归纳出各自的特点。进而在总结各种不同类型的自动文摘系统的特点的基础上，将基于潜在语义分析和篇章多级依存结构的文摘方法相结合，提出了一种综合型的自动文摘系统的设想。

潜在语义分析(Latent Semantic Analysis，LSA)是一种用于自动地实现知识提取和表示的理论和方法，它通过对大量的文本集进行统计分析，从中提取出词语的上下文使用含义。在技术上，它同向量空间模型类型类似，都是采用空间向量表示文本，但通过 SVD 分解等处理，消除了同义词、多义词的影响，提高了后续处理的精度。

篇章多级依存结构分析（Text Multilevel Dependency Structure，TMDS）是一种基于结构的自动文摘分析方法。如果把各个部分视为节点，并在两个有语义联系的部分之间引一条边，那么我们就得到了一个关联网络。它清楚的表示了文章的整体结构；同时篇章结构比语言表层结构深入了一大步，根据篇章结构能够更准确地探测文章的中心内容所在，因而基于篇章结构的自动文章能够避免机械文摘的许多不足，保证文摘质量。

本文提出的文摘方法综合利用了两种方法。首先通过对文本进行潜在语义分析，对文本矩阵进行相应的奇异值分解，重构语义矩阵；然后采用基于篇章多级依存结构的文摘分析方法，对重构的语义矩阵表示的文本内容进行深入的分析，抽取重要的句子生成文摘，这样就弥补了潜在语义分析在词法和句法分析上的不足；同时过滤和去除了语义噪音，缩小了问题的规模。

关键词：自然语言处理，自动文摘，潜在语义分析，篇章多级依存结构

ABSTRACT

In this thesis，the author first introduces the latest development of Automatic Summarization System in domestic and abroad, which shows the lack of the automatic summarization system research. Then the author introduces some basic concepts about automatic abstract system. Secondly, some basic concepts about Abstract and automatic summarization system are introduced, and the main formal models and methods of system are compared and analyzed, such as statistics based, meaning based, concept based, knowledge based etc. We induce their characteristics and put forward a kind of comprehensive automatic summarization system based on latent semantic analysis and text multilevel dependency structure.

Latent Semantic Analysis (LSA) is a completely automatic theory and method of the acquisition and representation of knowledge, which extracts the contextual-usage meaning of words by statistical computations applied to a large corpus of text. LSA is similar to Vector Space Mode (VSM), representing textual materials with space vectors. LSA can advance the accuracy of subsequent processes by using a truncated Singular Value Decomposition (SVD) to remove the influences of synonymy. In this paper, the authors introduce the basic ideas, characters and implementations of LSA，and discuss the applications based on LSA.

Text Multilevel Dependency Structure (TMDS) is one kind of method used in automatically realizing to withdraw and expression the knowledge. If regards each part as the pitch point, and draw a line in two parts that are semantically relate to another one, then we obtained a connection network. It clear expression article overall construction; At the same time the text structure penetrated a stride compared to the language surface structure, can accurately survey the central content of one article according to the chapter structure. Thus, the automatic summarization based on the text structure can avoid much shortage of the mechanical digest, guarantee digest quality.

A new text summarization method is proposed. It process documents not only based on latent semantic analysis, but also based on text multilevel dependency structure. The method first analysis the latent semantic structure of texts, make single value decomposition on text-matrix, reconstruct the semantic matrix; then a method based on text multilevel dependency structure is adopted, deeply analysis the content of the semantic matrix, abstract the important sentences to generate Abstraction and make up the shortage of latent semantic analysis on structure and syntax.

Keywords: natural language process, text summarization, latent semantic analysis, text multilevel dependency structure

目　　录

第一章 绪论

二十一世纪以来，随着科学技术的发展，人类从工业化社会步入了信息时代，计算机已经成为这个时代标志性的产物，人类对信息的处理也提出了更高的要求。

由于以 Internet 为主题的信息高速公路的不断普及和发展，信息技术已经渗透到我们社会生活的各个角落，正以前所未有的速度和能力改变者人们的生活和工作方式，人们正处在一个“信息爆炸”的时代。一方面，因特网上蕴涵着海量信息远远超过人们想象；另一方面，面对信息的海洋，人们往往束手无策，无所适从。计算机文摘系统被人们认为是信息资源处理的有效的手段之一，它为人们快速浏览信息确定自己的兴趣点提供了有力帮助，设计更准确有效的中文自动文摘方法已经成为热门课题，它渗透到计算应用的各个方面，市场前景十分巨大，并显示出强劲的势头。

1.1 国内外研究现状

目前，文摘系统的研究已有很多，但缺乏统一的理论基础和定义，评价方法也很多；但是随着人们对文摘系统研究的重视，必将朝着更广泛、更深入的方向发展。

1.1.1 自然语言处理的基础

随着社会的日益信息化，人们越来越强烈地希望用自然语言同计算机交流。自然语言处理是计算机科学中的一个引人入胜的、富有挑战性的课题。从计算机科学特别是从人工智能的观点看，自然语言处理的任务是建立一种计算机模型，这种计算机模型能够给出像人那样理解、分析并回答自然语言（即人们日常使用的各种通俗语言）的结果。

现在的计算机的智能还远远没有达到能够像人一样理解自然语言的水平，而且在可预见的将来也达不到这样的水平。因此，关于计算机对自然语言的理解一般是从实用的角度进行评判的。如果计算机实现了人机会话，或机器翻译，或自动文摘等语言信息处理功能，则认为计算机具备了自然语言理解的能力。

自然语言处理[1]（NLP，natural language processing）与自然语言理解（NLU，natural language understanding）是同义词，都是人工智能的一个分支，就是

研究如何能让计算机理解并生成人们日常所使用的(如汉语、英语)语言，使得计算机懂得自然语言的含义，并对人给计算机提出的问题，通过对话的方式，用自然语言进行回答。目的在于建立起一种人与机器之间的密切而友好的关系，使之能进行高度的信息传递与认知活动。自然语言理解系统可以用作专家系统、知识工程、情报检索、办公室自动化的自然语言人机接口，有很大的实用价值。

自然语言理解研究在电子计算机问世之初就开始了，并于 50 年代初开展了机器翻译试验。当时的研究方法还不能称作带有“智能”。到了 60 年代乔姆斯基的转换生成语法得到广泛的认可，生成语法的核心是短语结构规则，分析句子结构的过程就是利用规则自顶向下或自底向上的句法树生成过程。

由于认识到生成语法缺少表示语义知识的手段，在 70 年代随着认知科学的兴盛，研究者又相继提出了语义网络、CD 理论、格框架等语义表示理论。这些语法和语义理论经过各自的发展，逐渐开始趋于相互结合。到 80 年代一批新的语法理论脱颖而出，具有代表性的有词汇功能语法(LFG)、功能合一语法(FUG)和广义短语结构语法(GPSG)等。

这些基于规则的分析方法可以称之为自然语言处理中的“理性主义”。现有的手段虽然基本上掌握了单个句子的分析技术，但是还很难覆盖全面的语言现象，特别是对于整个段落或篇章的理解还无从下手。

与“理性主义”相对的是“经验主义”的研究思路，主要是指针对大规模语料库的研究。语料库是大量文本的集合。计算机出现后，语料可以被方便地存贮起来，利用计算机查找也很容易。随着电子出版物的出现，采集语料也不再成为困难。最早于 60 年代编制的 Brown 和 LOB 两个计算机语料库，分别具有 100 万词次的规模。进入 90 年代可以轻易列举出的语料库有几十个之多，像 DCI、ECI、ICAME、BNC、LDC、CLR 等，其规模最高达到 109 数量级。

我国自然语言理解的研究起步较晚，比国外晚了 17 年。国外在 1963 年就建成了早期的自然语言理解系统，而我国直到 1980 年才建成了两个汉语自然语言理解模型，都以人机对话的方式来实现。八十年代中期，在国际新一代计算机激烈竞争的影响下，自然语言理解的研究在国内得到了更多的重视，"自然语言理解和人机接口"列入了新一代计算机的研制规划，研究单位增多了，研究队伍也壮大了。

自然语言处理中，如何研究众多词汇之间的关系，是一项庞大的语言工程，需要大量的人力，物力和财力。

美国普林顿大学认知科学实验室的米勒和贝克威斯等人，于 1985 年开始致力于建造词汇关系网络的工作，建成了词网(WordNet)[2]。词网是由心理学家和计算

机科学家共同建造的一部在线的英语词汇参照系统，它是根据词义而不是根据词形组织词汇信息，可以说是一部基于心理语言学原理的语义词典。WordNet 目前有近 95600 个不同的词性，其中包括 51500 个简单词和 44100 个搭配词，这些词构成 70100 个词义（或者说同义词集合）。它与其他标准词典最显著的不同在于 WordNet 将词汇分成五个大类：名词、动词、形容词，副词和虚词。实际上，WordNet 仅包含名词、动词、形容词和副词。虚词通常是作为语言句法成分的一部分，WordNet 忽略了英语中较小的虚词集。词网中的名词按层次关系组织，形容词按 N 维超空间的方式组织，动词按推演关系组织。

我国学者董振东研制了“知网”（Mind-Net）[3]。知网是一个词典知识描述系统，描述的词汇包括汉语和英语两种语言，这两种语言是相对独立的，他们词语与词语之间的对应是建立在相同的属性描述的基础上的。目前，知网有汉语词汇 33069 条，英语词汇 38774 条。

计算机语言学中分析自然语言的方法主要有两种：一种是基于规则的方法，一种是基于统计的方法。实践证明，这两种方法各有千秋，尽管基于统计的方法对于大规模真是文本的处理比较合适，但是，我们决不能忽视基于规则的方法。基于规则的自动句法分析的理论和方法主要有断语结构语法、语言串分析法、递归转移网络和扩充转移网络、通用句法分析器和线图分析法、范畴语法、链语法、依存语法和配价语法，管辖和约束理论、词汇功能语法、功能合一语法、中文信息 MMT 模型、蒙太谷语法、广义短语结构语法、中心语驱动的短语结构语法、定子句语法等。这些理论和方法都是计算机语言学中经常使用的，它们是学习和研究时应具备的计算机语言学的基础知识。

自然语言的计算机处理，除了进行词法分析和句法分析之外，还要进行语义分析[4]。

关于语义分析和句法分析的关系，在现有的自然语言处理系统中还有不同的处理办法，有的系统采用“先句法后语义”的办法，有的系统采用“句法语义一体化”的办法。

所谓“先句法后语义”，就是在自然语言的分析系统中，首先进行独立的句法分析，得到表示输入句子的句法表示式，然后在经过独立的语义分析，获得输入句子的语义表示式。在句法分析中，虽然也要利用附加在词和词组的上的某些必要的语义信息，但主要的依据式词法和句法信息。这一系统的程序设计不依赖于某个特定的领域，具有较好的可移植性和可扩展性。

所谓“句法语义一体化”，是指在自然语言分析系统中，不单独设置一个句法

分析模块，而是句法分析和语义分析并行，或者根据某些语义模型，直接从输入句子求出其语义表示式，这一类系统往往可以有效的处理某些有语法错误或信息不全的句子，根据语义线索直接获得对句子的语义解释，但是，由于句法信息不充分，语义分析往往难于奏效。

不论采用哪一种办法，语义分析都是必不可少的。所以，语义分析和句法分析一样，他们都是自然语言处理的最基本的功能模块[5]。

中国的计算机语言学不可避免的要研究汉字和汉语的自动处理问题。

用计算机对汉语进行自动分析，面对的问题和困难印欧语系的语言如英语、俄语、德语、法语等要多一些，除了自然语言自动分析研究面对的共性问题外，汉语理解还有自己的特点和难点，这些特点和难点主要表现在语言的词汇分析，结构分析和语义分析等三个方面：

一、 有关词汇分析上的关键问题：

1. 汉语的书面书写形式是连续书写的，词与词之间没有自然的界限。因此汉语的自然语言理解首先要解决单词的自动切分问题，而自动切词中，交集型歧义和组合型歧义的判定以及未登陆词的处理，这些都是比较困难的。
2. 汉语常用词多为兼类词，兼类词的判定就是自动词性标注，这是汉语句法分析不可避免的问题。
3. 汉语的名词没有明显的标志，翻译成英语时难于判定其是单数还是复数，必须根据汉语句子中的显性标志或隐性标志来判断对应以汉语名词的英语名词是否用复数形式。
4. 汉语中的动词没有明显的时态标志，其时间的表达是隐性的。
5. 汉语中存在“离合动词”，分析时往往要把相关的成分连接起来，机器词典要做特殊处理。
6. 汉语形容词的比较方式多样，而这些词义本身又可能具有其他的功能和含义，从而形成兼类或歧义。
7. 汉语的量词特别丰富，有名量词和动量词之分。量词的分析和判断也是汉语自动理解中的一个难题。
8. 汉语的名词、动词、形容词、数词、量词都有重叠式，不同词类的重叠式功能不一样，其含义也不尽相同，不同的含义需要计算机作特殊判断，分析时需要做分别处理。

二、 有关结构分析方面的关键问题：

1. 汉语既无词尾形态标记，又基本上没有形态变化，大多数汉语的实词本身不能明确的表达语法意义，因此，许多语言学家认为汉语的词法结构中的语法关系主要考词序列和虚词来表示。然而事实并非如此，仅仅靠词序和虚词不能解决汉语语法结构的形式描述问题。
2. 汉语中名词词组结构复杂，如不加“的”限定性定语，分析时就常常出现结构歧义和非语法形式的问题[6]，给名词词组的识别带来困难。
3. 联动词和兼语式是汉语的两种特殊句式，在这样的特殊句型以及由多个动词构成的句子中，由于若干个动词或动词词组相互连接时没有明显的形式标志，主要动词淹没在一大堆动词中，形成 $VP_1+VP_2+VP_3+\cdots\cdots+VP_n$ 的格式[7]，计算机往往难以确定其中的主要动词，而如果主要动词判断有误，整个结构的分析必定失败。在兼语式中，兼语又做主语，又做宾语，使得句子中除了原来主语之外，又出现了一个兼做宾语的新主语。因此，句子出现一个以上的主语，和印欧语中传统的“主语＋谓语”那样的一个主语和一个谓语单纯的相互结合的句式有很大不同，也给自动语法分析带来极大困难。
4. 汉语句子中常常出现省略主语的现象，使得句子中主语成分的确定变得非常困难。
5. 汉语句子的被动句往往不用被动式表示，主动形式和被动形式无明显差别，受事主语句是汉语中的普遍句式，在处理中极易主动被动关系颠倒的错误。
6. 汉语中存在主谓谓语句，在处理时容易导致结构层次的混乱。
7. 汉语中形容词作谓语时往往不用“是”，汉英翻译时必须作特殊处理。
8. 汉语中名词可以作直接谓语，这样往往使作谓语的名词同做主语的名词紧密地连在一起，使主语和谓语之间的界限十分模糊，从而背计算机错误地判断为“名词＋动词”的名词词组，极易造成分析的失败。
9. 汉语中有“把字句”，具有处置的含义，分析时首先确定“把”字管辖的范围，这个界限往往不容易确定。
10. 汉语中的紧缩句是由复合句紧缩而成的，汉语分析时，只有把紧缩句拆分成复合句，这是比较复杂的工作。
11. 汉语时一种分析型语言，语义分析在汉语研究中起着举足轻重的作用，在进行处理时要给予足够的重视。
12. 汉语常用词多义想象普遍，多义词的判断规则十分繁琐，难于发现一

般性的规则。

13. 汉语的基本句式“主－谓－宾”结构与英语相似，都是 NP+VP+NP，表层结构的分析并不困难，但是，表层的句法结构远远不能满足汉英机器翻译的需要，词与词以及词组与词组之间的句法关系和语义关系才是问题的核心。
14. 汉语中存在大量的歧义现象，歧义是自然语言的计算机处理面临的一个严重问题。

上述两个方面的 22 个问题产生的根源在于汉语本身不同于印欧语言的特点。从根本上说，汉语具有以下明显的不同于印欧语言的五个特点：

第一，汉语缺乏印欧语言那样丰富的形态。

第二，汉语的语素、单词和词组之间的界限模糊。

第三，汉语的词类和他们的句法成分之间没有明确的一一对应的关系。

第四，汉语的句子成分和语义关系之间也没有明确的一一对应关系。

第五，汉语书面语没有分词连写，使得自动分词成为汉语自动分析的一个特殊问题，由此而引起的汉语书面文本的分词歧异、未登陆词的处理，成为汉语自动分析的又一大难点。

1.1.2 自然语言处理的发展态势

我国的有关科研单位和专家，从来没有停止过攻克中文信息处理难关的努力，在国家的几个科学攻关计划中都列有信息处理项目。这些项目都是以解决计算机对自然语言进行理解问题，也就是以开发智能型的汉语分析系统为奋斗目标。当前这类研究基本上都是在语料——主要是词——的统计概率的基础上进行的。许多专家已经感觉到，统计概率的路已经走到尽头，必须另辟蹊径，这“ 蹊径”就是语义，以词义为基础，与句法规则结合，以句为突破的单位。朝着这个目标努力，到目前为止，正在进行的众多研究项目，大体可以分为三种风格，或者说是三种思路、三个流派。

1.传统计算语言学

第一个流派是以传统计算语言学为基本理论，从词素分析入手，进而研究词-短语（词组）——语段——句子。概括地说，传统计算语言学的种种理论和方法，都以语料统计为基础。但是，只靠统计概率是不能统摄复杂多变的语言现象的，

因此还需要结合语言规则。为此，我国学术界从西方计算语言学的众多理论和方法中吸收了许多营养，例如短语结构语法、扩充转移网络、从属关系语法和配价语法等。由许嘉璐主持的国家社会科学“九五”重大项目“信息处理用现代汉语词汇研究”的立项和进展或许可以说是当前这一领域研究最集中突出的例子。这一课题是纯粹的基础性研究,而且应该说,单就词汇领域而言,它也还是不够完整的基础性研究。虽然这个课题是中文信息处理技术所需要解决的重要问题,但是要把这样的一些成果集成，形成可供使用的平台，还需要做相当艰苦的努力；而且即使这些成果集成了，也只是为今后的研究奠定一定的基础，因为要让计算机理解汉语的词，还需要解决词的意义如何概括、表达（ 用计算机可以“读”的符号）等问题。在解决了词的问题之后，才能进入句的领域，而要解开句子的奥秘，除了要弄清楚句子结构规律，关键问题也是语义。到目前为止，包括“信息处理用现代汉语词汇研究”在内的研究还没有正式进入意义领域。而对语义，以统计概率为主要方法，是难有作为的。在已有成果的基础上今后应该走什么路？国内外都还没有找到有十分把握的途径。

2.HNC 理论

出于对传统研究方法（词→短语→句→句群→篇章）是基于西方语言而建立的，其总体与汉语实际不适应的考虑，黄曾阳先生提出了概念层次网络理论（HNC）。HNC 理论认为，计算语言学界源于图灵标准而采用的句法分析和句法语义分析所提出的标准各有偏低和偏高的不足，不是描述人的语言感知过程的适当模式，因为“思维的机制绝不是语法或句法，而是概念联想网络的建立、激活、扩展、浓缩与存储”，从而提出计算机对汉语的处理不应该以图灵检验为标准，而应该以对语言模糊的消解能力为第一标准。“自然语言的语句呈现出无限和不确定的表现特征，……在其背后是否存在一种有限和确定的语句结构？人们对此进行过多方面和多层次的探索。”传统语言学、乔姆斯基理论、依托于数理逻辑理论的句法语义分析、依托于隐马尔科夫模型等的各种统计处理，各有自己的答案。对自然语言特性的把握必须是微观和宏观并重的，对语句特性的把握更是如此。上述四种答案“与语言微观和宏观特性的联系大体依次呈现出反变和正变的趋向。”HNC 的答案是：应该描述语言感知过程，为此，应从语言的深层入手，以语义表达为基础，把自然语言所表述的知识划分为概念、语言和常识三个独立的层面，建立语义完备性的概念表述数学表示式和语句的语义表述模式。人的语言交流过程，就是消解“模糊”的过程。因此，HNC 把消解模糊作为自然语言理解初级阶段的标准（就书面而言，有词的多义模糊、语义块构成的分合模糊、指代冗缺模

糊），即以消解模糊为攻克的第一步。HNC 认为，汉语以“字义基元化，词义组合化”方式构造新词，因此可以构建概念表述体系，亦即概念层次网络。HNC 同时认为自然语言无限的语句可以用有限的句类物理表示式来表达。“语句的宏观特性可以用语句的句类表示式来表达，语句的微观特性可以用语义块的构成表示式来表达。” 他们据此设计了局部联想网络解决词汇层面问题，设计了全局联想网络解决句类和语义块问题（句类是语句的语义类型）。据 HNC 课题组的研究，自然语言共有 7 个句类：作用、过程、转换、效应、关系、状态和判断。每个句类有自己的句类表示式，基本句类表示式共 57 个。此外，自然语言还有单个全局特征语义块的混合类和两个或多个全局特征语义块的复合句类。理论上二者应有 3192+57×56+57×3192+3192×3192=10377192 种。但是常见的混合句类只有理论值的十分之一左右，在计算机上是可以解决的。从理论上说，HNC 的句类分析是对大脑语言感知过程的初步模拟，应该接近人的语言过程，但这需要长时间的逐步完善。现在 HNC 课题组正按照“语义块感知和句类假设、句类检验、语义块构成分析”三步曲策略努力工作，期望能研究和开发出具有“自知之明”（即能在译文疑点处自动做出标记并给出多种解决方案）的汉/英互译系统。这一策略在规模较小的知识库支持下已经取得可喜结果。而其准确性最终要在更大的知识库支持下，在大规模语料中运行才能得到信度较高的印证。这正是这一课题当前工作的难点所在。

3.基于内涵模型论的语义分析

这一流派的理论设计，是陆汝占教授提出的。其出发点是考虑到对中文信息处理的研究单纯走语法的路已经难以有突破性结果，归根结底，要深入到语义层面。朝着意义精细方向考虑，就会产生兼类过多和概括力不够以及歧义、模糊、不确定等困难。换言之，汉语表达式的意义仅指称外延对象，没有涉及内涵性质，因此存在一个语句中的同一词语表达式的多个出现，都指称相同的外延对象。怎么办呢？应该在一个逻辑句义框架下来分析词汇及其分类，只要能明白表达句义，不必过于精细，也就是用逻辑框架来处理词汇理论。基于这一考虑，该课题组将汉语表达式抽象成数学表达式，恰当地表示内涵和外延义，然后把这些语义表示在计算机内进行处理，亦即把汉语表达式与计算机数据结构之间直线联结，改变为汉语表达式——抽象数学表示——数据结构三者的间接联结。课题组称之为基于形式方法——模型论的汉语语义计算理论。根据这一理论设计，句义分析的流程为：语句→切分→标注→句法分析→句法树→同构的语义树→逻辑公式→模型解释。显然，从“切分”到“句法树”，与受图灵检验启发而进行的研究一致；其

特色就在于建立“同构的语义树”，特别是进入“逻辑公式”并做出“模型解释”。陆汝占教授认为，语句要转换成逻辑式，应从汉语语句谓语动词结构着手。因为句法分析是语义分析的前提，句法分析又要靠语义特征。具体设想是：先构造一种句子的逻辑式之间的中介形式“函子”（functor），以表示谓语动词连同支配成分一起构成的语句核心，表现句义的基本要素。函子加上时态、模态算子就可以表示语态，构成句子的基本逻辑含义。对于计算机自动处理中文信息来说，汉语的缺省（省略和隐含）都是难点。因此基于内涵模型论的理论对这一点格外重视，但是至今除了利用上下文语境知识外，也还没有找到很好的解决策略，而语境知识的形式化也是十分复杂的问题。说从“切分”到“句法树”和现在通常的解决方法一致，只是就总体和顺序而言，实际上基于内涵分析的语义解释理论对于“词”、“句”等有着自己的理解。

以上三个流派都正在进行过程中，进展情况不一。第一个流派，不同单位和个人已经在一些局部取得了较好的成绩，面临着如何集成和如何解决词义、句子问题；第二个流派设想和计划比较庞大，在规模不够大的知识库内，已经得到部分技术实现，面临着继续扩大知识库、进行相当于“中试”或一定规模生产的过程，以便检验和完善其理论和技术设计；第三个流派，理论设计还较粗略，虽然用这一理论已经解决了一些实用问题，但是要证明它可以适用于整个现代汉语，还需要进一步推敲、实验、细化。值得注意的是，第二、第三种思路都很重视我国传统语言学，特别是训诂学的经验和成果，或从中得到启发，或借用其对词语的训释。这是有道理的。

1.1.3 自动文摘系统的研究状况

自动文摘系统早在 50 年代末，Luhn 首次设计了一个自动文摘系统[8]，引起了世人的极大关注。六年后，Luhn 发表了一篇阐述自动生成文摘方法的论文，这片论文连同他在情报检索领域中所作的其他工作一起，为后续的同类工作提供了一个开端。

自动文摘系统的研究大体上可分为两个阶段：第一阶段是从 50 年代末到 70 年代初的机械文摘时期；第二阶段是从 70 年代初到现在的理解文摘时期。所谓机械文摘是指以文章的结构、词频等知识提取出文摘，而理解文摘是指对文章的内容，从句子到结构的理解中提取出文摘来。

经过 20 年的机械文摘的研究，发现文摘受词典结构和启发函数设计的限制，

导致文摘质量不高。于是，人们采取保守的态度，对受限领域的文本进行语法语等分析，以及一些推理手段。成功的例子并不多，即使有成功的例子，也只能处理受限的文本。

这一时期的代表人物是 Luhn，他在 1958 年发表的关于自动编制文摘的文章，开创了自动文摘的样板性研究方案。Luhn 把词汇分成两大类[9]：通用词和内容词。通用词又称为功能词，通常包括连接词、代词、介词、冠词、助动词，以及某些形容词和副词，除此以外的所有词为内容词。功能词重要性被指定为 0，词频统计只对内容词进行，并把同根的内容词加以合并，词频超过某一事先设定的阈值 V 的内容词被认定是可以代表文章主题的有效词。

美国的 Baxendale 采用 3 中方法从文章中选词和词串：删除功能词、从论题句中选择内容词、从正文的介词短语中选词。她发现段落的论题是段落首句的概率为 85%，是段落末句的概率为 7% 。因此，有必要提高处于特殊位置的句子的权值；他还认为“介词短语似乎比任何其他简单的语言结构更能密切地反映文章的内容”，应该立足于词组和词串，而不是孤立的单词。

目前的许多自动摘录系统都综合考虑了两种或多种形式特征。比如，新加坡南洋(Nanyang)大学研制的图书馆新闻删节系统(Library Newspaper Cutting System) ，提供了题名法、位置法、关键词法和指示性短语法 4 种自动摘录方法供用户选择。但是在多种特征的结合方面尚待深入研究。

1997 年,日本的 Tadashi Nomoto 等人提出了一种基于语料库的自动摘录方法。他们将一批文献分为训练集和测试集，对训练集中每篇文献内的每个句子自动建立一个包括“该句在文本中的位置(Location in Text)”、“与标题相近似的程度(Similarity to Title)”等属性的属性集，并人为地将句子分为两类，一类是文摘句，另一类不是文摘句。然后在训练集的基础上建立统计模型，对测试集中每个句子是否能够作为文摘句的判别问题，将转化为依据决策树对该句进行分类的问题[10]。这种设计思想是在语料库语言学的影响下提出的，它让计算机自动地从训练集中提炼各个特征的结合函数，为多种形式特征的综合利用开辟了一条新的道路。但是，由于它毕竟是建立在文本表层的形式特征基础之上，缺乏深度，所以发展潜力将受到限制。

70 年代末 80 年代初，美国耶鲁大学的 Schank 在脚本的基础上研制了 SAM(Script Applier Mechanism)系统。该系统应用脚本分析简单的故事，在此基础上对故事进行总结。

美国耶鲁大学的 DeJong 于 1979 年研制了著名的 FRUMP(Fast Reading

Understanding and Memory Program)系统，该系统用于快速阅览英文新闻资料，是理解文摘系统的样板。

FRUMP 由预言器和验证器两部分组成。预言器利用梗概剧本预测当前情形下可能出现的一个或一组事件，验证器的任务是去证实这些被预测的事件，并给出实际信息。FRUMP 的应用范围受内部存储的梗概剧本的限制，如果文章中没有该系统所期望的内容则无法生成任何摘要，会有被误导，以致望文生义的可能[11]。美国 J. I. Tait 的 Scrable 系统对 FRUMP 系统进行了改进，它要求输入的资料在处理前先转换成 CD(Conceptual Dependency Structrue)结构，在此基础上分析和确定被预测的信息与未预测的信息之间的关系，并将这两部分信息合理地组织成一篇完整连贯的文摘。然而由于 CD 结构过于复杂，所以实现起来困难较大。

80 年代末，美国 GE 研究与开发中心的 Lisa F. Rau 等研制了 SCISOR 概念信息缩写、组织和检索系统(System for Conceptual Information Summarization ，Organization and Retrieval)。SCISOR 属于典型的理解文摘，它处理的对象是有关“公司合并”的新闻报导。SCISOR 首先采用关键词过滤和模式匹配的方法对待处理文献进行主题分析，以便判定该报道的内容是否与“公司合并”有关；然后采用与领域无关的自底向上的分析器 TRUMP(TRansportable Understanding Mechanism Package)识别每个句子的结构，生成类似于框架(Frame)的概念表示；最后运用自顶向下的预期驱动的分析器 TRUMPET(TRUMP Expectation Tool)从概念表示中提取预期的内容[12, 13]。80 年代初，德国康斯坦茨大学的 Hahn 等人研制了 TOPIC 系统，该系统针对微处理器领域的科技文献，以框架作为知识表示的基础，通过对全文的语法语义分析生成不同长度的摘要。

基于理解的文摘方法需要对文章进行全面的分析，生成详尽的语义表达，这对于大规模真实文本而言是很难实现的。与之相比，信息抽取(Information Extraction)只对有用的文本片段进行有限深度的分析，其效率和灵活性显著提高。

信息抽取的自动文摘以文摘框架(Abstract Frame)为中枢，分为选择与生成两个阶段。文摘框架是一张申请单，它以空槽的形式提出应从原文中获取的各项内容。英国 Lancaster 大学 Paice 等人在 1993 年提出的选择与生成文摘法，实质上就是信息抽取方法。目前该系统主要针对“小麦实验”方面的文章，但研究者们正在努力使它能够方便地移植到其它领域。由于文摘框架的编写完全依赖于领域知识，所以信息抽取仍然是受领域限制的，只不过文摘框架比理解文摘中的脚本等要简单得多，更易于编写。信息抽取要相应用于多个领域，就必须为每个领

域都编写一个文摘框架，在处理文本时先进行主题识别，根据主题调用相应的文摘框架。另外，单凭特征词或特征短语的提示作用来填充文摘框架并不是非常准确的，而且由于语言的灵活多样，一些有价值的文本片段可能没有明显的特征。最后，由于使用模板生成文摘，使得文摘的语言千篇一律，十分呆板。

后来，人们在研究中发现篇章是一个有机的结构体，篇章中的不同部分承担着不同的功能，各部分之间存在着错综复杂的关系。篇章结构分析清楚了，文章的核心部分自然能够找到。如果将一个语言单元的各个子单元视为节点，并在两个有语义联系的子单元之间引一条边，那么我们就得到了一个关联网络。在网络中，与一个节点相连的边数称为该节点的度。节点的度越大，则节点在网络中的重要性越高。将最重要的若干子单元抽取出来，即可构成文摘。

前苏联的 E. F. Skoroxod′ko 将文章视为句子的关联网络，句间的关系建立在词间的同义关系基础之上，和很多句子都有联系的中心句被确认为文摘句[14]。美国 Cornell 大学的 Salton 等人则将文章视为段落的关联网络。文献中的每个段落被赋予一个特征向量，两个段落特征向量的内积作为这两个段落的关联强度。如果两个段落的关联强度超过给定阈值，则认为两个段落有语义联系。和很多段落都有联系的中心段被提取出来组成一篇文献摘要。对于篇幅较长的文章，句子之间的关联网络将十分庞大，其时空开销都将是难以承受的。相比之下，段落之间的关联网络要小得多。另外，和由句子组装起来的文摘相比，由段落拼接起来的文摘连贯性显著提高。不过，由于最重要的段落中也可能包含一些无关紧要的句子，所以基于段落抽取的文摘显得不够精练。

我国从 1985 年开始介绍国外自动文摘方面的研究情况，从 80 年代末开始研究自动文摘实验系统，至今也有 10 余年的历史了。在形式特征方面，汉语和西文主要区别是汉语词间没有空格，因而存在着自动分词问题。汉语自动分词是一项经过多年研究仍未圆满解决的难题，以致于南京大学信息管理系的李明提出了从汉字频率统计出发提取文摘的权宜之计，从而回避自动分词问题。然而因为汉语中真正负载信息的是词而不是字，所以如果分词技术能够满足大规模真实文本处理的需要，那么以词为基础的自动文摘必然优于以字为基础的自动文摘。实际上，大多数中文文摘系统都要对文本进行分词处理，只是由于采用的分词方法不同，使得分词精度有所不同。此外，汉语的词汇极为丰富，同一个概念可以用很多不同的词汇表达，这给词频统计带来了一定的困难。上海交通大学王永成教授从 80 年代末就开始研究自动摘录技术，1997 年研制了 OA 中文文献自动摘要系统。该系统集成了位置法、指示短语法、关键词法和标题法等多种方法，是一个实用的系

统。哈尔滨工业大学王开铸教授于1992年研制了基于自然语言理解的文摘实验系统MATAS，1994年研制了自动摘录类的HIT2863Ⅰ型自动文摘系统。1996年提出了基于信息抽取和文本生成的自动文摘方案，1998年完成了基于篇章多级依存结构的HIT2863Ⅱ型自动文摘系统。

近两年来，从事这项研究的单位不断增加。北京邮电大学信息工程系钟义信教授采用的文摘方法类似于Paice的选择与生成文摘法，目前主要针对计算机病毒方面的文章。山西大学郭炳炎教授也在开展自动文摘的研究，他们采用基于统计的方法分析文本结构。据悉，IBM中国研究中心和大陆微软公司都在研制中文自动文摘的产品。

1.2 本课题的研究背景、目的与意义

1.2.1 本课题的研究背景

自动文摘系统的研究已成为一个国际性大课题，世界上许多国家的科学工作者都投入到文摘生成系统的研究之中。对于中文文摘系统的研究，也成为汉语主要使用国家的重大研究项目，特别是我国大陆、台湾、香港以及东南亚地区，研究自动文摘的系统的学者也越来越多。

中文自动文摘系统的研究已经成为我国二十一世纪的重点项目，在当前呈现“知识爆炸”现象的信息社会中，知识更新更快，老化快。据有关资料估计，科学技术信息的年增长率将达到40％以上，每5～8年增长10倍。信息的发觉、管理、开发、利用和更新将成为提高社会劳动生产率的主要手段。

目前，自动文摘的研究还远远不能满足信息社会对信息处理的需要，关于它的理论研究在我国开展的时间也不长。产生这种现象的主要原因无论是基于统计、理解、结构、还是基于信息抽取的文摘方法生成的文摘质量并不令人十分满意。

为此，我们研究的思路和目的是：首先对整篇文本进行分句、分词的预处理，在此基础上进行语义模型分析；然后采用篇章的结构分析来确定文档中各个句子间的深层语义关系；在线性加权两种分析方法的分析结果确定句子的最终权值，从而选择文摘句生成摘要；最后采用指代消歧技术对生成的文摘进行平滑处理，使生成的文摘更加流畅，减少冗余。

1.2.2 本课题的研究目的与意义

研究是在文本词汇的计算机处理、自动语义分析和自动句法分析已完成的基础上，进行潜在语义分析和篇章多级依存结构相结合的方法提取文章文摘句的研究，消除两种方法的弊端，发挥他们各自的优势，从而有效地提高了文摘的质量。

研究拟解决的关键问题是如何解决文本的奇异值分解、重要句子的提取和篇章多级依存结构中句子的加权。在文章中通过对文本进行分词、分句处理后，采用潜在语义分析的方法对线性语义项进行奇异值分解，过滤掉噪音，提取出重要的语义信息。将基于潜在语义分析和基于篇章多级依存结构的两种研究方法相结合，通过对文本的奇异值分解，提取重要语义信息，在此基础上进行篇章机构分析的提取算法的描述和讨论，对中文自动文摘进行了深入的研究，为进一步讨论一种综合型的中文自动文摘系统打下基础

第二章 文摘和自动文摘系统

自动文摘技术自它产生之日起，就受到广大科学工作者的重视，经过 40 多年的研究与发展，形成了一定的理论和方法，自动文摘技术得到不断的发展和完善，越来越多的公司、企业从事自动文摘技术的研究和开发，彼此的交流与合作也在增多，自动文摘技术已成为二十一世纪中国十大科技突破口之一。本章通过对文摘的定义、自动文摘的方法的介绍，使人们对自动文摘系统有一个全面的认识。

2.1 文摘的定义

随着互联网的普及、信息获取途径的增加，每天都有不断涌现的海量信息。为了从这些海量信息中快速、准确地获取有用信息，文档的自动摘要处理变得越来越重要。通过阅读文摘而不是全文能极大地加快信息过滤速度，帮助人们了解概况或确定是否应详读原文。

国际标准《文献工作—出版物的文摘和文献工作》中的文摘是指：“一份文献内容的缩短的精确的表达而无须补充解释或评论，且对写文摘的人来说没有差别。”

美国国家标准学会《文摘编写标准》中给出文摘的一个定义：某一文献内容简要而准确的表述，不加解释和评论，也不区分这片文摘是谁写的。

在我国国家标准《文摘编写规则》中，文摘被定义为：以提供文献内容梗概为目的，不加评论和补充解释，简明、确切地记述文献重要内容的短文。

在这三个定义中，都提出了“不加评论和解释”的要求，即对文摘的客观性的要求。

Johnson(1970)研究发现:尽管有时人们对于文章中句子的重要程度也有不同意见，但是人们对于文章中哪些是最重要的以及哪些是最不重要的部分则意见非常一致，然而对于哪些不是非常重要的则分歧就比较大。根据这个现象，获得“理想摘要”方法的立论依据就是:如果多个专家对同一篇文献所做出的摘要满足一定的同一率，则可以用专家们的共同意见来代表这篇文章的最重要部分，构成“理想摘要”。否则，得到的“理想摘要”的可靠性就值得怀疑。这种方法最大的问题就在于“理想摘要”很难确定获得。如果文章的风格内容比较特别，使得多名专家不易做出一致的决定，从而导致专家们的赞同率较低，则从根本上就否定了利

用这种方法的可行性。另外，目前对文摘系统还出现了一种日益强烈的需求，即要求文摘系统可以提供面向用户具有个性化的摘要，个人输入自己的需求从而得到不同的摘要结果。这种个性化的摘要就更不能用多位专家的共同意见来代表。

譬如，对于江泽民总书记的一个报告，有些领导分管农业，有些分管工业，他们所要求的摘要就有分工的偏重，而不是笼统地为一般人看的纯主题的摘要。而我们认为摘要是准确客观全面的反映某一文献中心内容的简洁连贯的短文。

实际上，文摘的确难于被准确定义，国际著名的模糊数学大师在讨论自动摘要问题时也认为文摘也难于被准确定义。文摘在中文中也称摘要、概要、提要、梗概、简介等，在英文中则有 summary、brief、compendium、epitome 等，使用什么术语并不重要，只要摘出的内容满足客观的要求即可。

2.2 自动文摘方法

自动文摘的研究已经走过 40 年历史，它的价值充分暴露出来。1993 年 12 月在德国 Wadern 召开了历史上第一次以自动文摘为主题的国际研讨会。1995 年，国际期刊 Information Processing & Management 出版了一期题为 SummarizingText 的专刊，编者在序言中指出，这一期专刊的出版标志着自动文摘的时代已经到来。到目前为止，自动文摘有 4 种主要的方法：基于统计的自动文摘、基于理解的自动文摘、信息抽取和基于结构的自动文摘[15]。

2.2.1 基于统计的自动文摘

基于统计的自动文摘将文本视为句子的线性序列，将句子视为词的线性序列。它通常分 4 步进行：(1)计算词的权值；(2)计算句子的权值；(3)对原文中的所有句子按权值高低降序排列，权值最高的若干句子被确定为文摘句；(4)将所有文摘句按照它们在原文中的出现顺序输出。

在基于统计的自动文摘中，计算词权、句权、选择文摘句的依据是文本的 6 种形式特征:

1. 词频(Frequency)。能够指示文章主题的所谓有效词(Significant Words)往往是中频词。根据句子中有效词的个数可以计算句子的权值，这是 Luhn 首先提出的自动摘录方的基本依据。V.A.Oswald 主张句子的权值应按其所含代表性“词串”的数量来计算，而 Doyle 则重视共现频度最高的“词对”。美国 IBM 公司在 1960 年前后研制了一套文摘自动生产程序 ACSI2Matic，

该程序在句权的计算方面对 Luhn 的方法进行了改进。1995 年美国 GE 研究与开发中心的 Lisa. F. Rau 等人完成了 ANES(Automatic News Extraction System)系统，该系统采用相对词频作为词的权值[16]。

2. 标题(Title)。标题是作者给出的提示文章内容的短语，借助常用词表(Stoplist)，在标题或小标题中剔除功能词或只具有一般意义的名词，剩下的词和原文内容往往有紧密可以作为有效词。
3. 位置(Location)。美国的 P. E.Baxendale 的调查结果显示：段落的论题是段落首句的概率为 85%，是段落末句的概率为 7%[17]。因此，有必要提高处于特殊位置的句子的权值。
4. 句法结构(Syntactic Structure)。句式与句子的重要性之间存在着某种联系，比如文摘中的句子大多是陈述句，而疑问句、感叹句等则不宜进入文摘。
5. 线索词(Cue)。Edmundson 的文摘系统中有一个预先编制的线索词词典，词典中的线索词分为 3 种：取正值的褒义词(Bonus Words)，取负指的贬义词(Stigma Words)，取零值的无效词(Null Words)。句子的权值就等于句中每个线索词的权值之和。70 年代初，俄亥俄州立大学的 James A. Rush 教授和他的学生开发了 ADAM(Automatic Document Abstracting Method)系统。ADAM 强调的是排斥句子的标准而不是选择句子的标准，词控表(WCL)中大多数词是否定性的。
6. 指示性短语(Indicative Phrase)。1977 年，英国 Lancaster 大学的 Paice 提出根据各种“指示性短语”来选择文摘句的方法。和线索词相比，指示性短语的可靠性要强得多。

上面我们介绍了文本的 6 种形式特征，即：F 词频、T 标题、L 位置、S 句法结构、C 线索词、I 指示性短语。这 6 种特征是自动摘录的依据，它们从不同角度指示了文章的主题，但都不够准确，不够全面。如果能够将上述各种特征“有机”地结合起来，即以 W =f(F，T，L，S，C，I)作为计算句子权值的公式，那么摘录的质量可望进一步提高。问题的关键在于函数 f 如何确定。

Edmundson 用一个简单的线性方程 $W = a_1C + a_2K + a_3T + a_4L$ 将 4 种基本的句子选择方法集成在一起。W 代表句子的最终权值，C 代表线索词(Cue)权值，K 代表根据词频计算而得的关键词(Key)权值，T 代表题名词(Title)权值，L 代表位置(Location)权值，a_1 、a_2 、a_3和 a_4是调节参数。这种将不同性质的因素简单地线性叠加的方式缺乏充分的理由，实践表明确实不够理想。在经过 10 年的探索之后，

Edmundson 断言：今后的文摘自动化方法必须考虑文献正文的句法特征和语义特征，而不能简单地依赖粗糙的统计数据。

2.2.2 基于理解的自动文摘

基于理解的文摘方法是以人工智能，特别是自然语言理解技术为基础而发展起来的文摘方法。这种方法与自动摘录的明显区别在于对知识的利用，它不仅利用语言学知识获取语言结构，更重要的是利用领域知识进行判断、推理，得到文摘的意义表示，最后从意义表示中生成摘要。

一.基于理解的自动文摘通常有以下步骤:

1. 语法分析。借助词典中的语言学知识对原文中的句子进行语法分析，获得语法结构树。
2. 语义分析。运用知识库中的语义知识将语法结构描述转换成以逻辑和意义为基础的语义表示。
3. 语用分析和信息提取。根据知识库中预先存放的领域知识在上下文中进行推理，并将提取出来的关键内容存入一张信息表。
4. 文本生成。将信息表中的内容转换为一段完整连贯的文字输出。

二.篇章意义的机内表示：

篇章意义的机内表示是原文分析的结果和文摘生成的依据，它在基于理解的文摘系统中处于中枢地位。不同系统采用的篇章意义机内表示有所不同。

1. 脚本。70 年代末 80 年代初 ，美国耶鲁大学的 Schank 在脚本的基础上研制了 SAM(Script Applier Mechanism)系统。该系统应用脚本分析简单的故事，在此基础上对故事进行总结。美国耶鲁大学的 DeJong 于 1979 年研制了著名的 FRUMP(Fast Reading Understanding and Memory Program) 系统，该系统用于快速阅览英文新闻资料，是理解文摘系统的样板。FRUMP 由预言器和验证器两部分组成。预言器利用梗概剧本预测当前情形下可能出现的一个或一组事件，验证器的任务是去证实这些被预测的事件，并给出实际信息。FRUMP 的应用范围受内部存储的梗概剧本的限制，如果文章中没有该系统所期望的内容则无法生成任何摘要，会有被误导，以致望文生义的可能。
2. 概念从属结构。美国 J. I. Tait 的 Scrable 系统对 FRUMP 系统进行了改进，它要求输入的资料在处理前先转换成 CD(Conceptual Dependency

Structrue)结构，在此基础上分析和确定被预测的信息与未预测的信息之间的关系，并将这两部分信息合理地组织成一篇完整连贯的文摘。

3. 框架。80年代末，美国GE研究与开发中心的 Lisa F. Rau 等研制了 SCISOR 概念信息缩写、组织和检索系统(System for Conceptual Information Summarization，Organization and Retrieval)。SCISOR 属于典型的理解文摘，它处理的对象是有关“公司合并”的新闻报导。SCISOR 首先采用关键词过滤和模式匹配的方法对待处理文献进行主题分析，以便判定该报道的内容是否与“公司合并”有关；然后采用与领域无关的自底向上的分析器 TRUMP(TRansportable Understanding Mechanism Package)识别每个句子的结构，生成类似于框架(Frame)的概念表示；最后运用自顶向下的预期驱动的分析器 TRUMPET(TRUMP Expectation Tool)从概念表示中提取预期的内容。
4. 一阶谓词。意大利 Udine 大学的 Danilo FUM 等人在 80 年代初研制了 SUSY(SUmmarizing System)缩写系统，该系统以一阶谓词逻辑为基础，取得了较好的效果，体现出了逻辑方法的潜力。SUSY 系统由两部分组成。第一部分称为纲要生成器(Schema Builder)，它收集用户需求，形成摘要纲要(Summary Schema)和文本纲要(Text Schema)。第二部分包括分析器(Parser)和缩写器(Summarizer)。分析器自底向上地分析原文，建立起一阶谓词形式的机内表示。缩写器首先使用文本纲要和加权规则产生加权的内部表示，然后使用摘要纲要和选择规则修剪加权的内部表示，最后从输入文本中检索出基本单元(词、短语或整句)，将它们装配成摘要[18]。

2.2.3 信息抽取

基于理解的文摘方法需要对文章进行全面的分析，生成详尽的语义表达，这对于大规模真实文本而言是很难实现的。与之相比，信息抽取(Information Extraction)只对有用的文本片段进行有限深度的分析，其效率和灵活性显著提高。

信息抽取的自动文摘以文摘框架(Abstract Frame)为中枢，分为选择与生成两个阶段。文摘框架是一张申请单，它以空槽的形式提出应从原文中获取的各项内容。例如，针对计算机病毒类的文章可以提出如下的框架:

病毒{病毒名称；
病毒传染对象；
病毒类属；
病毒攻击对象；
}

在选择阶段，利用特征词从文本中抽取相关的短语或句子填充文摘框架。例如，在文本中发现“…感染可执行文件…”字样，则可以将特征词“感染”后面的短语“可执行文件”作为病毒的感染对象填入文摘框架。

在生成阶段，利用文摘模板将文摘框架中的内容转换为文摘输出。文摘模板是带有空白部分的现成的套话，其空白部分与文摘框架中的空槽相对应。例如“该病毒的感染对象是(病毒传染对象)”是模板中的一个句子，因为在文摘框架中登记的病毒感染对象为“可执行文件”，因此在文摘中将输出这样的句子：“该病毒的传染对象是可执行文件”。

2.2.4 基于结构的自动文摘

篇章是一个有机的结构体，篇章中的不同部分承担着不同的功能，各部分之间存在着错综复杂的关系。篇章结构分析清楚了，文章的核心部分自然能够找到。但是语言学对于篇章结构的研究还很不够，可用的形式规则就更少了，这使得基于结构的自动文摘到目前为止还没有一套成熟的方法，不同学者用来识别篇章结构的手段也有很大差别。

1. 关联网络。如果将一个语言单元的各个子单元视为节点，并在两个有语义联系的子单元之间引一条边，那么我们就得到了一个关联网络。在网络中，与一个节点相连的边数称为该节点的度。节点的度越大，则节点在网络中的重要性越高。将最重要的若干子单元抽取出来，即可构成文摘。
2. 修辞结构。90 年代初，日本 Toshiba 公司的 Kenji Ono 等基于修辞结构研究自动文摘。他们将修辞关系归纳为举例〈EG〉、原因〈RS〉、总结〈SM〉等 34 种，首先依据连接词等推导出一种类似于句法树的修辞结构树，然后对修辞结构树进行修剪，将保留下来的内容根据它们之间的修辞关系组织成一篇连贯的文摘[19]。这种方法的不足在于：修辞关系的识别依赖于连接词，如果文章中连接词的数量很少，那么数修辞关系就无法识别了。
3. 语用功能。这种方法主要是针对科技文献。科技文献的写作有比较严格的

规范，文献中不同部分承担着不同的语用功能，根据语用功能可以将文章的主体部分识别出来构成文摘。日本北海道大学的 Maeda 将句子的信息功能分为：背景(B)、主题(T)、方法(M)、结果(R)、例子(E)、应用(A)、比较(C)和讨论(D)，并认为 T、M、R 和 D 是主干，应进入文摘；E、A、C 和 B 是枝叶，应排除在文摘之外[20]。美国纽约 Syracuse 大学的 Liddy 通过对人工文摘的大量调查归纳出经验文摘(Empirical Abstract)的基本结构：背景——目的——方法——结果——结论——附录，其中每一项内容中又包括了一些细则。如果将文摘中承担这些功能的片段识别出来，就可以组成文摘。和用其它方法生成的文摘相比，根据语用功能提炼出来的文摘更符合科技文献文摘编写的标准。如果想把这种方法推广到科技文献以外的文本中去，则需要对各类文章的结构深入研究。其实即使是科技文献也有各种类型，理论文章、实验文章和综述文章的结构区别也很大。

2.3 本章小结

综上所述，自动文摘的这几种主要方法各有千秋。自动摘录能够适用于非受限域，这符合当前自然语言处理技术面向真实语料，面向实用化的总趋势，但是由于它局限于对文本表层结构的分析，所以经过近 40 年的发展已接近技术极限，文摘质量很难再有质的飞跃。理解文摘牺牲领域宽度，换取了理解深度，它作为理论探索的价值很高，但实用性较低，在可预见的未来中前景暗淡。信息抽取能够通过重编文摘框架使文摘系统适应新的领域，它的文摘质量并不比理解文摘逊色，而适用范围更宽了。基于结构的自动文摘不涉及领域知识，所以它和自动摘录一样是面向非受限域的，同时由于引入了篇章结构方面的知识，使文摘质量有了较大的提高。结合两种文摘方法的优点，提出一种综合的文摘系统是目前自动文摘方法研究的趋势。我们在本章所讨论的文摘和自动文摘方法的基本概念体系以及国内外发展动态的基础上，将在下一张讨论潜在语义分析和篇章多级依存结构分析在自动文摘中的应用。

第三章 基于 LSA 和 TMDS 的自动文摘

在本章中，我们要介绍基于潜在语义分析（LSA）和篇章多级依存结构（TMDS）的自动文摘方法技术产生的背景、基本思想、特点和实现方法。

3.1 潜在语义分析

当前的信息检索技术，都是通过关键字硬匹配来实现的，而汉语的表达方式是灵活多样的，一个概念可以有多种不同的表达方式，不同语境下，一个词也可能表达有多种不同的意思。采用硬匹配技术必然会导致检索精度的下降，从而降低整个系统的性能。要改变这种现状，就要充分考虑到语义和语境信息，潜在语义分析技术的引入，有效地解决了同义词和多义词的问题，通过识别文本中的同义词，可以将信息检索效率提高 10—30％[21]。

3.1.1 潜在语义分析产生的背景

传统的信息检索模型，如布尔模型、概率模型和向量空间模型，无论采用哪种模型，都是基于关键字硬匹配进行检索的，这样就可能产生词的同义和多义现象。所谓词的同义现象是指，不同的用户根据个人的需要、所处的环境、知识水平以及语言习惯等不同，对同一事物的表达方式也不一样，实验表明，对于同一事物，用相同词语表述的用户不到 20％，这样就导致用户的查询与文本索引项表面上不一致，但实际上两者却是匹配的，造成了漏查现象，使检索的查全率大大下降。查全率是返回的相关文本在文本集中的比率，它是衡量检索系统的一个很重要的性能指标。所谓词的多义现象是指相同的词在不同的语境中表达的意义并不相同，这样就使得用户查询和文本索引项表面上一样，但两者却并不相关，这样就把本来无关的文本作为检索到的相关文本返回给了用户，造成了查准率的下降。查准率是相关文本在返回给用户的检索结果中的比率，查准率是衡量检索系统性能的另外一个重要指标，查全率和查准率两者综合考虑，反映系统性能。词的同义和多义的存在给文本处理带来了极大的不便，解决此类问题成为文本处理领域的焦点。通常解决词的同义现象采用词义扩充或基于电子词典的方法，也就是提供附加的索引项参加匹配运算。这种方法的弊端是附加的索引项可能含有多个词义从而在提高查全率的同时使查准率大大降低。解决词的多义现象的一个常

见的方法是利用受控的词汇表和人工介入作为转换机制，但是其代价昂贵，效果也无法令人满意；另外一种方法是利用布尔交集来消歧，结果也不理想。

除了上述的同义和多义现象，传统的信息检索还存在一个问题，那就是其技术多是基于 G.Salton 的向量空间模型 VSM 研制的，而 VSM 的基本假设条件是各个分量的相互正交，但是自然语言中词汇和短语多具有一定相关性，很难满足这个假设条件，这也使检索结果受到很大的影响。

正是为了更好地解决上述问题，20 世纪 90 年代 T.K.Landauer，S.T.Dumais 提出了一个新的检索模型－潜在语义分析，简称为 LSA(Latent Semantic Analysis)[22]。LSA 可看作是一种扩展的向量空间模型，它利用统计计算导出的概念索引进行信息检索。LSA 基于这样一种断言：即文本库中存在隐含的适用的语义结构，这种语义由于部分地被文本中词的语义和形式上的多样性所掩盖而不明显，LSA 通过对原文本库的项/文本矩阵的奇异值分解计算，并取前 K 个最大的奇异值及对应的奇异向量构成一个新矩阵来近似表示原文本库的项/文本矩阵，由于新矩阵削减了项和文本之间语义关系的模糊度，从而有利于信息检索。

3.1.2 潜在语义分析的基本思想

潜在语义分析（LSA）是一种用于知识获取和展示的计算理论和方法[23]，出发点就是文本中的词与词之间存在某种联系，即存在某种潜在的语义结构。这种潜在的语义结构隐含在文本中词语的上下文使用模式中。因此采用统计计算的方法，对大量的文本中进行分析来寻找这种潜在的语义结构，它不需要确定的语义编码，仅依赖于上下文中事物的联系，并用语义结构来表示词和文本，达到消除词之间的相关性，简化文本向量的目的。

一.项/文本矩阵

在潜在语义分析 LSA 模型中，一个文本可以表示成一个 m×n 的项/文本矩阵 A,这里 n 表示文本中的句子数;m 表示文本库中包含的不同的特征项(词)的个数。也就是说，文本中的不同的特征项（词）对应于矩阵 A 的一行，而每一个文本则对应于矩阵 A 的一列。W（｜W｜＝M）是表示文档 D 中的关键词集合，S（｜S｜＝N）表示文档 D 中的句子集合，列向量 a_{ij} 表示在文本中句子 j 的词频向量，那么项/文本矩阵 A 表示为：$[a_{ij}]$。

$$A=\begin{array}{c} \\ W_1 \\ W_2 \\ \vdots \\ W_m \end{array}\begin{array}{c} \begin{array}{cccc} S_1 & S_2 & \cdots & S_n \end{array} \\ \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \end{array}$$

其中，a_{ij}表示第 i 个词在第 j 个句子中出现的频度。由于词和句子的数量很大，而单个文本中出现的特征项的个数是有限的，故项/文本矩阵 A 为高阶稀疏矩阵。

(1).局部权值 L(i，j)，表示第 i 个词在第 j 句子中的权重。

(2).全局权值 G(i)，表示第 i 个词在整个文本中的权重。

这样 a_{ij}可以由下式求得：

$$a_{ij}=L(i,j)\cdot G(i) \qquad (3-1)$$

局部权值 L(i，j)和全局权值 G(i)有不同的取值方法。表 3.1 列出了局部权值常用取值方法:

表 3.1 局部权值 L(i，j)的计算方法

方法名	词频法	0/1 二值法	对数词频法
公　式	Tf_{ij}	0/1	$Log(Tf_{ij}+1)$
备　注	项 I 在文本 j 中出现的频度	项在文本中出现为 1,否则为 0	

全局权值的常用取值方法如表 3.2 所示:

表 3.2 全局权值 G(i)计算方法

方 法 名	Normal	Gfidf	Idf	Entropy
公　式	$\sqrt{1/\sum_j TF_{ij}^2}$	Gf_i/df_i	$Log2(\frac{ndocs}{df_i})+1$	$1-\sum_j \frac{p_{ij}\log(p_{ij})}{\log(ndocs)}$

其中 Tf_{ij}表示特征项 i 在文本 j 中出现的频度；Gf_i表特征项 i 在整个文本库中出现的频度；df_i为文本库中包含特征项 i 的文本数；ndocs 为文本库中包含的文本总数；$p_{ij}= Tf_{ij}/Gf_i$。

实验表明，对数词频法取局部权植，Entropy 法取全局权植效果比较好。

下面用一个简单的示例来说明 LSA 在中文信息处理中应用的效果。由于该

例采用的样本数较少，无法体现出词汇权重的统计意义，因此在此不考虑词汇在文档中的权重问题，且在比较文档之间、词汇之间相似度时，采用它们在高维下表示的相关系数来度量。现有个文档，其中 4 个是关于物理学的，另 5 个是关于主成分分析的，文档内容如表 3.3。

表 3.3 原始文本

编号	文本
Phy1	力是物质间的一种相互作用，运动状态的变化是由这种相互作用引起的
Phy2	死力可用物体的质量和该物体由静止状态转入运动状态时所获得的速度的乘积来量度
Phy3	牛顿首先引入了质量的概念，而把它和物质的重力区分开来，物质的重力只是地球对它的引力作
Phy4	任何两个物体都是相互吸引的，引力的大小跟两个物体质量的乘积成正比
PCA1	主成分分析（PCA）法寻找较少的综合指标来代表原来较多的指标
PCA2	主元分析法是一种数据压缩并从中提取有用的信息的方法
PCA3	主成分分析法是解决数据相关并降低数据维数的一种非常有效的统计方法
PCA4	是一种将分散在一组变量上的信息集中到某几个综合指标（主成分）上的探索性统计分析方法
PCA5	当变量间的相关关系量度不明显时，做主成分分析意义不大

提取 9 个文档中的 14 个关键词，建立词——文本矩阵 A，如表 3.4 所示

表 3.4 词汇－文本原始矩阵

词　汇	Phy1	Phy2	Phy3	Phy4	PCA1	PCA2	PCA3	PCA4	PCA5
物质	1	0	2	0	0	0	0	0	0
运动	1	1	0	0	0	0	0	0	0
作用	2	0	1	0	0	0	0	0	0
物体	**0**	**1**	**0**	**2**	**0**	**0**	**0**	**0**	**0**
质量	0	1	1	1	0	0	0	0	0
量度	0	1	0	0	0	0	0	0	0
引力	0	0	1	1	0	0	0	0	0
主成分分析	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**1**
指标	0	0	0	0	2	0	0	1	0
主元分析	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**0**
数据	0	0	0	0	0	1	2	0	0
信息	0	0	0	0	0	1	0	1	0
统计	0	0	0	0	0	0	1	1	0
相关	0	0	0	0	0	0	1	0	1

从这个例子中可以看到，词——文本矩阵往往是一个稀疏的矩阵，包含的信息过于分散。我们可以通过奇异值分解进行降维，使之不再在稀疏。

二. 奇异值分解 SVD

潜在语义分析重点应用了矩阵的奇异值分解(Singular Value

Decomposition，SVD)。SVD 是数理统计中常用的方法之一，大量应用在不受限的最小立方问题、矩阵阶次估计和规范相关分析等问题的解决方案中；在这里我们用来对文本进行近似计算。

比如，对于生成的m×n的矩阵A，我们认为m≥n（词的数目总是大于等于句子的数目），那么对矩阵A的奇异值分解可以定义为：

$$A = U \sum V^T \qquad (3-2)$$

在这里，U=[u_{ij}]是一个 m×n 的列正交矩阵。它的列向量称为左奇异向量；Σ＝diag(σ_1，σ_2，$\sigma_3\cdots\sigma_n$)是一个 n×n 的对角阵，它的对角线的元素都是按照降序排列的非负奇异值；V=[v_{ij}] 也是一个 n×n 的正交矩阵,它的列向量称为右奇异值向量，如果矩阵 A 的秩 rank（A）= r，那么Σ满足：

$$\sigma_1 \geqslant \sigma_2 \geqslant \sigma_3 \cdots \geqslant \sigma_r \geqslant \sigma_{r+1} = \cdots = \sigma_n = 0 \qquad (3-3)$$

对矩阵 A 进行奇异值降维分解 (设 m > n ，rank(A) = r，存在 K，K< r 且 K< < min(m ， n))，则在 2 范数意义下，A 的秩（K）的近似矩阵 A_k 为：A ≈A_k = $U_k\Sigma_k V_k^T$，U_k 和 V_k 均为正交矩阵，它们的列分别被称为矩阵 A_k 的左右奇异向量，Σ_k 是对角矩阵，对角元素被称为矩阵 A_k 的奇异值。矩阵 A 的 SVD 降维分解如图 3.1 所示。

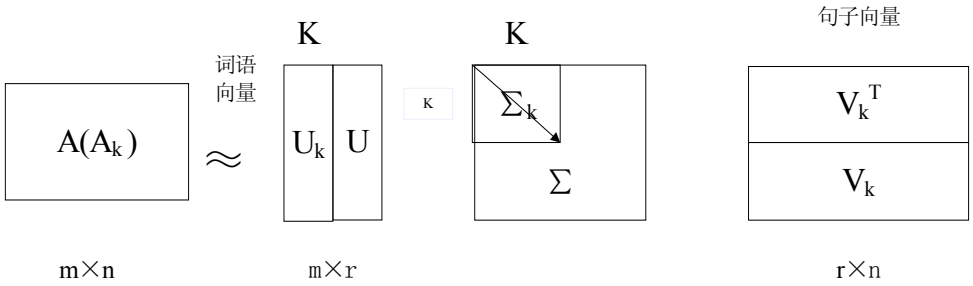

图 3.1 奇异值分解图示

从某种意义上来说,SVD 是一种用于发掘一组相互无关联的索引变量(因素) 的技术，从而使每个词/文本都可以利用左右奇异值向量，表现为单个 k 维空间向量，并可以削弱噪音。

从转换的观点来看，奇异值分解在 m 维由词频向量权重构成的有限空间和 r 维的奇异值向量空间之间建立了一种映射关系。也就是说矩阵 A 中的每一个列向量，它描述了句子 j 的词频向量权重，对应于矩阵 V^T的列向量 Ψi＝[v_{i1}，$v_{i2}\cdots v_{ir}$]^T；矩阵 A 中的每一个行向量，它表明了词 i 在每个文本中出现的次数，对应于矩阵 U

的行向量 $\Phi_j=[u_{j1}, u_{j2}\cdots u_{jr}]$。这里 Ψi 中每一个向量 v_{ix} 和 Φ_j 中每一个向量 u_{iy} 分别被称为第 x 个和第 y 个奇异值向量的索引。

从语义的角度来看，经过奇异值分解后的矩阵保留了相同的语义结构描述，又对原文本向量进行了降维，表示为 K 个线性独立的基本向量或概念，文本中每个词和句子都可以被这些向量或概念索引来连接，奇异值分解的一个不同于传统方法的特点是能够精确描述词与词之间的相互关系[24]，在同一个空间中表示词语和文本，词—词、词—句子、文本—文本的相似度，可以通过它们在语义空间的位置向量距离来衡量：同义词或包含不同词语但主题语义相近的文本的空间位置相近，非相似的词语或文本空间位置较远，并且原来使用模式中有关词语具体使用的非主要的、无关紧要的信息都被忽略掉，只关注该文本集的主要语义信息。

例如，在类似文本中出现的词，在 K 维词语中也会比较接近。将此 K 维空间理解成概念空间，则表示了这些词在概念上是相似的或同义的。考虑“电脑”、“计算机”、“软件”和“动物”四个词，这里“电脑”和“计算机”是同义的，“程序”是和“电脑”“计算机”相关的概念，而“动物”则是完全不相关的概念。在传统的检索系统中，若文本中没有直接出现项“电脑”，则对关于“计算机”的文本和对关于“动物”的文本查“电脑”所得的查询结果是一样的，都不会被命中。但用户更希望在查询“电脑”时能把关于“计算机”的文本找出来，或把关于“软件”的文本也找出来，只是相关度相对于关于“计算机”的文本要低一些，但绝不希望把关于“动物”的文本找出来。而潜在语义分析技术的引入很好地解决了此类问题，通过奇异值分解形成的潜在语义空间可以表示出这些词之间的内在联系。具体地说，和“电脑”在文本中同时出现的项中有许多也会出现在项“计算机”出现的文本中，如“硬件”、“程序”、“网络”、“操作系统”、“USB”，“CPU ”，“显示器”、“程序员”等等，因而它们在 K 维空间中会有类似的表示，而“软件”的上下文语境会和“计算机”、“电脑”在某种程度上一致，而“动物”则会与其完全不同。因而在 K 维空间表示中，“软件”和“电脑”、“计算机”距离更接近，而和“动物”距离较远，从而，更加凸现出项和文本之间的语义关系。因此，当一个词在文本中很显著或者连续出现，那么这个词将被一个奇异向量捕获并描述[25]。任何含有此项的句子都会被此向量所映射，并有最大的索引值与之对应。那么我们可以这样认为，矩阵中每一个奇异值向量描述了一个突出的主题和内容，大量符合要求的奇异值表明这些主题或内容在文本中的重要程度。这种方法有效地解决了词语使用中多义性和同义性的问题。

奇异值分解 SVD 的几何解释如下图所示：

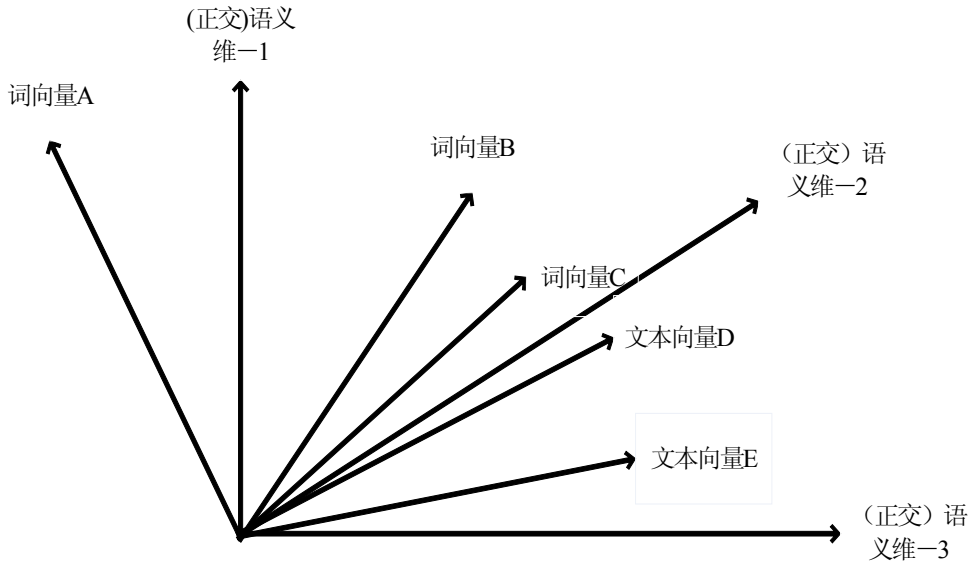

图 3.2 项和文本在语义空间上的表示

LSA 利用词语之间的关联性，通过对文本库中词语的上下文使用模式进行统计转换，获得一个新的低维空间，这个空间就是潜在的语义结构空间。把项和文本在同一个语义空间中表示出来。由上节可知，任意一个项/文本矩阵，经过 SVD 分解有：$A=U\Sigma V^T$，选择适当的 K 值，得到降维的分解式：$A\approx A_k=U_k\Sigma_k V_k^T$。其中，奇异值向量矩阵的行可以看作是项和文本在 K 维空间的表示，即项和文本的空间坐标。

项—项、项—文本、文本—文本之间的相似度，可以通过它们在语义空间的位置向量距离来衡量。同义词(如图 3.2 中的 B、C)或包含不同词语但主题语义相近的文本(如图 3.2 中的文本 D、E)距离较近，非相似的词语或文本空间位置较远(如图 3.2 中的词 A、B)。

下面我们以表 3.4 所示的词汇－文本原始矩阵为例。原始矩阵中，“主成分分析”和“主元分析”两个词汇的相关系数为－0.250。“物体”和“主元分析”的相关系数为－0.177。计算奇异值分解，并保留 2 个奇异值，再进行奇异值分解反运算得在进行奇异值分解反运算得 A 的近似矩阵，如表 3.5 所示：

表 3.5 原始矩阵在二维空间中的重构矩阵

词　汇	Phy1	Phy2	Phy3	Phy4	PCA1	PCA2	PCA3	PCA4	PCA5
物质	0.998	0.558	1.284	0.753	-0.033	-0.030	-0.067	-0.030	0.036
运动	0.444	0.266	0.571	0.337	0.010	0.0049	0.02.	0.005	0.036
作用	0.918	0.541	1.181	0.692	-0.033	-0.030	-0.067	-0.030	0.031
物体	**0.585**	**0.356**	**0.754**	**0.448**	**0.037**	**0.026**	**0.077**	**0.025**	**0.068**
质量	0.734	0.441	0.946	0.559	0.018	0.010	0.038	0.010	0.061
量度	0.173	0.129	0.226	0.147	0.136	0.103	0.281	0.102	0.124
引力	0.570	0.339	0.734	0.432	-0.002	-0.005	-0.005	-0.005	0.034
主成分分析	**-0.025**	**0.09**	**-0.019**	**0.046**	**0.546**	**0.417**	**1.125**	**0.412**	**0.451**
指标	-0.032	0.056	-0.032	0.022	0.387	0.296	0.798	0.293	0.139
主元分析	**-0.010**	**0.015**	**-0.010**	**0.005**	**0.107**	**0.082**	**0.221**	**0.081**	**0.088**
数据	-0.055	0.099	-0.054	0.040	0.687	0.524	1.415	0.518	0.565
信息	-0.020	0.029	-0.020	0.011	0.214	0.163	0.440	0.161	0.175
统计	-0.032	0.057	-0.032	0.023	0.396	0.302	0.816	0.299	0.326
相关	-0.014	0.070	-0.008	0.038	0.046	0.310	0.835	0.306	0.355

在初始矩阵中，向量“主成分分析”和“主元分析”的相关系数只有－0.250，表现不出两者的相似性。在降秩矩阵中，向量“主成分分析”和“主元分析”的相关系数已非常接近 0.997，可见降秩后含义相近的词相关性得到加强。而另一方面，在初始矩阵中向量“物体”和“主元分析”相关系数为－0.353，降秩后，相关系数降到－0.765，可见含义不相关的词汇之间的相关性被减弱。

3.1.3 潜在语义分析的特点

潜在语义分析技术 LSA 是一种通过分析大量的文本集，自动生成特征项一语义之间映射规则的方法，LSA 认为词语在文本中的使用模式中存在着潜在的语义结构，同义词之间应该具有相同的语义结构，多义词的使用必定具有多种不同的语义结构。LSA 就是通过统计方法，提取并量化这些潜在的语义结构，进而消除同义词、多义词的影响，提高文本表示的准确性。

在 LSA 空间结构中，文本和词语依据语义上的相关程度被组织存放：分散在不同文本中的同义词空间位置相邻。LSA 方法对语义空间的维度进行约简，消除语义表达中的“噪音”，且词语含义是词语多种含义的带权平均。因此，相对于传统的统计模型和向量空间模型，它具有以下特点：

（1）.LSA 假设隐藏在词语中的隐含意思(也就是潜在语义空间的这些语义维)可以更好地刻画文本真实含义。由于表示部分语义维的含义时，在不同的文本集中使用了不同的词语组，使得对这种语义空间表示法的认识受到了阻碍。LSA 力图通过适当的数据处理，达到恢复原始的正交语义结构空间，以及其中的原始的

语义维的目的。

（2）.LSA 利用潜在的语义结构表示词条和文本，将词条和文本映射到同一个 k 维的语义空间内，均表示为 k 个因子的形式，向量的含义发生了很大的变化。它反映的不再是简单的词条出现频率和分布关系，而是强化的语义关系。在保持了原始的大部分信息的同时，克服了传统向量空间表示方法时产生的多义词、同义词和单词依赖的现象。同时，在新的语义空间中进行相似度分析，比使用原始的特征向量具有更好的效果，因为它是基于语义层而不仅是词汇层[26]。

（3）.由于词条和文本被映射到同一 k 维的语义空间，所以在 LSA 模型中不仅能够进行传统的词条与词条、文本与文本之间的相似关系分析，而且能够分析词条与文本之间的相似关系，与传统的向量空间模型相比，具有更好的灵活性。

（4）.对于原始的词条—文本矩阵，通过 LSA 分析提取出 k 维语义空间。在保留大部分信息的同时使得 K<＜ min(m ， n)，这样用低维词条、文本向量代替原始的空间向量，可以有效地处理大规模的文本库。

（5）.LSA 不同于传统的自然语言处理过程和人工智能程序，它是完全自动的。所谓自动，就是 LSA 不需要人工干预，不需要预先具有语言学或者知觉相似性知识(不使用人为构造的字典、知识基础、语义网络、文法、词法、句法剖析器等，它的输入只是原始的未经处理的文本序列)。它完全是根据普通数学学习方法，提取合适的维度语义空间，结合其他理论方法，达到有效展示对象和文本内容的目的。通过对大量的文本分析，LSA 可以自动地模拟人类的知识获取能力，甚至分类、预测的能力。

潜在语义分析虽然能够很好解决文本同义词和多义词的问题，但是它仍然存在以下不足：

（1）.LSA 在进行信息提取时，忽略词语的语法信息(甚至是忽略词语在句子中出现顺序)，认为语法结构在文本的语义表达中处于次要的地位。它仍然是一种简单地通过所有词语向量的线性总和来产生文本向量，表示文本的含义的方法。然而，句子的语法结构包含了词语之间的更深层次的语义关联信息，忽略这种语义关联关系影响了 LSA 对文本内容的把握能力。因此，现在的一些研究开始将 LSA 模型和语法信息相结合，以提高 LSA 的性能。

（2）.LSA 处理的对象是可见变量(文本集中出现的词语、文本)，它不能通过计算得到词语的暗喻含义，以及类比推论含义。为解决这些因素对 LSA 获取知识能力影响，需要在 LSA 的基础上结合其他的理解模型。

3.2 篇章多级依存结构

篇章多级依存结构 TMDS(text multilevel dependency structure)分析是一种基于结构的分析方法，篇章是一个有机的整体，段落和段落之间、句子和句子之间存在着一种依存关系，通过对这种关系的分析，找到整篇文章中最核心的内容，去除冗余的部分，克服其他文摘方法的不足，提高文摘的质量。

3.2.1 篇章多级依存结构产生的背景

新兴的话语语言学对篇章的结构十分重视，有的文献将文章的表层结构分为并列结构和链式结构两种，将文章的深层结构分为并列式、推衍式和混合式三种。但语言学关于篇章结构的观点是站在人的立场上阐述的，能够用来自动识别篇章结构的可形式化的规则很少。

在计算语言学界，1991 年加拿大 York 大学的 J.Morri 和多伦多大学的 G.Hirst 提出了词汇集聚理论。词汇集聚就是一种把文本中相关的词构成一个链(简称词链)的过程，词链是文本中有关同一事件的单元结构的链接，通过查找词链可以确定文本结构。

1996 年前后，美国著名的情报科学专家 Salton 将自动分类的思想引入文本结构分析，他为每个段落赋予一个特征向量，用向量的内积作为两个段落的相关性的度量，于是一篇文本通过段落与段落的聚合构成了一个篇章结构网络。

我国山西大学关于文本结构自动分析的研究进一步发展了 Salton 的方法。在《哈尔滨工业大学工学博士学位论文》第五章《篇章宏观结构分析》一文中，提出了篇章宏观结构分析算法，作者将章节内部各个段落之间依存关系分为四种：

(1)并列关系(BL)：两个段落的重要性相同，依存弧从前一段落指向后一段落，弧上权值为 0。

(2)总分关系(ZF)：一个段落起总说作用，称为总说段；另一个段落起分说作用，称为分说段；依存弧从总说段指向分说段，弧上的权值大于 0。

(3)困果关系(YG)：前一句是结果，后一句是前一句的原因。

(4)转折关系(ZZ)：因果、转折两种关系的含义与复句间的因果、转折关系相同，依靠关联词来识别。

将全文内部各个章节之间依存关系分为两种：

(1)并列关系(BL)：两个章节的重要性相同，依存弧从前一段落指向后一段落，弧上权值为 0。

(2)总分关系(ZF)：一个章节起总说作用，称为总说章；另一个章节起分说作用，称为分说章；依存弧从总说章指向分说章，弧上的权值大于 0。

哈尔滨工业大学计算机科学与工程系李挺、王开铸在此基础上提出了一种基于篇章多级依存结构的自动文摘方法[27]，这种方法既克服了机械文摘的表层性，又克服了理解文摘的领域局限性。文中给出了篇章多级依存结构的形式化描述，证明了篇章多级依存结构具有非常适合于自动文摘的优点，给出了如何识别、化简篇章结构，如何从压缩了的篇章结构中生成摘要的方法。

3.2.2 篇章多级依存结构的基本原理

我们把文本看成一个图，图的根节点就是全文节点，它的下级节点为章节节点，依此类推有段落节点、复句节点、单句节点，图中至少有一个全文节点（即全文至少有一个章节组成），单句节点使叶节点，并且只有单句节点是叶节点，因为单句节点是构成篇章的物理实体，而其他节点都是逻辑单位，如果某个逻辑结节点下没有了物理节点，那么它自身也就不存在，每个非叶节点的下级节点按照它们在文章中出现的先后顺序从左到右排列；图中上层与下层之间由部整关系弧连接，层与层之间由依存关系弧连接；不同语言层次中存在不同类型的依存关系，高层语言单元之间依存关系的类型较少，低层语言单元之间依存关系的类型较多。

章节层的依存关系包括：并列、承接、总分等；段落层的依存关系包括：并列、承接、总分、转折、因果等；复句层的依存关系包括：并列、承接、递进、总分、选择、转折、因果等；单句层的依存关系包括：并列、承接、递进、总分、选择、转折、因果、充分条件、必要条件、无条件、让步等。

由此可见，篇章多级依存结构 TMDS 是一个有向图，它由代表语言单元的节点和代表语言单元之间相互关系的弧按照特定的方式结合而成，图中根据上层和下层之间和层与层之间的关系，每条弧的权值的大小不同，参见图 3.3。

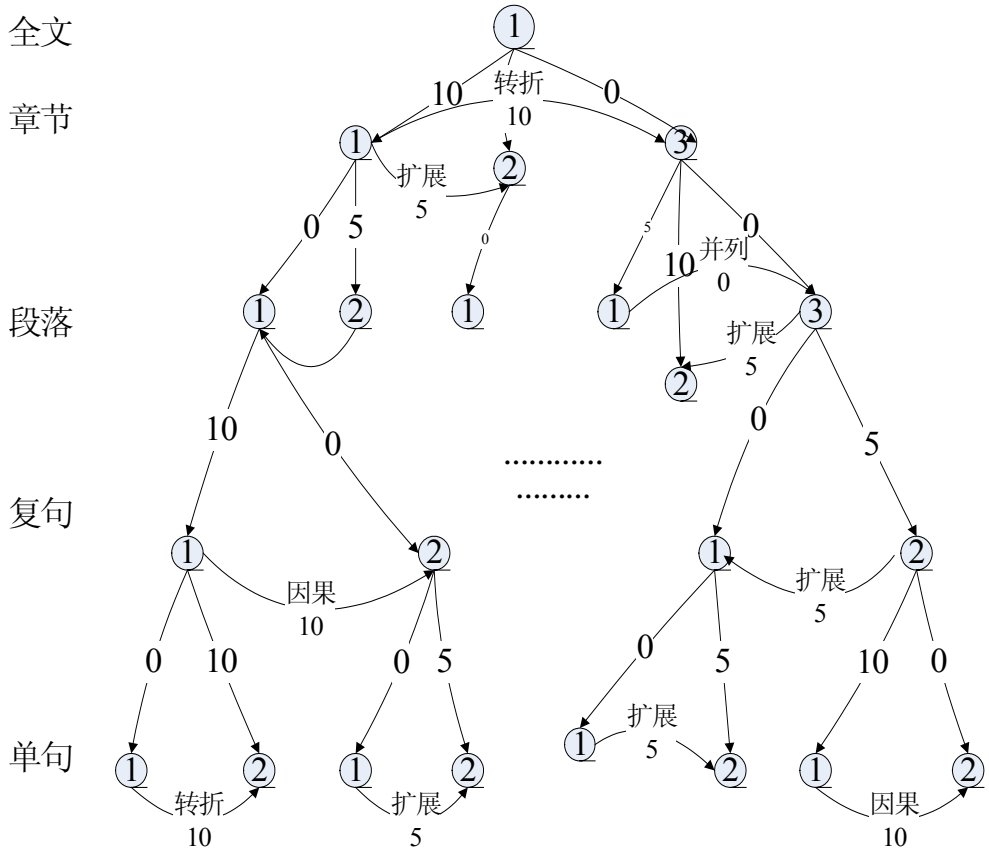

图 3.3 篇章多级依存结构示意图

一.TMDS 的基本概念：

TMDS 是一个可以用一个三元组表示的有向图，TMDS=(V ， E_P， E_D)，V 是分层节点集，E_P 是部整关系弧集，E_D 是依存关系弧集。

(1).分层节点集 V

TMDS 中的节点代表语言单元，节点的全集为 V。篇章中的语言单元可以分为全文、章节、段落、复句和单句 5 个层次，相应地，节点全集 V 也可以划分为 5 个节点子集。

(2).部整关系弧集 E_P

如果全文由 m 个章节组成，则从全文结点出发引出 m 条有向弧分别指向各个章节结点，如果某个章节由 n 个段落组成，则从该章节结点出发引出 n 条有向弧分别指向各个段落结点，如此类推。所有这些有向弧均代表了篇章中相邻上下层语言单元之间的部分-整体关系，因此称为部整关系弧(简称 P 弧)。

(3).依存关系弧集 E_D

隶属于同一上级节点的各个下级节点之间存在依存关系(dependency relation)，例如章节内部段落之间的总分关系，复句内部单句之间的因果关系等。在 TMDS 中采用依存关系弧(简称 D 弧)来表示依存关系，D 弧从依存关系中起支配作用的语言单元(如总说段、果句等)出发，指向从属语言单元(如分述段、因句等)。

(4).多级树 MT

节点全集 V 中的所有节点通过 P 弧相互连接构成一棵有向树，称为多级树 MT(multilevel tree)，$MT=(V, E_P)$。在 MT 中，单句节点都是叶节点，并且只有单句节点才能是叶结点，这是因为单句节点是构成篇章的物理实体，而其它各级节点都是逻辑单位，如果某个逻辑节点之下已没有物理节点，则该逻辑节点也不应存在。每个非叶节点的下级节点按照它们在文章中出现的先后顺序从左到右排列。

二.TMDS 权值的计算：

TMDS 的提出是为自动文摘服务的，自动文摘需要计算出各语言单元在全文中的重要程度。一个语言单元在全文中的重要程度决定于两个方面：一是它所在的上级语言单元在全文中的地位，比如同一个复句如果处在重点段落中就比处在次要段落中更重要一些；二是它在上级语言单元中的地位，比如一个段落中有两个复句，前者是后者的原因，后者是前者的结果，那么后者显然比前者更重要。而它在上级语言单元中的地位，又是由它与周围同级语言单元之间的关系所决定的。

在 TMDS 中，节点代表了语言单元，节点的权值就是语言单元的重要程度的量化体现。一个节点同时处于多级树和某一棵依存树中，它在多级树中的上级节点的权值代表上级语言单元的重要性，它与其上级节点之间 P 弧的权值代表它在上级语言单元中的地位，P 弧的权值由该节点的支配节点所处的地位以及该节点与其支配节点之间地位上的差距(D 弧的权值)所决定。

由上面的分析可知:节点的权值取决于 P 弧的权值，P 弧的权值取决于 D 弧的权值，所以应首先进行 D 弧的加权，然后进行 P 弧的加权，最后给节点加权。加权的具体方法如下:

(1).D 弧的加权

在依存关系中，支配成分的重要性显然高于从属成分的重要性。在 TMDS 中，用 D 弧的权值作为支配节点与从属节点之间重要性差别的度量，D 弧(u，v)的权值记为 $f_D(u, v)$，f_D是从 E_D到自然数集 N 的映射，$f_D(u, v)$越大，u 与 v 的重要性差

别越大。$f_D(u, v)$取决于 R(u，v)，比如因果关系中，果句比因句重要得多，而在递进关系中，前后两句重要性的差别不大。$f_D(u, v)$还与(u，v)所在的语言层次有关，高层语言单元的依存弧权值相对高一些。

(2).P 弧的加权

如果我们能够建立起 TMDS，并依据依存关系的类型等得出每一条 D 弧的权值，那么我们就能够计算出 P 弧的权值。P 弧联系着上下两级语言单元，P 弧的权值反映了下级语言单元在上级语言单元中的地位，权值越小，地位越高。设(w，v)∈EP，v 所在的的依存树为 DT，则 P 弧(w，v)的权值 $f_P(w, v)$的计算方法如下:

① 若 v 是 DT 的根节点，或者说 v 是 w 的下级根节点，则 $f_P(w, v)= 0$。

② 否则，如果 v 的支配节点为 u，则 $f_P(w, v)= f_P(w, u) + f_D(u, v)$。

(3).节点的加权

节点 v 的权值 q(v)的计算方法如下:

① 若 v 是全文结点，则 $q(v)= 0$。

② 否则，如果 v 的上级节点为 u，则 $q(v)=q(u) + f_P(u, v)$。

节点的权值反映了该节点在全文中的地位，权值越小，地位越高。弧和节点的权值计算方法参见图 3.3(图中带箭头的直线代表 P 弧,带箭头的曲线代表 D 弧)。

最后,构建 TMDS 的目的在于通过对 TMDS 的化简获取文摘,我们称原文的 TMDS 为 DTMDS，化简得到的文摘的 TMDS 为 ATMDS。ATMDS 是文摘的 TMDS，这就要求它一方面包含原文中最重要的内容，另一方面仍具备 TMDS 的特征。TMDS 中的物理节点是单句节点,ATMDS 中的单句节点集是由 DTMDS 中权值最小的若干单句节点组成的，这些节点所代表的单句就是全文中最重要的单句，用这些单句来组织文摘就保证了文摘将包含原文中最重要的内容。TMDS 结构有两个特征，一是各层节点纵向地构成多级树，二是每个节点的各子节点横向地构成依存树。同时，化简后得到的 ATMDS 也满足这两个特征。因此，ATMDS 仍然是篇章多级依存结构。

3.2.3 篇章多级依存结构的特点

基于篇章多级依存结构的自动文摘方法的优点是既能避免自动摘录的不连贯性，又能避免基于理解的自动文摘和基于信息抽取的自动文摘受专业领域知识限制的缺陷。特别是当遇到多主题或篇幅很长的文献时，并将文章视为段落的关联网络的方法能很好地进行摘录，再配合以仿人算法，所得的自动文摘的相关性和连贯性都是其他方法无法比拟的。然而，这种方法也有自身的不足之处，最重要

的缺陷就是不能做到让计算机真正理解文献的主题内容。这种方法只是在人工智能和 NLP 领域内无法取得突破性进展时而产生的一种替代方法。同时这种方法比较适用于写作思路比较清晰、开头结尾概括了文档的主要内容的科技性文献和新闻，对于有隐含意义题材的文章如散文、诗歌和小说并不适用。

3.3 融合两种方法生成文摘

现在研究如何简单而无冲突地融合基于 LSA 和 TMDS 地分析结果，以提高系统生成文摘地质量。

首先，利用 LSA 从原始的文本中抽取包含主要信息的句子，对这些从文本中抽取的句子重新组织，按其在原文中的顺序排列，并用符号表示。由于 ATMDS（文摘篇章多级依存结构）具备 DTMDS 的全部特征，它记录着文摘中各级语言单元之间的依存关系，所以采用 ATMDS 来表示降维后的文本结构，将 ATMDS 中权值最小的文摘基本单元（这里为单句）加入摘要，直到达到特定长度。摘要的长度（长度是以句数而非字数来计算）由用户确定，通常为原文长度的 10％—30％。

最后，将 ATMDS 中的单句按照它们在原文中的顺序依次输出，并根据 ATMDS 中记录的各种依存关系将相应的关联词语嵌入单句之间，同时利用省略和指代消除首语重复，获得具有一定逻辑性的较为连贯的文摘。

3.4 本章小结

本章对潜在语义分析和篇章多级依存结构的基本概念、原理、特点给予了详细的介绍，并在此基础上，提出了一种基于潜在语义分析和篇章多级依存结构相结合的自动文摘方法，其特点是利用潜在语义空间的项、段落、文本的相似度运算，而不是单纯的词频。方法可操作性强，不依赖于领域，完全建立在文本集上，除分词字典外，无需任何其他知识库的支持。

篇章多级依存结构将文章视为句子的关联网络，与很多句子都有联系的中心句被确认为文摘句。它弥补了奇异值分解在语法和句法上的不足，使生成的文章更加连贯，不受专业知识领域的限制，但是对于篇幅较长的文章，句子之间的关联网络将十分庞大，其时空开销都将是难以承受的。因此，在未来的工作中我们要想法提高算法的效率，使之能更有效的处理大规模动态变化的文本集。

第四章 系统的设计实现与试验分析

本章主要介绍基于 LSA 和 TMDS 的自动文摘系统的实现的过程，对系统的整个模块进行全面的介绍，特别是潜在语义分析中奇异值的分解的实现，篇章多级依存结构中文本结构分析的实现；同时，也对文章摘要的评估方法进行介绍，并进行了试验和数据分析。

4.1 系统的设计实现

基于 LSA 和 TMDS 的自动文摘系统是一个较为复杂的文摘系统，它的基本思路是：利用潜在语义分析对词—文本矩阵降维，去掉语义噪音；采用篇章多级依存结构分析对经过降维处理后的文本先后进行文本结构分析，提取特征词，确定单句、复句、段落等之间的关系；利用结构过滤掉不重要的句子和段落形成文摘篇章多级依存结构树，将树中的单句按照原文的顺序输出就得到一篇文摘。

4.1.1 系统的主要功能

本系统的主要功能是，首先对整篇文本进行分句、分词的预处理，应用常用词表进行常用词的过虑；在此基础上进行潜在语义分析，保留相同的重要的语义结构，过滤掉噪音；然后采用篇章多级依存结构分析来确定各个句子间的深层语义关系，通过计算句子的权值来确定其在文章重的重要程度，按比例抽取权值较小的句子生成摘要；并根据 ATMDS 中记录的各种依存关系将相应的关联词语嵌入单句之间，同时利用省略和指代消除首语重复，这样即可获得具有一定逻辑性的较为连贯的文摘使生成的文摘更加流畅，减少冗余。

4.1.2 系统主要模块的设计

系统由四个主要模块组成，各个模块的总体结构如图 4.1 所示。

1．文档的预处理

进行文档分析之前，要对为本进行预处理，对文本进行词的切分，从文本中抽取词汇和短语。系统采用的是本试验室开发的一种自动分词程序，它采用 11 段

的思想[1]对文本中的句子进行分词和词性的标注，抽取文本的原始词频向量，生成文本的词汇表，而将虚词、介词、连词、特高频词、特低频词等组成一个常用词表。其次，保留词汇表中那些对表示文本内容作用大的词汇（如名词、动词），把常用词表中的词汇从文本的词汇表中过滤掉。由于词处理是中文自然语言处理的基础，词处理结果的好坏直接影响文摘生成的质量，所以完成以上工作后，我们再进行语义分析。

2.语义模型分析

根据预处理后的文本表示，构建一个词－句子矩阵并进行奇异值分解，对分解后的矩阵再进行降维分解和重构。

3.结构分析

对从语义矩阵中抽取的含有语义关系的句子进行结构分析，确立各句子之间和段落中各语言单元之间的关系，用篇章多级依存结构树来表示。

4.摘要生成

根据文摘篇章多级依存结构树表示的句子间的相互关系，选择重要的句子生成摘要；同时利用省略和指代消除技术整合各文摘句，以生成一个较为连贯的文本摘要。

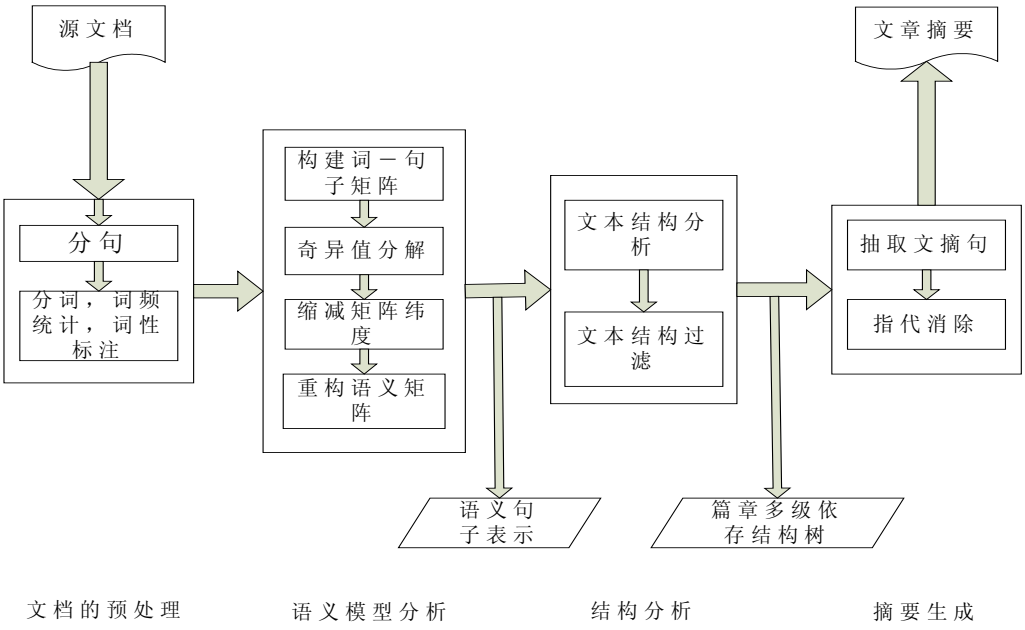

图 4.1 基于 LSA 和 TMDS 的自动文摘系统结构图

[1] 11段的思想是在语法分析的基础上，把句子分成背景、主语部分、谓语部分、宾语部分以及标点符号五个部分。其中，主、谓、宾三个部分分别占用三个段，每部分的前后两段分别是修饰中心词的定语、状语等内容。

4.1.3 词—句子矩阵的实现

词—句子矩阵是潜在语义分析方法的基础，也是建立语义向量空间的基础。这个矩阵是奇异值分解矩阵分解的输入。

形成词—句子矩阵需要进行的步骤有：分词—计算词频—形成矩阵。在本文中采取的分词算法是正向最大匹配。图 4.2 是这一过程的流程图描述：

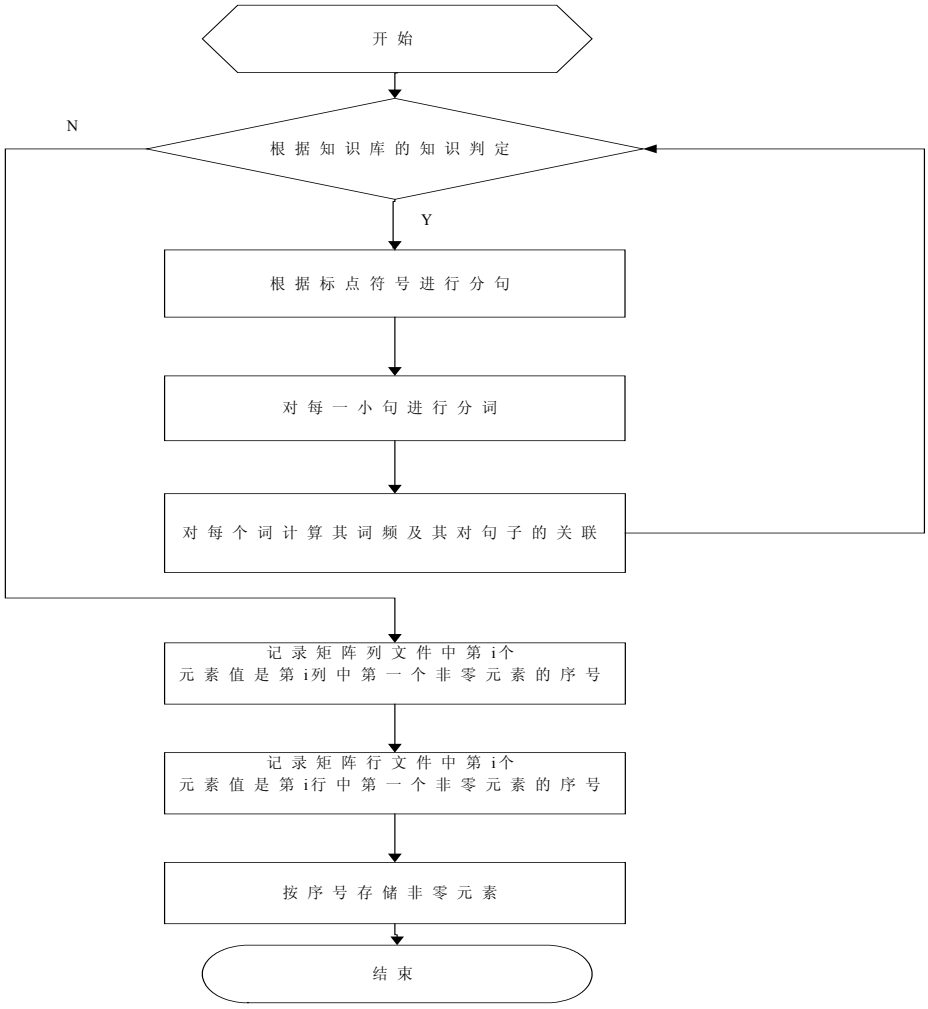

图 4.2 产生词—句子矩阵的流程图

词—句子矩阵是一个大的稀疏矩阵，本文中采取了一种将稀疏矩阵分解为三个文件存储的办法：第一个文件 matrix_value 存储的是矩阵中所有的非零元素，其中非零元素按照矩阵列遍历的顺序排列。第二个文件 matrix_row 存储的是 matrix_value 中的每一个非零元素对应的初始矩阵中的行数，其存储顺序与 matrix_value 中的相同。第三个文件 matrix_col 中，第 i 个数值存储

的是初始矩阵中第 i 个列中第一个非零元素在所有非零元素中的序号，最后一个数值是整个初始矩阵中非零元素的总数。

算法描述如下:

```
for 知识库中的每条知识 do
    根据标点符号进行分解，分解出 sentences
    for 每一个 sentences do
        dividterm (sentences)  //对每一小句做中文分词
                               //获得词 termj，放入 termlist 中
    end for
    for termlist 中的每个词 do
        统计词频
    end For
    relationvectors(RelVector)//记录该句子 ID 号，以及该句子中
                              //包含的词的 ID 号和出现频次
end for
s=1
for 每一个关系向量 RelVector do
    将 s 存入 matrixcol 文件
    for 每一个词 do
        将词频写入 matrix_value 文件
        将词的 ID 号写入 matrix_row 文件
    end for
end for
```

其中：

SentenceString 类型，存储每一小句和在文中的顺序代码

dividterm (String argu)中文分词的实现方法，返回值是 termlist

termlist 存放 term 对象的集合

RelVector，vector 类型存放词和句子的关系

Matrix_col 存放矩阵每一列中第一个非零元素的序号，该序号是该非零元素在所有非零元素中的顺序号

Matrix_row 存放每一个非零元素所在的行，存储顺序与 matrix_value 相同

matrix_value 按非零元素的顺序存放非零元素值

通过对文本的分析，我们把具有文本中最能表达文章主要内容的词，句子从文本中抽取出来，以矩阵的数学形式加以表示，并且用数组把分析的结果保存起来，为下一步实现起奇异值分解作好准备。

4.1.4 奇异值分解的实现

奇异值分解对初始矩阵进行操作：先求出矩阵的秩，再求出矩阵的所有非零奇异值，奇异值的个数小于等于矩阵的秩；根据奇异值构成一个对角线矩阵Σ，Σ的主对角线上的元素是奇异值其它元素为零；根据奇异值计算出奇异值对应的奇异向量构成两个奇异矩阵，分解得到三个矩阵(一个对角线矩阵，两个奇异矩阵)，这三个矩阵相乘就能得到初始矩阵。

上述获得的是三个高维的矩阵，在系统中，可以对其进行简化：获取调整过的 k 值(整数)，根据 k 值来确定奇异值分解得到的三个矩阵的维数，U_k，Σ_k和 V_k^T 中的 k 就是选取的 k 值。通过选取恰当的 k 值可以减少三个矩阵的维数，而且简化后的三个矩阵的乘积与初始矩阵是非常近似的，对于进行下一步计算的误差影响不大，不会影响词与词、句子和句子的相似度的精度。但是 k 值如果选的不好，简化后的矩阵就丢失了原始矩阵所含有的一些关系，因此根据这几个矩阵计算出来的相似度就会有很大的误差会影响到最后答案的选择。

进行奇异值分解时主要是求出矩阵的奇异值然后根据奇异值算出两个奇异矩阵，设 A 的非零奇异值有 r 个，分解生成的三个矩阵是 U_{mr}，Σ_{rr}和 V_{nr}^T。下面介绍计算奇异值和奇异矩阵的方法。

根据矩阵 A 新生成一个矩阵 B,B 是 A 和 A 的转置的乘积即 $B=AA^T$或者 $B=A^TA$，这样 B 的特征值的非负平方根就是 A 的奇异值，而且 AA^T和 A^TA 有相同的特征值，因此通过求 AA^T(或 A^TA)的特征值的平方根就可以得到矩阵 A 的奇异值。矩阵 AA^T 的特征向量称为左奇异向量，m 个左奇异向量构成 U_{mr}；矩阵 A^TA 的特征向量称为右奇异向量，n 个右奇异向量构成 V_{nr}^T。因此根据特征值求出 AA^T和 A^TA 的特征向量，就能得到两个奇异矩阵 U_{mr}和 V_{nr}^T。

这样求矩阵 A 的奇异值就转化为计算矩阵 B(由矩阵 A 构造)的特征值。求 B 的特征值时首先将 B 化成三对角形式，然后计算化简矩阵的特征值，也就是 B 的特征值。根据 B 的构造方法，A 的奇异值通过对 B 的特征值取平方根就可得到，然后根据奇异值进行下一步计算。根据在进行奇异值分解前确定的 k 值，对矩阵进行化简得到:U_k，Σ_k和 V_k^T三个简化的矩阵。由于越大的奇异值上携带越多的信息量，

确定 k 个奇异值的方法是在生成的所有奇异值中按奇异值的大小顺序，选出 k 个并求出对应的两个奇异矩阵。在矩阵 V_k^T 中用列向量表示句子，用行向量表示词在所有句子中的右奇异值向量，第 k 个奇异值表示的主要内容就对应于矩阵 V_k^T 第 k 行中索引值最大的那个向量的列号表示的句子（列号和文本中句子数目和顺序一一对应），即最重要的句子。我们把这些句子和句子的序号保留在 SentenceNum 中，作为候选语义句子集合。

由于初始矩阵通常很大，因此奇异值分解得到的矩阵也会很大，不能直接放到内存中来保存。U_k，Σ_k 和 V_k^T 的存储方法：含有奇异值的矩阵 Σ_k 可以用一个数组来存放，数组的大小是 k，每个元素的值就是对角线上的元素，或者用一个 k×k 的一维数组存放矩阵的所有元素；得到的两个奇异矩阵 U_k 和 V_k^T 中的数据用两个文件来存储，矩阵的行对应文件中的行。

基于以上的讨论，我们给出一种基于奇异值分解的抽取语义句子的方法，并用符号表示生成的句子，其算法描述如下：

1.设 N＝1。

2.对矩阵 A 进行奇异值分解，得到一个新的奇异值矩阵 Σ 和右奇异值向量矩阵 V^T，在奇向量空间中，每个句子 j 用矩阵 V^T 的列向量 $\Psi i=[v_{i1}, v_{i2}\cdots v_{ir}]^T$ 来表示。

3.选择矩阵 V^T 中第 N 个右奇异向量。

4.选择第 N 个右奇异值向量中最大的索引值所对应的句子，把它送入候选文摘集中。

5.如果 N 达到了我们预期的数量，则中止操作，否则 N 的值加“1”，转向步骤 4。

在步骤 5 中，找出第 N 个右奇异值向量中最大索引值，等同于找到列向量 Ψi 中第 N 个基本元素 V_{ik} 是最大的。

图 4.3 是这一过程的流程图描述：

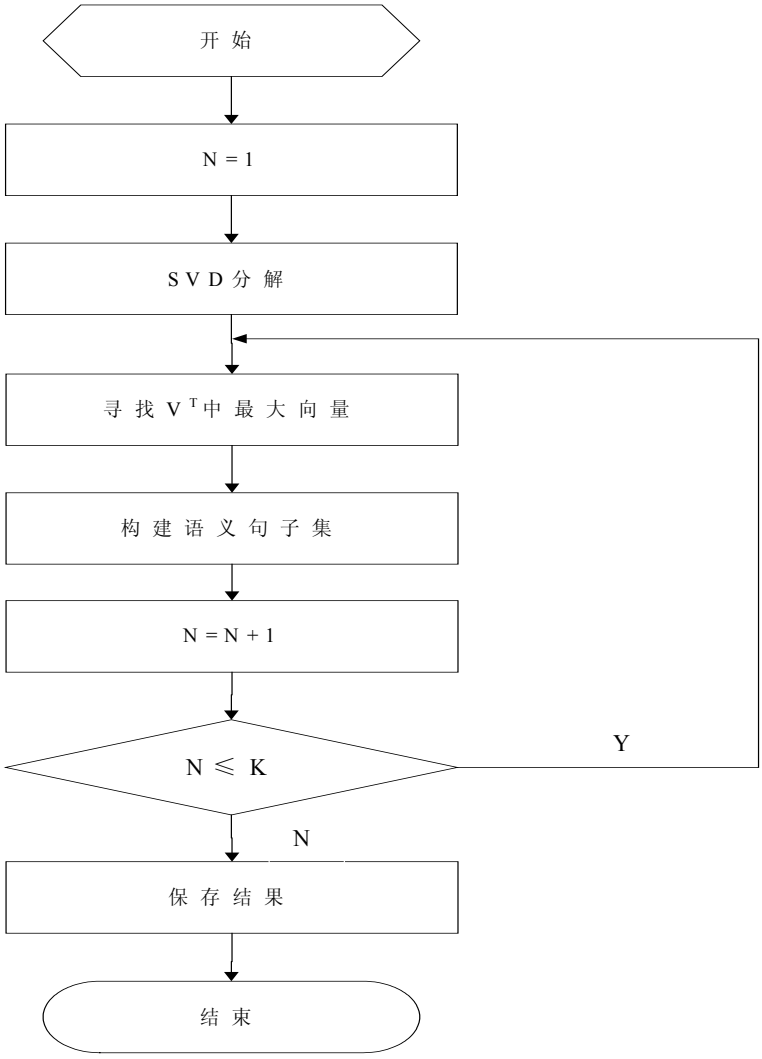

图 4.3 奇异值分解的流程图

从算法和流程图中我们可以看到奇异值分解是同通过对参数 N 的调整实现对矩阵的降维，问题的关键是奇异值的分解计算和右奇异向量的选取，以及 K 值大小的确定，这三部分决定了新的语义矩阵是否能正确表达原矩阵的中心内容。篇章多级依存结构分析就是在此基础上进行的，因此，奇异值分解的好坏直接影响到整个文摘的质量。

4.1.5 文本依存结构分析的实现

在潜在语义分析中，我们对文本进行了分句处理，保存在 SentenceString 中，经过奇异值分解和降维，过滤了大部分的语义噪音，把能够表达文本主要内容的句子在原文中按顺序编号，保存于 SentenceNum 中，这样新的语义文本作为结构

分析的输入，它在原有文本中的位置和顺序并不改变。

结构分析模块分为两个部分，一是原文结构分析，包括预处理、语言单元切分、依存关系分析、建立 DTMDS；二是结构的过滤，包括依存结构的加权、关系弧的加权、节点的加权、节点的过滤。其详细步骤如下：

①预处理：主要是指词的处理，包括词的切分、词频统计、特征词的抽取，词类和义项的标注等。首先我们要对文本进行词的切分，从文本中抽取词汇和短语。系统采用的是本试验室开发的一种自动分词程序，它能够对文本中的句子进行分词和词性的标注，抽取文本的原始词频向量，生成文本的词汇表，而将虚词、介词、连词、特高频词、特低频词等组成一个常用词表。其次，保留词汇表中那些对表示文本内容作用大的词汇（如名词、动词），把常用词表中的词汇从文本的词汇表中过滤掉。

②语言单元切分：切分的目的在于确定各级语言单元(包括单句、复句、段落和章节)的边界，从而建立 TMDS 的主体结构——多级树 M T（Multilevel Tree）。

③依存关系分析：分析原文中每一语言单元内部树形依存结构， 并确定各子单元之间依存关系类型，即获得了依存树 DT(Dependency Tree)。

⑤建立 DTMDS：以多级树 M T 为主体，将各级依存树 DT 组合起来，建立原文的篇章多级依存结构 DTMDS(Document TMDS)。

⑥依存结构的加权：根据依存关系的类型可以为依存树 DT 中的每一条依存关系弧确定权值。

⑦部整关系弧的加权：部整关系弧所指的子节点在依存树中的重要性越高，则部整关系弧的权值越小。

⑧节点的加权：每个节点的权值等于从该节点出发沿部整关系弧返回 M T 根节点的路径上各条部整关系弧的权值总和。

⑨过滤：将所有的叶节点(单句)按权值从低到高排序，根据用户对文摘句数的需要截取相应数量的低权值节点，将 DTMDS 中与这些节点无关的节点和弧全部删去，剩下的节点和弧即为 DTMDS 的子树——文摘篇章多级依存结构 ATMDS (Abstract TMDS)。

⑩选取文摘句：将 ATMDS 的叶节点对应的单句抽取出来，按其在文章中的顺序排列，补充相关的连接词，生成一篇连贯的文章摘要。

基于 TMDS 的自动文摘方法的目标是按任意比例生成报道性文摘，并且文摘力求抓住原文的核心内容，语句连贯，不遗漏主题。

简化得到的每个叶节点代表一个单句，每个单句又代表一个命题，那么一篇

文章就是许多命题按照相互依存的方式逐级构成的系统，每个命题在系统中所处的位置是该命题重要性的标志，选择重要的命题（权值越小重要程度越高），并按照它们之间原有的结合方式组成新的篇幅较短的文章，就是文摘。

下面给出一个简单的文本摘要的例子：

【原文】

研究生全面收费尚未通过国家有关部门审批

记者从有关高校了解到，今年北京大学、清华大学、武汉大学、华中科技大学、复旦大学、上海交通大学、同济大学、西安交通大学和哈尔滨工业大学 9 所院校将参与研究生培养机制改革试点，率先实行研究生全面收费。

据记者了解，目前，试点项目还没有通过国家有关部门的实施审批，试点内容什么时候实施现在还难以确定。

北京大学、清华大学的相关负责人昨天表示，目前学校还没有得到教育部的通知，明年是否全面收费不能确定。即使实行全面收费，学校也将会完善“奖、贷、助、补、减”的资助体系，加大资助力度，家庭贫困的学生不会因此被拒之门外。

目前，一些新设的研究生教育如ＭＢＡ早已实行收费，研究生教育不是义务教育，收费是大势所趋。但某试点高校的有关负责人表示，如果实行“全面收费”，学生交的学费也将大大小于研究生教育成本。“因为将本着‘谁受益，谁交费’的原则。培养研究生学生本人受益，学校和国家也在受益，所以将有多个单位共同承担教育成本。”

【10％文摘】

据记者了解，目前，试点项目还没有通过国家有关部门的实施审批。

【20％文摘】

据记者了解，目前，试点项目还没有通过国家有关部门的实施审批，试点内容什么时候实施现在还难以确定。即使实行全面收费，学校也将会完善“奖、贷、助、补、减”的资助体系。

【30％文摘】

记者从有关高校了解到，今年北京大学、清华大学、武汉大学、华中科技大学、复旦大学、上海交通大学、同济大学、西安交通大学和哈尔滨工业大学 9 所院校将参与研究生培养机制改革试点，率先实行研究生全面收费。据记者了解，目前，试点项目还没有通过国家有关部门的实施审批，试点内容什么时候实施现

在还难以确定。北京大学、清华大学的相关负责人昨天表示，即使实行全面收费，学校也将会完善“奖、贷、助、补、减”的资助体系。

通过这个例子我们可以看到：基于潜在语义分析和篇章多级依存分析的自动文摘方法生成的文摘能较好的表达文章的中心意思，上下文比较连贯，符合实际文摘要求。我们将在下一节详细介绍如何对生成的文摘的质量进行评估。

4.2 实验分析

本节中，主要针对上述的分类过程和算法设置一些系统评估。文本摘要的评估方法大致可以分为两类[28]：内部评测和外部评测。采用的评测语料库是国家语委现代汉语平衡语料库。

4.2.1 评测用语料库

本文采用的评测用语料库是国家语委现代汉语平衡语料库。国家语委现代汉语平衡语料库是一个经过人工分类好的平衡语料库，它包含了人文与社会科学、自然科学及综合三大板块。其中，人文与社会科学包括了政法、历史、社会、经济、艺术、文学、军体、生活 8 个大类；自然科学包括了数理、生化、天文地理、海洋气象、农林、医药卫生 6 个大类；综合板块包括了应用文、其他 2 个大类。每个类别包含的文章数量多少不等，多的有 533 篇，少的有 28 篇；每篇语料大约 2000 字左右，内容由报纸、教材、综合刊物等上面的内容构成。

4.2.2 内部评测

内部评价(Intrinsic)方法，它是通过直接分析评价自动摘要系统生成的摘要，通过比较来判断摘要中包括了多少原文的主题内容及摘要的流畅度等。例如，将自动摘要系统生成的摘要与原文比较、自动摘要与人工生成的“理想”摘要比较、自动摘要与不同自动摘要系统生成的摘要比较等。

本文采用的是一种内部评价方法。我们采用精确率(P)和召回率(R)来评价系统生成摘要的质量。假设 T 表示专家文摘抽取的句子数目，S 表示文摘系统抽取的句子数目，那么精确率和召回率按如下公式计算:

$$P=\frac{|S\cap T|}{|S|},\quad R=\frac{|S\cap T|}{|T|} \tag{4-1}$$

例如：某篇文章在文摘长度占文章比例 10%时，系统抽取出文摘句子数为 8 句，该文章的专家文摘抽取的句子数量为 12 句，同时存在于文摘系统和专家文摘句中的句子数量为 5 句，则系统在该文章的文摘长度为 10%时：

召回率(R)＝5/12=0.417

准确率(P)＝5/8=0.625

如果用 CR 表示摘要比率，DR 表示降维比率（DR=k/r，即降维后矩阵的秩和原始文本矩阵的秩的比值），通过试验发现对于不同 CR 值，它的 DR 值不尽相同，当 CR 值为 10％、20％，30％时，DR 的平均值分别为 0.8、0.7、0.6，这个范围内文摘的质量较好[29]。

测试试验:

采用国家语委语料库在对已经做了专家文摘的 10 篇文本(包含报刊、经济、新闻报道几个方面)进行自动文摘后，每类文本的平均评测参数情况如表 4.1 所示。

表 4.1 LSA+TMDS 文摘方法的性能评价

摘要比率	参　数	系　统	降维比率
10％	精确率 P	0.5138	0.8
	召回率 R	0.1859	0.8
20％	精确率 P	0.4834	0.7
	召回率 R	0.3468	0.7
30％	精确率 P	0.4337	0.6
	召回率 R	0.4337	0.6

可见随着 CR 的增加，P 的值越小，R 的值越大，这说明 CR 越大，生成的文摘越能反映文章的中心思想。另外，选择合适的 DR 值，LSA+TMDS 就会取得良好效果，因此在文章摘要中采用奇异值分解能够把文摘从关键字分析的水平提高到语义分析的水平，从文本中获得更为精确的语义信息。

4.2.3 外部评测

外部评价(Extrinsic)方法，它是一种间接的评价方法，将自动文摘应用于某一个特殊的任务中，根据摘要功能提高这项任务的效果来评价自动文摘系统的性能。例如，用户使用摘要确定原文主题的程度或用户基于摘要能回答的原文有关问题的程度等。由于时间的关系，没有做相关的试验。

4.3 本章小结

本章对基于潜在语义分析和篇章多级依存结构的自动摘要方法系统结构以及各个模块的功能进行详细的介绍，同时也介绍了评价文摘系统的主要方法；对文摘系统的试验结果进行了分析，说明此方法的可行性。在今后的工作中将深入研究语言生成问题，结合未登陆词识别、领域自动判别等技术，进一步改进生成的自动摘要的质量，使其更接近人工摘要的自然性、流畅性。

第五章 全文总结和研究展望

5.1 研究工作总结

本文回顾了自动文本摘要的研究现状，介绍了常用的几种自动文本摘要的方法；从理论上探讨了文本摘要的几种方法：基于统计、基于理解、信息抽取、基于结构的自动文摘模型。论述了获取的自动文摘系统的关键技术；尝试性地利用潜在语义分析以及篇章多级依存结构的分析，来完成文本文摘句的提取，并提出一种综合型的中文自动文摘系统的设想；阐述了系统整体性能以及各模块设计；为今后的研究提供了可能的理论基础。

纵观全文，本文的主要工作包括以下几个方面:

1.介绍了国内外关于自动文摘系统等方面的发展动态，总结了各种不同类型的自动文摘系统的优点和不足，并在此基础上提出了建立一种综合型的自动文摘系统的设想。

2.介绍了文摘和自动文摘系统的基本概念体系，并针对目前几种主要的自动文摘系统形式化模型和方法:基于统计、基于理解、信息抽取和基于结构等模型和方法进行了比较和分析，归纳出各自的特点，进而在总结各种不同类型的自动文摘系统的特点的基础上提出了一种综合型的自动文摘系统。

3.给出了潜在语义分析的基本概念和实现方法，并把它应用于文摘系统中，提取主要的语义信息，对文本中噪音信息进行过滤，在一定程度上提高了文摘的质量。

4.讨论了篇章多级依存结构在自动文摘中的优势，介绍了基本概念和实现方法，并在潜在语义分析的基础上进行结构分析，克服了潜在语义分析的不足，使生成的文摘更加流畅，包含了文章的主要内容。

5.2 存在的问题和以后的研究方向

该系统对某些文本取得了比较好的效果，但其中还存在一些问题，下一步的工作还可以从以下几个方面展开:

1.在今后的工作中将深入研究语言生成问题，结合未登陆词识别、领域自动判别等技术，进一步改进生成的自动摘要的质量，使其更接近人工摘要的自然性、

流畅性。

2.提高算法的效率，以解决两种方法融合后在时间和空间开销较大的问题。

3.将现有成果与网页分类、信息检索、信息过虑结合起来进行研究。

参考文献

姚天顺，朱靖波，杨莹等《自然语言理解》北京：清华大学出版社，2002，P1－7

陈群秀，《一个在线义类词库：WordNet》，语言文字应用，1998，P345

董振东，《语义关系的表达和知识系统的建造》，语言文字应用，1998，P67

冯志伟，《自然语言的计算机处理》，上海外语教育出版社，1996，P256－300

石安石，《语义论》，商务印书馆，1993，P218－230

冯志伟，《论歧义结构的潜在性》，中文信息学报，1995，9，P143

石纯一，黄昌宁等，《人工智能原理》，清华大学出版社，1993，P159

A. Mathis, Techniques for the Eutluation and Imprauement of Computer Priduced Abstracts. Ohio State Universtiy. Dec. 1972, PB214675

Luhn, H.P., The automatic creation of literature abstracts, IBM J. of Res.And Development, 1958，P159-165

[10]H. P. Edmundson.New methods in automatic abstracting. Journal of the Association for Computing Machinery ,1969 ,16(2) :264～285

[11]马希文 ,李小滨 ,徐越 .自然语言处理与自动文摘.智能技术与系统基础 ,1988 ,99～117

[12]L. F. Rau, P. S. Jacobs and Uri Zernik. Information Extracting and Text Summarization Using Lingustic Knowledge Acquistion. Information Processing & Management ,1989 ,25(4) :419～428

[13]P. S.Jacobs and L. F. Rau. Scisor: Extracting Information from OnLine News. Communication of the ACM ,1990 ,33(11) :88～97

[14]G. Salton ,J.Allan ,C.Buckleym and Amit Singhal.Automatic Analysis,Theme Genaration ,and Summarization of Machine2Readable Texts. Science ,1994 ,264(3) :1421～1426

[15]刘挺,王开铸 自动文摘的四种主要方法[J] . 情报学报,1999 ,18(1) :10 - 191.

[16]R.Brandow ,K.Mitze and L. F. Rau.Automatic Condensation of Electronic Publications by Sentence Selection.Information Processing & Management ,1995 ,31(5) :675～685

[17]哈罗德·博,查尔斯·L·贝尼埃合著 ,赖茂生 ,王知津合译.文摘的概念与方法.书目文

献出版社 ,1991

[18]D. Fum ,G. Guida and C. Tasso. Forward and Backward Reasoning in Automatic Abstracting. COLING 82 ,1982 ,83～88

[19]K.Ono ,K. Sumita and S. Miike.Abstract Generation Based on Rhetorical Structure Extraction. COLING 94 ,Kyoto ,August 5～9 ,1994 ,344～348

[20]T.Maeda.An Approach toward Functional Text Structure Analysisof Scientific and Technical Documents. Information Processing & Management ,1981,17(6) :329～339

[21]M W Berry,S T Dumais,G W O ’Brien.Using Linear Algebra for Intelligent Information Retrieval [J].SIAM Review ,December 1995.

[22]LANDAU ER T K, DUMA IS S T. A solution to Plato’s problem: the latent semantic analysis theory of the acquisition, induction and representation of know ledge[J]. Psychol Rev, 1997, 104: 211 240.

[23]Landauer T K,et al.A Solution to Plato ’s Problem:The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge [J] .Psychological Review,1997 ,104:211 2240.

[24]S.Deerwester, S.Dumais, G.Furnas,T.Landauer,and R.Harshman.“Indexing by latent semantic analysis, ” Journal of the American Society of Information Science,vol.41,pp.391-407,1990.

[25]M.Berry, S.Dumais, and G.O’Brien, “Using linear algebra for intelligent information retrieval, ” Tech.Rep.UT-CS-94-270,University of Tennessess, Computer Science Department,Dec.1994.

[26]林鸿飞,姚天顺.基于潜在语义索引的文本浏览机制[J] . 中文信息学报,2000 ,14(5) :49256.

[27]刘挺，王开铸.基于篇章多级依存结构的自动文摘研究[J].计算机研究与发展，1999,36（4）479：－488.

[28]Karen Sparck Jones , etc. Automatic Summarizing Factors and Directions Advances in Automatic Text Summarization , Cambridge MA :MIT Press ,1998.

[29]Jen-Yuan Yeh , Hao-Ren Ke , Wei-Pang Yang , I-Heng Meng. Text summarization using a trainable summarizer and latent semantic analysis，Information Processing and Management 41 (2005) 75－95

致　　谢

衷心感谢我的导师杨国纬教授，感谢他三年来的辛勤培养，悉心帮助。杨老师敏锐的思维、渊博的知识给我留下了深刻的印象，他严谨的治学态度和高尚的品德让我终生受益匪浅。

同时感谢教研室的刘启和、程博、李凡、卜强、颜俊华等师兄和师姐；感谢教研室的所有同学：王慧慧、李立燕、丁豪、周波、杨宾伟、李立、任君燕等，他们对我的工作给予了很大的支持和帮助。

感谢我的家人对我在读书期间的支持和帮助!

最后衷心感谢为评阅本文而付出辛勤劳动的各位专家和学者!

附　录 1

语料 1. 未来人类：基因开道

【原文】在新的世纪，新的研究成果——基因为你带来新的生活：

人的寿命将大大延长，拥有更健康的身体。工作到８０岁才退休，平均寿命将达１２０岁。

人类和癌症等众多尚无治疗手段的“不治之症”说“再见”，人生病的机率大大降低，因为许多疾病刚“露头”，就被检测出来而被消灭。

投资热点将从电脑等信息产业转向基因产业。基因“比尔·盖茨”将横空出世，基因公司炙手可热，笑傲美国华尔街。

有得必有失。人类在得到很多益处的同时，也将失去人的“隐私”权。找工作，要有基因证明；买保险，也要出示基因材料。最终，你将发现，你离开了基因，你寸步难行。

不过，基因仍然是利大于弊。未来人类：基因为你开道……。

【10％文摘】
在新的世纪，新的研究成果——基因为你带来新的生活：

【20％文摘】
在新的世纪，新的研究成果——基因为你带来新的生活。有得必有失。

【30％文摘】
在新的世纪，新的研究成果——基因为你带来新的生活。投资热点将从电脑等信息产业转向基因产业。有得必有失。

语料 2. 香港各界群起抨击吕秀莲

【原文】本报香港 4 月 8 日电。死硬的“台独”分子吕秀莲，4 月 2 日在接受香港记者采访时，狂妄地宣扬其“台独”主张，引起香港各界人士的愤怒。他们群起抨击吕秀莲的荒谬主张，指出她的这番言论，是背弃自己的祖先，蓄意向全体中

国人挑战，这样的民族败类应当人人共讨之。

全国政协委员、香港社团服务中心董事总经理刘迺强说，台湾新领导人当选后，中央政府抱着“听其言，观其行”的态度，观察其后的发展。吕秀莲所发表的言论赤裸裸地暴露了其鼓吹“台独”的嘴脸。她认为台湾与祖国大陆是“远亲与近邻”的关系，台湾民众是“国共内战”的替罪羊。这与事实根本是不相符的。台湾是中国的一部分，台湾人就是中国人，本来就是一家人，而不是什么“亲”和“邻”的关系。在目前海峡两岸关系极为敏感的情况下，她发表的“台独”言论，只会恶化两 岸关系。世界上只有一个中国，就是中华人民共和国，台湾是中国的一部分，已是举世公认的事实，是任何人也无法改变的。她的这番话改变不了这个事实，只能引起台湾同胞和全体华夏子孙对她的厌恶和反感。鼓吹“台独”的人，绝没有好下场。

全国政协委员、恒通资源集团有限公司董事长施子清说，吕秀莲最近对香港媒体发表的谈话，是不负责任的胡言乱语，她根本否定了自己是炎黄子孙，把祖先都忘了。她分裂祖国的企图已暴露无遗。现在全世界都承认台湾是中国的领土，过去国共两党曾经有分有合、政见不一，但是都没有偏离“一个中国”的原则。台湾新领导人为国家民族计、为台湾人民计、为自身的政治前途计，都应当回到“一个中国”的原则上来，为改善两岸关系做点有益的事情。他说，香港已成功实行“一国两制”，中央政府不干预特区自治范围内的事情，回归后的香港比过去更民主，经济发展上更得到祖国内地的大力支持。台湾如能实现与祖国大陆的和平统一，台湾人民同样可以维护自己的利益，发展的空间会更广阔。

香港理工大学的陈文鸿认为，吕秀莲在这个时候发表这种言论，有其险恶用心，企图离间两岸同胞血肉相连的关系，制造仇恨情绪，以制造机会，企图在其大权在手时，公开向祖国大陆挑战。吕秀莲声称“台湾是一个主权独立的国家”，广大香港市民是不能接受的，相信也不能为大多数台湾同胞所接受。国家主权不可分割。台湾与祖国大陆目前虽然还没有统一，但毕竟是一个国家的内部问题。吕秀莲的说法，是她一贯鼓吹“台独”理念的流露。她的荒谬主张是全体中国人不能接受的。

香港中西区区议员陈财喜说，吕秀莲发表的“台独”言论，是非常危险的，这简直是“玩火”。香港市民都认为“台独”是非常危险的，祖国大陆对台湾问题的立场是非常明确的，承认一个中国的原则，一切都好商量，否则，一切无从谈起。

【10％摘要】

本报香港 4 月 8 日电。死硬的“台独”分子吕秀莲，4 月 2 日在接受香港记者采访时，狂妄地宣扬其“台独”主张，引起香港各界人士的愤怒。全国政协委员、香港社团服务中心董事总经理刘迺强说，台湾新领导人当选后，中央政府抱着“听其言，观其行”的态度，观察其后的发展。台湾新领导人为国家民族计、为台湾人民计、为自身的政治前途计，都应当回到“一个中国”的原则上来，为改善两岸关系做点有益的事情。

【20％摘要】

本报香港 4 月 8 日电。死硬的“台独”分子吕秀莲，4 月 2 日在接受香港记者采访时，狂妄地宣扬其“台独”主张，引起香港各界人士的愤怒。全国政协委员、香港社团服务中心董事总经理刘迺强说，台湾新领导人当选后，中央政府抱着“听其言，观其行”的态度，观察其后的发展。吕秀莲所发表的言论赤裸裸地暴露了其鼓吹“台独”的嘴脸。台湾新领导人为国家民族计、为台湾人民计、为自身的政治前途计，都应 当回到“一个中国”的原则上来，为改善两岸关系做点有益的事情。香港已成功实行“一国两制”，中央政府不干预特区自治范围内的事情，回归后的香港比过去更民主，经济发展上更得到祖国内地的大力支持。

【30％摘要】

本报香港 4 月 8 日电。死硬的“台独”分子吕秀莲，4 月 2 日在接受香港记者采访时，狂妄地宣扬其“台独”主张，引起香港各界人士的愤怒。全国政协委员、香港社团服务中心董事总经理刘迺强说，台湾新领导人当选后，中央政府抱着“听其言，观其行”的态度，观察其后的发展。吕秀莲所发表的言论赤裸裸地暴露了其鼓吹“台独”的嘴脸。她认为台湾与祖国大陆是“远亲与近邻”的关系，台湾民众是“国共内战”的替罪羊。台湾新领导人为国家民族计、为台湾人民计、为自身的政治前途计，都应当回到“一个中国”的原则上来，为改善两岸关系做点有益的事情。香港已成功实行“一国两制”，中央政府不干预特区自治范围内的事情，回归后的香港比 过去更民主，经济发展上更得到祖国内地的大力支持。香港市民都认为“台独”是非常危险的，祖国大陆对台湾问题的立场是非常明确的，承认一个中国的原则，一切都好商量，否则，一切无从谈起。

语料 3. 吃快餐也要讲究健康

【原文】2004 年伊始，进入中国 16 年的肯德基开办了它在中国的第 1000 家餐厅，与此同时，他们也把经过长期酝酿的《健康食品政策白皮书》公布于世，面对快餐都是垃圾食品的质疑，中国肯德基极具前瞻性的姿态，变被动为主动。吃快餐也要讲究健康的倡议，让大众对于饮食的认识从吃饱、吃好逐渐向吃得健康转变。有一个正确的认识是饮食健康、科学的前提，毕竟吃油炸鸡肉还是吃水果沙拉最终选择权在消费者手中。

近年来，在美国这样的发达国家，很多著名快餐企业在健康问题上都受到指责，原因是发达国家社会中越来越严重的肥胖以及由此产生的一系列健康问题。不良饮食结构引起的健康问题正是经济迅速发展的中国即将、甚至是已经需要面对的难题。

根据 1999－2000 年全美健康营养调查显示，美国 65％的成年人体重超过正常水平，患有肥胖症的成人达到 31％。预计肥胖症比例 2008 年将达到 39％。美国疾病防治中心主任认为，肥胖是当代美国最大的健康问题之一，而更令人担忧的是青少年肥胖问题也日趋严重，1999－2000 年全美健康营养调查显示 15％的儿童和青少年有肥胖症。肥胖给美国带来许多严重的社会问题，首先肥胖的最大危害是导致其他疾病发病率增加，在过去的 10 年中，美国人 2 型糖尿病发病率急剧上升，在某些社区中，占了糖尿病新发病例总数的 50％以上。2002 年，美国卫生总署进行的一项调查显示，每年肥胖症及其相关疾病上投入的费用已经达到 1170 亿。究竟谁是这个问题的始作俑者，大众争论不休，遗传因素，运动过少，内分泌失调，精神紧张、压力过大引起的神经、精神因素，生理因素，环境因素等都是导致肥胖的原因，但体重迅速增加的关键因素首推饮食习惯。更多人认为快餐是导致肥胖的罪魁祸首，遍布社区的快餐店、便宜的价格、充足的分量都对消费者有很强的诱惑力，而快餐营养成分的局限性使不少人开始认同它为“垃圾食品”。

虽然饮食不健康造成的肥胖问题在中国远不如在美国那样严重，但根据中国卫生部疾病控制司 2003 年的报告显示，中国的肥胖症患病率近年来呈上升趋势，而且增长速度迅速。目前体重超重者已经达到 22.4％，其中 3.01％为肥胖者。肥胖症预防和控制已经成为中国公共卫生事业刻不容缓的任务。

中国肥胖发病率与经济发展有密切关系，其总的规律是大城市高于小城市；中小城市高于农村；随着中国向工业化社会飞速发展，首先膳食结构发生了很大改变，传统的膳食结构以粮食为主，副食主要是新鲜食品，不做精细加工，甜食

摄入量较少。对于大多数中国人来说，一日三餐中摄取的热量有 65％来自谷物，动物性食品占的比例较少。不过近些年来，中国人的体力活动也在减少，职业性体力劳动和家务劳动量减轻，静态生活的时间增加。例如美国北卡罗来纳大学的科林•贝尔教授 曾经对中国社会进行了一项调查，研究交通方式的改变对肥胖的影响，结果表明在拥有汽车的家庭肥胖比例要比没有汽车的家庭高 80％。

对于中国人来说，最重要的是如何越过发达国家已经经历的肥胖系列问题，多年来中国肯德基已经开始致力于饮食健康的科学研究。2000 年肯德基在中国成立了食品健康咨询委员会，聘请了 8 位食品健康专家，委员会定期为公司介绍食品和营养学科方面的最新动向和科研成果，对健康食品的研发提供建议和指导。肯德基认为，消费者对于食物的最基本要求是安全、卫生，而与洋快餐相比，目前中国很多中小型餐饮企业在安全和卫生方面并不能让人满意；第二建立良好的饮食和生活习惯是保持、提高身体健康的重要组成部分，肯德基多年来一直也将继续以全国的 1000 家餐厅为窗口，向消费者介绍营养健康小知识；第三有益于健康的食品仍然需要味道可口，肯德基始终将产品研发放在重要位置，上世纪九十年代中就成立了自己的产品研发团队和实验厨房。

【10％文摘】

2004 年伊始，进入中国 16 年的肯德基开办了它在中国的第 1000 家餐厅，与此同时，他们也把经过长期酝酿的《健康食品政策白皮书》公布于世，面对快餐都是垃圾食品的质疑，中国肯德基极具前瞻性的姿态，变被动为主动。吃快餐也要讲究健康的倡议，让大众对于饮食的认识从吃饱、吃好逐渐向吃得健康转变。近年来，在美国这样的发达国家，很多著名快餐企业在健康问题上都受到指责，原因是发达国家社会中越来越严重的肥胖以及由此产生的一系列健康问题。

【20％文摘】

2004 年伊始，进入中国 16 年的肯德基开办了它在中国的第 1000 家餐厅，与此同时，他们也把经过长期酝酿的《健康食品政策白皮书》公布于世，面对快餐都是垃圾食品的质疑，中国肯德基极具前瞻性的姿态，变被动为主动。有一个正确的认识是饮食健康、科学的前提，毕竟吃油炸鸡肉还是吃水果沙拉最终选择权在消费者手中。有一个正确的认识是饮食健康、科学的前提，毕竟吃油炸鸡肉还是吃水果沙拉最终选择权在消费者手中。近年来，在美国这样的发达国家，很多著名快餐企业在健康问题上都受到指责，原因是发达国家社会中越来越严重的肥胖以及由此产生的一系列健康问题。不良饮食结构引起的健康问题正是经济迅速

发展的中国即将、甚至是已经需要面对的难题。

【30％文摘】

2004 年伊始，进入中国 16 年的肯德基开办了它在中国的第 1000 家餐厅，与此同时，他们也把经过长期酝酿的《健康食品政策白皮书》公布于世，面对快餐都是垃圾食品的质疑，中国肯德基极具前瞻性的姿态，变被动为主动。有一个正确的认识是饮食健康、科学的前提，毕竟吃油炸鸡肉还是吃水果沙拉最终选择权在消费者手中。近年来，在美国这样的发达国家，很多著名快餐企业在健康问题上都受到指责，原因是发达国家社会中越来越严重的肥胖以及由此产生的一系列健康问题。不良饮食结构引起的健康问题正是经济迅速发展的中国即将、甚至是已经需要面对的难题。美国疾病防治中心主任认为，肥胖是当代美国最大的健康问题之一，而更令人担忧的是青少年肥胖问题也日趋严重，1999－2000 年全美健康营养调查显示 15%的儿童和青少年有肥胖症。肥胖给美国带来许多严重的社会问题，首先肥胖的最大危害是导致其他疾病发病率增加，在过去的 10 年中，美国人 2 型糖尿病发病率急剧上升，在某些社区中，占了糖尿病新发病例总数的 50%以上。虽然饮食不健康造成的肥胖问题在中国远不如在美国那样严重，但根据中国卫生部疾病控制司 2003 年的报告显示,中国的肥胖症患病率近年来呈上升趋势，而且增长速度迅速。

附　录 2

攻读硕士研究生期间曾发表的论文

张峰，杨国纬，《基于潜在语义分析和结构分析的自动文摘方法》.四川师范大学学报自然科学版增刊.成都.2005.10