摘　　要

随着社会的进步和互联网技术的发展，网络信息量的频繁剧增，当今社会面临着信息大爆炸。当大量的信息像潮水般涌向人们时，传统人工处理信息的手段已经远远不足。为了解决这一问题，科学界提出文摘自动生成的技术。

自动文摘通常被视为自然语言处理的一项任务。文摘是准确全面地反映某一文章中心内容的简洁连贯的短文，与索引相比更能满足信息获取的要求。我国对自动文摘技术的研究目前还在初级阶段，但此技术所具有的重要作用是不可低估的，必将在未来的信息处理领域得到广泛的应用。

本论文基于现阶段的研究现状下，运用统计自然语言处理方法，首先对文章进行自动分词，利用停用词表对分词结果进行过滤，并利用知网（HowNet）获得概念，建立概念向量空间模型。通过计算词语重要度和句子重要度，系统得到一个粗略的文摘。最后再进行冗余计算，得到本文章的文摘。

本文在上述研究的基础上，设计了基于概念向量空间模型的自动文摘系统，实现了机器自动生成文摘的各个模块的功能，证实了本文利用概念统计的方法比基于词频统计的方法得到的文摘，能更准确含概原文章的中心内容。

关键词：自动文摘　知网　概念向量空间模型　　自然语言处理

ABSTRACT

Along with the advancement of society and technology of the World Wide Web is developing, the information of the network is growing exponentially, society is facing exploding of the information nowadays. When the large volume of information emerge people like tidewater, it is too deficiency to use tradition human professional to dispose the information. In order to resolve this problem, the science domain advance the technology of text Automatic Summarization.

Automatic Summarization usually is regarded as a item task of nature language. Summarization can express a certain article's center content accurately and whole, it is composed by some succinct and coherent sentences. Compare with index, summarization can satisfy the request of information-obtained. Researching the technology of automatic summarization of our country is in a elementary phase yet, but the significant function of this technology cannot underestimate, and it must be extensively used in future information disposal domain.

Aiming at the present situation, this paper uses statistical nature language disposal method, it carrys out automatic participle firstly, uses cease word list to filtrate the result of automatic participle, and obtains the conception by using HowNet, to establish the conceptual vector space model. By carrying out the weight of word and sentence, system can get a summary abstract. And it accounts the redundancy to obtain this paper's summarization finally.

This paper bases on the research above-mentioned, it devises a system of automatic summarization based on conceptual vector space model, it realizes computer automatic summarization's function of every module. And this paper approves that basing on conceptual statistical method is better than word frequency statistical method, it can contain original text center content more exactly.

Key words: Automatic Summarization HowNet Conceptual Vector Space Model Nature Language Disposal

长春理工大学硕士学位论文原创性声明

本人郑重声明：所呈交的硕士学位论文，《自动文摘技术的研究与应用》是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：梁晓丽 2007年3月27日

长春理工大学学位论文版权使用授权书

本学位论文作者及指导教师完全了解“长春理工大学硕士、博士学位论文版权使用规定”，同意长春理工大学保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅。本人授权长春理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，也可采用影印、缩印或扫描等复制手段保存和汇编学位论文。

作者签名：梁晓丽 2007年3月27日

指导导师签名：[illegible] 2007年3月27日

第一章 绪 论

§1.1 自动文摘的研究意义和相关概念

§1.1.1 研究意义

随着信息时代的到来，电子文本的大量涌现和 Internet 网的广泛使用，人们在欣然享受着海量信息所带给我们的资讯震撼的同时，开始逐渐意识到要想在这信息的海洋中迅速有效地找到满足自己特定需求的信息是多么的困难和无助，因而迫切渴望能借助一些有效的工具来应对这场信息过载危机。信息的过滤、搜集与综合成为极有潜力的研究课题，而这些智能业务中最引人注目的便是自动文摘。它的实现基础是要构建一个文本理解系统。本文构建了一个文本理解系统，并以系统最后生成文摘的质量作为评判系统理解能力的依据。

自动摘要是计算机语言学和情报科学共同关注的课题，其本质是信息的挖掘和信息的浓缩。从理论上讲，对自动摘要的研究将有助于探讨人类理解、概括自然语言文本，并从中获取知识的认识模型。自动摘要被认为是计算机实现自然语言理解的重要标志之一。从应用角度讲，在文献电子化和 Internet 迅速发展的今天，自动摘要系统的使用将大幅度降低编制摘要的成本，缩短文摘的出版周期，为人们廉价、迅速和准确地获得所需要的信息提供方便。

信息检索技术的出现在一定程度上缓解了信息过载的压力。然而，鉴于现有的信息检索技术所能达到的信息查询的准确率和召回率还差强人意，与人们的实际需求还相距甚远。因此，如何能从众多检索结果，尤其是以文本形式存在的成千上万的检索结果当中行之有效地找到与用户的当前需求最相关的信息便成为了一个众所关注的热点问题。

自动文摘作为解决当前信息过载问题的一种辅助手段，正日益受到国内外学术界和工业界的密切关注，从近年来频繁召开的有关自动文摘的专题学术会议、工作组以及评测大赛就可窥见一般。

自动文摘研究之所以如火如荼的开展，关键就在于研究人员已经充分意识到它能在一定程度上弥补信息检索技术在应对信息过载危机时所表现出来的种种缺憾。这种弥补具体表现在以下两个方面：

一、质量良好的文摘能在一定程度上取代原始文本的被检索地位，作为原始文本的一个替代品参与检索，从而能有效地缩减检索信息的时

间；

二、质量良好的文摘能用于检索结果的可视化，使得用户无需浏览原始的大量检索结果便能轻松地取舍信息，从而能有效地节省信息的浏览时间，提高需求信息的命中率。

由此可见，自动文摘必将为辅助解决当前日趋严重的信息过载问题而提供越来越成熟的技术支持和更加强劲的应用保障。

自动文摘是一类特殊的自然语言理解问题。语言的层面模型观点指出，语言具有三个主要层面：结构层面、意义层面和功能层面。由于对语言各层面的研究至今尚很不充分，自动文摘就难免面临诸方面难以逾越的障碍。首先在意义层面上，由于语言可以有许多比喻性用法，对其意义进行了不同的引申，语句里各词的词义不是几个范畴能包括的，故准确地把握语言的意义十分困难，其次在功能层面上，由于语言的功能过于广泛致使歧义问题十分突出。因此，基于目前的语言研究水平，只有采取一些避开这些困难的有效对策才能使当前对自动文摘的研究不至于重蹈旧辙。

§1.1.2 文摘相关概念及目的

文摘是准确全面地反映某一文献中心内容的简洁连贯的短文，与索引相比更能满足细心获取的要求。所谓自动文摘就是利用计算机自动地从原始文献中提取文摘[1]。

文摘可分为：

1)指示型文摘：对原文内容的一种指示性的介绍，不涉及到具体的细节内容。其目的在于帮助用户做出是否需要对原文做深入阅读的判断；

2)信息型文摘：提供对原文细节内容的一种浓缩的表达，以帮助用户仅通过阅读文摘便能抓住原文的核心内容，从而大大地节省阅读的时间，提高阅读的效率；

3)评论型文摘：提供对原文内容的一种评论，以帮助用户了解原文作者想要表达的主观意图。

进行自动文摘的主要目的是：（1）自动文摘是表明文章主题的一个摘要内容。当出现在文章的第一页或仅以摘要形式被作为一个链接时，它可以明确的表达出文章撰写的主要目的。这样可以使读者很快地肯定或否定这篇文章是否是他们感兴趣的内容，而决定是否需要去读其中的详细内容；（2）当文摘被建立索引时，可以让读者很快找到自己真正需要的相关文章，而不必将时间浪费在不相关文章的阅读上；（3）当文摘被搜索引擎标记上域信息后，可以使用户进行的搜索更加高效，

以在最短的时间里找到与查询关键字相关内容的文档列表。

§1.1.3 国内外研究现状

关于自动文摘的研究，起始于1958年IBM公司的H.P. Luhn所做的工作[2]。到目前为止，已经有国内外众多学者和研发机构投入到此项富有市场前景和研究价值的课题中来，并取得了一系列丰硕的成果。国内对自动文摘的研究起始于80年代末，上海交通大学王永成教授领导的课题组所做的工作是当时的典型代表[3]。目前我国在该领域的研究仍处于初级阶段，尚有很大的发展空间。

纵观自动文摘的研究历程，归纳起来可以分为三个主要的发展阶段[4]、两种主流的研究方法以及两种广泛采用的评价策略[5]，现详述如下：

1）三个主要的发展阶段

阶段一：50年代末~60年代末

代表性的工作：(Luhn,1958)， (H.P.Edmundson,1969)等。

Luhn于1958年发表了世界上第一篇关于计算机自动编制文摘的经典论文“The Automatic Creation of Literary Abstracts”，从此揭开了自动文摘研究的序幕[2]。他提出了一种基于关键词频率统计的自动文摘方法，即通过统计文本中的内容词的词频来描述内容词的重要度，并利用文本句子中包含的所有内容词的重要度来给各个句子打分，从中挑选出得分最高的若干句子构成摘要。他的伟大贡献在于首次提出了一种基于文本浅层特征统计的自动文摘方法，并将著名的Zipf定律成功地应用到自动文摘研究领域，取得了令人瞩目的效果。

1969年，Edmundson在Luhn提出的基于关键词频率统计的自动文摘方法的基础上，进一步提出了一个重要的改进设想[6]。即将文本的关键词、标题、位置以及提示词这四种浅层特征联合起来考虑，并通过对它们的综合统计来给每个句子打分，这个分值就作为句子重要性的度量值。他还系统地比较了综合应用这四种特征加权的方式所产生的摘要的效果，结果发现标题——位置——提示词综合加权策略取得了最好的摘要效果，而单纯使用关键词加权则效果最差。

总之，在自动文摘研究的早期，单纯的基于文本浅层特征的统计学方法占据了研究的主导地位，并曾一度统治了相当长一段时期。国内上海交通大学王永成教授所领导的课题组于1997年成功研制出中文自动文摘系统OA[7]。该系统在原理上就是综合采用了以上介绍的多种浅层特征集成的句子打分法，只是它主要针对的是中文文本而非英文文本。

阶段二：70年代初~80年代末

代表性的工作[8]：(Schank,1974)，(Dejong,1979)，(J.I.Tait,1982)，

(DaniloFUM,1982)，(Hahn,1988)，(Lisa F.Rau, 1989)等。

在 1974 年，耶鲁大学的 Schank 研制了 SAM 自动文摘系统。该系统采用脚本来分析简单的故事，并对故事进行归纳摘要[9]。

耶鲁大学的 Dejong 于 1979 年研制出了著名的 FRUMP 自动文摘系统。该系统利用语法知识来判定某个预期词在句子当中的位置，并通过句法分析来遍历整个文本以寻找标示为已知脚本的短语，从而建立起各种故事的梗概[10]。

1982 年，J.I.Tait 对原有的 FRUMP 系统进行了改进。他提出将所有的资料先转换成概念依存结构，然后再在此基础上通过分析、推测各种信息之间的关系来构成摘要[10]。

意大利 Udine 大学的 Danilo FUM 等研究人员在 1982 年成功研制出了 SUSY 文摘系统。该系统以一阶谓词逻辑作为文本的机内表达形式，利用纲要产生器和分析缩写器来装配出满足特定需求的摘要[11]。

德国康斯坦大学的 Hahn 等研究人员于 1988 年研制出 TOPIC 自动文摘系统，该系统针对的是微处理器领域的科技文本，它采用框架作为知识的载体，并通过联合语法、语义分析来生成各种长度的文摘[8]。

1989 年，美国 GE 研发中心的 Lisa F.Rau 等科研人员研制出了 SCISOR 自动文摘系统。该系统利用篇章主题分析以及复杂的句法结构分析等技术生成与摘要有关的框架概念，并采用某种预期驱动分析器从所有框架概念当中提取出预期内容，构成摘要。该系统主要处理的是“公司合并”方面的新闻[12]。

总之，在这个阶段，以人工智能技术，深层自然语言处理技术以及知识工程技术为代表的自动文摘方法逐渐占据了该领域的主导地位。在国内，哈尔滨工业大学的王开铸教授领导的课题组于 1992 年研制出的中文自动文摘实验系统 MATAS，即采用基于了深层自然语言处理的方法[13]。此外，哈尔滨工业大学的刘挺教授于 1996 年提出的中文自动文摘系统的设计方案即是采用上述基于信息抽取的框架知识表达来实现的[14]。北京邮电大学的钟义信教授领导下的课题组也充分利用了上述基于自然语言处理和知识工程的方法开发出了面向特定领域的中文自动文摘系统模型 Ladies[15]，该系统主要处理的是有关计算机病毒方面的中文文本，并取得了不错的效果。与之类似的还有东北大学与香港城市理工大学联合开展的有关自动文摘方面的研究，他们提出的中文自动文摘系统通过脚本来存储知识，通过用户交互手段来生成最终的摘要[16]。

阶段三：90 年代初~至今

代表性的工作：(Salton et al,1994)， (Kupiec et al,1995)，(Lin&Hovy,1997) ，(Jaime Carbonell&Jade Goldstein,l998)，(Yihong Gong&Xin Liu,2001)，(Conroy&Oleary,2001)等。

Salton 等研究人员在 1994 年通过统计文本段落之间的共享词汇数来计算段落之间的语义关联，构造文本的语篇结构图来辅助文本话语结构的自动分析，从而提出了基于语篇话语结构分析的抽取型自动文摘方法[17]。国内与之类似的工作是南京大学的王继成等研究人员在 2003 年所提出的基于篇章结构指导的中文 Web 文档自动摘要方法[18]。

1995 年，Kupiec 等研究人员开创了将机器学习技术用于自动文摘领域的先河[19][20]。他们采用基于朴素 Bayesion 理论的机器学习方法从科技论文和论文摘要的语料库中提取出对抽取重要句子有贡献的联合特征，并在此基础上充分利用已获得的联合特征来从科技文本中抽取一定数量的句子以构成摘要。

Lin 和 Hovy 在 1997 年尝试了用机器学习方法验证句子位置这一自然语言处理领域惯用的浅层特征对文摘句选取质量的影响[21]。

Jaime Carbonell 和 Jade Goldstein 在 1998 年探讨了如何将文本中包含的概念多样性引入到自动文摘的研究当中，从而使产生的摘要能尽可能地覆盖原文多个概念并包含较少的冗余。具体做法是通过采用一种称为最大边缘相关(MMR)的摘要模型来实现的[22]。

哈尔滨工业大学的刘挺等研究人员在 1999 年提出了一种基于篇章多级依存结构分析的自动文摘方法[23]，并通过实验验证了该方法的可行性和有效性。

Yihong Gong 和 Xin Liu 两位研究人员在 2001 年提出了两种句子抽取型的自动文摘方法[24]。一种是基于相关性度量策略，另一种是基于潜在语义分析(LSA)算法。基于相关性度量的文摘方法，它挑选文摘句的策略在于：先循环计算每个句子和文本之间的语义相似度，从中挑选出相似度最大的那个句子放入摘要。然后从剩余的句子集合中依次去掉已包含在刚入选摘要的那个句子中的所有词语，再通过重新计算剩余的句子和文本之间的语义相似度来选择出下一个具有最大相似度的句子入选进最终的摘要。而基于潜在语义分析的文摘方法则通过对句子—词语矩阵做 SVD 分解，进而挑选出分解结果矩阵的对角线上若干最大特征值所对应的句子入选最终的摘要。

2001 年，Conroy 和 Oleary 两位研究人员尝试了将隐马尔可夫模型引入自动抽取型摘要的研究当中[25]。

2001 年，上海交通大学的研究人员还尝试了以心理语言学为基础，构造基于主题敏感词分析的新闻文献自动摘要系统[26]。

总之，从 90 年代初至今，自动文摘研究在经历了相当长一段时期的发展之后，正朝着面向实用化，面向非受限领域文本处理的方向迈进，进入到一个前所未有的高潮期。与此同时，各种新颖的研究思想、研究成果和热点课题层出不穷。但总的来说，占主导地位的研究方法又逐渐

回归到以统计学的方法为主，以深层自然语言处理、信息抽取以及基于本体的知识工程方法为辅的混和型方法上了[27]。

2） 两种主流研究方法：

方法一：基于抽取的研究方法(Extraction Method)

尽管自动文摘的研究是从基于抽取的研究方法开始的，然而目前的绝大多数工作仍然采用了基于抽取的方法来从原文本中抽取句子或更大的语言单元以构成摘要，只是在具体的抽取方法上有所改进。从最初的单纯依靠原文本浅层特征的句子抽取方法逐渐过渡到采用更加复杂的句子抽取策略，如基于语料库的机器学习方法[28]，基于文本主题结构分析的方法[29]以及基于文本修饰辞分析的自动文摘方法[30]等。

方法二：基于泛化生成的研究方法(Abstraction Method)

近期，基于泛化生成的自动文摘方法获得了不少研究人员的关注，并取得了一定的成果。该方法主要利用了信息抽取、信息压缩、信息融合等多种泛化生成的核心技术。

信息抽取技术的思路主要表现在：通过预定义信息槽来存放待抽取信息。如针对计算机病毒类的文章，预定义信息槽往往设计为包括病毒名、发作时间、解决办法等；然后利用计算机自动地在原文本中定位有关的信息片断，最后将这些片断填充到各个对应的槽中以产生结果摘要。该技术的优点在于能产生较高质量的准确摘要，但缺点也不容忽视，那就是它的应用领域严格受限且开发这类文摘系统所需的代价昂贵。

信息压缩和信息融合技术的特点在于：充分利用了现有的自然语言产生技术来改造文本中的相关句子，并在一定程度上构造出新的句子。该技术具有代表性的工作是 Knight，Kevin 和 Marcu 在 2000 年所发布的研究成果[31]。他们采用了基于期望最大化的估计方法训练系统模型中的参数，然后通过训练阶段所获得的参数来产生相关的规则，并将它们用于压缩句子的句法分析树，从而产生出原文本的一个精简的文摘版本，而该版本所包含的每个句子能在最大程度上符合语法规范。

据统计，目前绝大多数的自动文摘方法往往都致力于基于抽取的文摘方法[32](即采用 Extraction 的文摘方法)，而非基于泛化生成的文摘方法[33](即采用 Abstraction 的文摘方法)。一方面，这是由理性的自然语言理解技术和知识工程技术的高度复杂性及其应用领域的严重受限性所造成；另一方面，这也与近年来统计学的研究方法、机器学习的研究方法以及模式识别的研究方法在自然语言处理一系列应用领域中所取得的不俗成绩密不可分。

基于抽取的文摘方法按抽取办法的不同可大致分为有指导型和无指导型。有指导型抽取方法的实现依赖于大量人工做的标准摘要，即业类俗称的金标准“Gold Standards”来帮助训练和确定摘要统计学模型

的特征参数。然而，由于人工摘要的置信度问题至今仍是一个悬而未决的问题，因而在很大程度上促使了研究人员对无指导型文摘办法的研究。而无指导型的文摘办法，其最大优势就在于：它的实现无需人工摘要的支持，仅从文本自身出发，利用统计学方法和启发式规则来确定文本中各个句子的权值并依此来挑选出文摘句。该办法还可以进一步被细分为无篇章结构分析型和基于篇章结构分析型。前一种办法的通常做法是:先给原文本包含的所有句子打分，然后挑选出得分最高的若干句子，并按照这些句子在原文中出现的语序先后关系依次输出它们以构成摘要。但细心的研究人员很快发现采用这种方法产生的文摘不仅主题覆盖不全而且冗余偏大，它往往只能抽取出文章中分布密度较大的主题，而忽视了其它主题的存在。针对此问题，南京大学的王继成等提出了基于篇章结构分析型的自动文摘方法，他们通过文本中相邻段落的用词重叠统计来计算相邻段落之间的语义距离，从而得出文章主题的一种划分。最后从各个划分好的主题下抽取出适量的句子来构成摘要。这种方法在处理篇章结构比较规范的文本时效果比较好，能有效地解决无篇章结构分析型文摘方法所凸显出的上述问题。然而，令人遗憾的是，当文本的写作风格比较自由，且主题分布灵活多样时，即一个主题可能分布在不相邻的若干个段落当中。在这种情况下，采用此方法的效果则会大打折扣。

3)两种广泛采用的评价策略

策略一：Intrinsic evaluation

这是基于摘要自身质量的一种直接式的评价策略。

策略二：Extrinsic evaluation

这是一种间接式的评价策略，即让摘要在自然语言处理的其它应用当中去取代其对应文本的原始地位，从而通过对该应用效果的影响程度来间接评价摘要的质量。

自动文摘的评价是一个非常棘手的问题，国内外学术界一直在努力探索着，力求寻找到一种行之有效的解决方案，但到目前为止似乎离预想中的目标还有相当长的一段距离，不过这也正好促使了对自动文摘的评价这一经典难题的前所未有的关注。一系列自动文摘领域颇具影响力的评价比赛正在受到越来越多的科研机构和研究人员的大力支持，而这必将促进自动文摘技术的蓬勃发展。

美国的 SUMMAC, DUC，日本的 TSC 以及中国的 863 计划中文信息处理与智能人机接口技术评测系列之自动文摘任务便是此类评价比赛中的典型代表。

§1.2 本课题研究的内容

关于自动文摘系统的研究，主要有基于意义的理解文摘和基于统计的机械文摘两种主要的研究方法。关于它的理论的研究远远滞后于信息社会中信息处理的发展要求。

产生这种现象的主要原因是由于基于意义的理解文摘和基于统计的机械文摘系统都存在着一些弊端。如对于基于意义的理解文摘，由于知识库建立的困难性，知识表示的复杂性，使得它只能面向某一应用领域，并且文摘质量并不十分令人满意;对于基于统计的机械文摘，大多采用的是基于词形统计的向量空间模型法。这种方法以词形作基础，认为词形是文章的最小意义单元。但是向量空间模型最基本的假设是向量各义项之间要正交，也就是意义不相关，而在真实文本中，存在着相当多的一词多义与一义多词现象，使作为义项的词语之间往往有很大的相关性。从而导致文摘的质量不高。

为此，我们提出了基于知网（HowNet）概念获取算法得到文本的主题语义概念，建立概念向量空间模型。这样，可以使得向量空间模型中各向量义项间保持正交关系，从而提高向量空间模型进行自动文摘的各项效能。

本课题研究主要内容包括文本词语的计算机处理、词语所表达概念的自动获取和句子语义相似度的计算分析，以及文本主题句的提取的研究。文本将基于统计的机械文摘、基于 HowNet 的词语概念获取和主题句和主题语义相似度计算等研究方法结合起来提高了文摘的质量。

§1.3 本文内容组织

本文各章安排如下：

第一章绪论，概述了文本自动文摘的意义和应用背景，介绍了文本自动文摘的国内外研究现状，以及研究存在的问题和提出的相关技术。

第二章是文本自动文摘模型的介绍，综述了当前文本自动文摘领域几种重要的模型，并作了相应的比较和分析。

第三章是基于概念向量空间模型的中文自动文摘研究。这一部分是文论文的核心部分，介绍了词语概念获取的主要工具 HowNet，阐明了建立一个稳定、可靠、高效的自动文摘系统里面的各项关键技术的实现。

第四章是系统实现与试验分析，介绍了自动文摘系统各模块以及相应模块所实现的功能；此外，对该系统进行了全面的测评，主要是通过

各种不同的方法和测试手段对设计实现的系统进行评估，指出了存在的问题以及初步的解决方案。

第五章对全文进行总结并展望了未来的工作。

第二章　自动文摘的相关模型

§2.1 向量空间模型

在自然语言处理的各个研究领域，对文本各级语言单元进行形式化的表达是一个既基础而又重要的问题。而形式化表达其根本目的就在于力图将各种无结构化的文本单元转换成便于计算机处理的结构化的表达形式，以支持后续一系列语言处理应用的需要。

向量空间模型(Vector Space Model)是 20 世纪 60 年代由 Gerard Salton 等人提出的。主要应用于信息检索、自动索引、分类、聚类、篇章分析等。其思想是把文本表示成向量空间中的点(称为向量)，用向量之间的夹角余弦作为文本间的相似度度量。当向量空间模型用于文本检索时，首先要建立文本和用户查询的向量表示，然后进行查询向量和文本向量间的相似度计算。

§2.1.1 文本向量空间表示[34]

对于计算机来说，中文文本就是由汉字和标点符号等最基本的语言符号组成的字符串，由字构成词，由词构成短语，进而形成句、段、节、章、篇等语言结构。用尽量简单并且准确的方法表示文档，是进行文本检索的前提。

在向量空间模型(VSM：Vector Space Model)中，文本的各级语言单元被映射成 N 维向量空间中的对应向量，而各个向量则通过文本中的特征的重要度来形式化表达。值得注意的是，这里所谈到的文本中的特征既可以指文本中所包含的字、词，也可以是更加复杂的特征，如概念、句法结构等，至于具体选用什么样的特征往往与实际的应用需求有关，不可一概而论。

VSM 表示方法是在文本中提取其特征项组成特征向量，并以某种方式为特征项赋权，如：文档 D 可表示成 $D(T_1,T_2,\cdots,Tn)$，其中 T_k 是特征项，$1<k<N$。由于特征项的重要程度不同，可用附加权重 W_k 来进行量化，这样文档 D 可表示为 $D(T_1,W_1;\ T_2,W_2;\ \cdots,T_N,W_N)$，简记为 $D(W_1,W_2,\cdots,W_N)$。这时说项 T_k 的权重为 W_k，$1\leqslant k\leqslant N$。如果把 T_1, $T_2,\cdots,T_N$ 看成是一个 n 维坐标系，而 W1, W2, …, W_n 是相应的坐标值，则 $D(W_1,W_2,\cdots,W_N)$被看成是 n 维空间中的一个向量。称 $D(W_1,W_2,\cdots,W_N)$为文本 D 的向量表示。

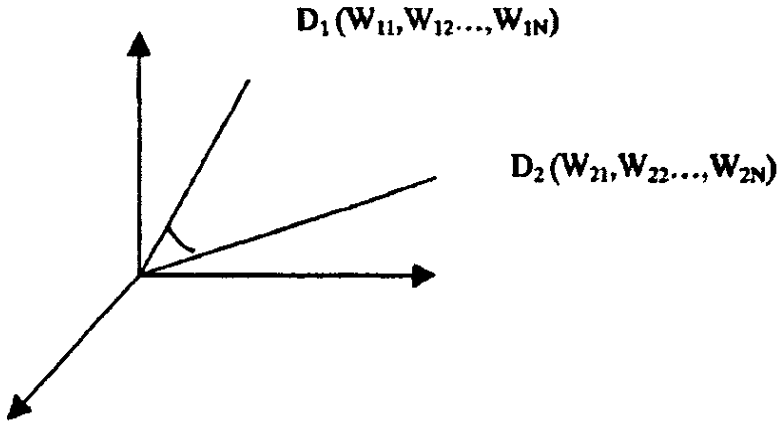

图 2.1　文本的向量空间表示

可以看出，对向量空间模型来说，有两个基本问题：即特征项的选择和项的权重计算。

§2.1.2 特征项选择

用来表示文档内容的项可以是各种类别，对汉语来说，有字、词、短语，甚至是句子或句群等更高层次的单位。项也可以是相应词或短语的语义概念类。

项的选择必须由处理速度、精度、存储空间等方面的具体要求来决定。特征项选取有几个原则：一是应当选取包含语义信息较多，对文本的表示能力较强的语言单位作为特征项；二是文本在这些特征项上的分布应当有较为明显的统计规律性，这样将适用于信息检索、文档分类等应用系统；三是特征选取过程应该容易实现，其时间和空间复杂度都不太大。实际应用中常常采用字、词或短语作为特征项。

§2.1.3 特征项权重计算

对于特征项权重的计算，经典的 tf. idf[35]方法考虑两个因素：1)词语频率 tf (term frequency)：词语在文档中出现的次数；2)词语倒排文档频率 idf(inverse document frequency)：该词语在文档集合中分布情况的一种量化，常用的计算方法是 $\log_2(N/n_k+0.01)$，其中 N 为文档集合中的文档数目，n_k 为出现该词语的文章数。

根据以上两个因素，可以得出公式：

$$W_{ik} = tf_{ik} \times \log_2(N/n_k + 0.01)$$

其中 tf_{ik} 为词语 T_k 在文档 D_i 中出现的次数，W_{ik} 为词语 T_k 在文档 D_i 中的权值，k=1,2,...,m(m 为词的个数)。

为了计算方便，通常要对向量进行规一化，最后由：

$$W_{ik} = \frac{tf_{ik} \times \log_2(N/n_k + 0.01)}{\sqrt{\sum_{k=1}^{m}(tf_{ik} \times \log_2(N/n_k + 0.01))^2}} \qquad (2.1)$$

以上公式的提出是基于这样一个考虑：对区别文档最有意义的特征词应该是那些在文档中出现频率足够高而在文档集合中的其它文档中出现频率足够少的词语。

§2.1.4 文本间的相似度度量

向量空间模型中的另一个概念是相似度(Similarity)：相似度 $Sim(D_1, D_2)$ 用于度量两个文档 D_1 和 D_2 之间的内容相关程度。当文档被表示为文档空间的向量，就可以利用向量之间的距离计算公式来表示文档间的相似度。常用的距离有向量的内积距离：

$$Sim(D_1, D_2) = \sum_{k=1}^{n} W_{1k} \times W_{2k} \qquad (2.2)$$

$$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)}} \qquad (2.3)$$

将 VSM 应用于不同的领域，其相似度的计算有所不同。例如，对于信息检索来讲，VSM 采用向量间的某种距离度量来反应文本对查询的满足程度。所有相似度的值最后能与真实情况相符，计算简便，同时最好能归一化到[0, 1]区间上，并且分布尽可能的均匀，使阈值的选择容易一些。

§2.2 存在的问题

从向量空间模型的特点可以看出，在特征项确定的情况下，特征项的权重计算是文档分类的关键，特征项权重计算常用的方法有布尔函数、开根号函数、对数函数、TFIDF 函数等，其中 TFIDF 函数应用最为广泛，其基本思路是使用频率因子 TF（Term Frequency）进行特征项的赋权，同时还要考虑文档集因子 IDF（Inverse Document Frequency），体现出查询内容与文档的相关度大小，一般采用使用出现频率的倒数来计算，$IDF = \log(N/n_i)$，其中 N 为文档集合，n_i 为查询内容在文档中出现的次数，但是 TFIDF 函数也存在缺点，它虽然考虑了出现特征项的文本在整个文档集中的比例，却不能很好地把握特征项在文本集合中分布的差异，所以影响了分类的最终效果。

VSM 的第一个问题是由于特征项在文档中的不同位置会代表不同的权重，而不同的关键词长度也会影响权重的大小。例如“汽车修理”一词在查询时，如果该词出现在文档的标题处，则其权重一定比出现在文章的摘要中要高，而出现在摘要中的权重一定要比出现在正文中要高；而且如果文档 D_1 的长度比文档 D_2 长，那么在 D2 中的权重也应该比 D1 要高，其相似度也应该大一些，对于中文文档，关键词的长度越长，则在文档中出现的机率就越小，所以较长的关键词要比较短的包含更多的信息。在实际情况中，如果同一特征项在不同文档中出现的次数不同，那么在出现频率较高的文档中，其权重应该较高（而不应该是统一权重值“1”），在传统的 TFIDF 函数中，每增加一个文档都要重新计算向量，导致查询速度降低，同时由于使用频率因子，在扩大查询范围时，不可避免的会影响到查询的准确性。

VSM 的另一个问题在于查询和文档向量间是依靠链接来判断的，而且判断的依据中简单的两者相同关键词的比较，但实际情况是，大量的关键词具有相同的语义，同一关键词也会有多种语义的解释描述（即产生了语义分歧）。例如“计算机”一词，也可以是“电脑”、“微机”等，对用户来说所指的可能是一个意思，但在 VSM 中这几个词是完全不同的概念。

这里用改进的 VSM 方法。可以看出，传统的 VSM 主要的缺陷就是特征项相互独立的要求与自然语言多样性的矛盾。实际上我们主要考虑两个方面的改进，一个是关键词的长度和出现在文档中的位置对权重的影响；另一个就是要考虑关键词的语义环境影响。

§2.3 加权的VSM算法改进

为了解决特征项在文本集合中分布的差异，提出改进的加权 VSM 算法，公式如下：

$$W_i = \lambda \times tf_i \times \log\left(\frac{N}{n_i} + 0.1\right) + \frac{tf_i}{l_i} \qquad (2.4)$$

其中λ为位置加权系数，表示文本在文档不同位置的加权处理参数，按照文本在文档中的位置不同，一般分为标题、摘要、关键词、正文、结论和超链接等6个位置，分别赋予不同的加权系数，由于Web文档信息都是通过链接来完成的，Web上的各种标记和链接包含了页面的结构信息，应该给予足够的重视和利用。

例如:在链接 r→s 中，r 的连接标记若为文档 D_1 锚文本文档 D_2 其中锚文本对目标 URL="http://www.china…" 会有比较准确的描述，而文档 D_1,D_2 就次之，所以对于出现在锚文本和文档 D_1，D_2 中的每一个特征项应赋于较高的权重系数。

另外一个关键的加权位置在一些语义的重点语句位置，如“综上所述”、“结束语”、“主要在于”等关键语句中，其值可以从辅助主题词表中获取（具体解释见后）。一般位置加权系数λ的计算可以考虑使用各分部分的频率与不同位置加权系数的乘积和来表示。

$$\lambda = tf_0 + tf_1 \times \lambda_1 + tf_2 \times \lambda_2 + tf_5 \times \lambda_5 + tf_4 \times \lambda_4 + tf_6 \times \lambda_6$$

其中 tf_0 为对正文关键词统计的词频数；tf_1，tf_2，tf_3，tf_4，tf_5 别为标题、摘要、关键词、超链接中的词频；λ_1，λ_2，λ_3，λ_4，λ_5 分别为其加权系数。

tf_i 为特征项频率；N为总文档数量；n_j 为包含特征项 W_i 的文档数；l_i 为文档长度，使用 $\frac{tf_i}{l_i}$ 来表示文本能够代表文档内容的能力，例如虽然“计算机”一词出现在文档标题和正文中的频率相同，但由于标题比正文文档长度要小的多，所以我们认为“计算机”一词在标题中的权重要比在正文中的权重要大的多。

第三章 基于概念向量空间模型关键技术的研究

§3.1 自动分词技术的研究

机器不同于人，它不可能智能地读懂文章内容。当然，我们在读文章时，也是从组成这篇文章的基础词着手，明白各个句子的的含义，再概括出各段落的大意，最后得出文章的中心思想。对一篇文章的处理，我们先从自动分词开始。下面，现介绍一下自动分词的算法。

§3.1.1 自动分词算法

我们可以将现有的分词算法分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

1、基于字符串匹配的分词方法

这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。

按照扫描方向的不同，串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大（最长）匹配和最小（最短）匹配；按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的几种机械分词方法如下；

1）正向最大匹配

正向最大匹配法是最早提出的自动分词方法，它的基本思想是先取一句话的前六个字查字库，若不是一个词，则删除六个字的最后一个字再查，这样一直查下去，至找到一个词为止。句子剩余部分重复此工作，直到把所有的词都分出为止。

2）逆向最大匹配

逆向最大匹配法也一样，不同的是它是从句子的最后六个字开始的，每次匹配不成功时去掉汉字串中最前面的一个字。

在应用中，方法有所变化。如下述算法我们初始不是取六个字而是取长度最短词的个数。

A1：一条汉语语句分划成单一字符 X1，X2，…，X。

A2：决定语词中可能出现的词最大字符长度 L_{max}，最小字符长度

L_{min}。

A3：逆向匹配，取语句最后的编 m 个字查词库，若查不到，加入一个字重复工作，直至字符数为 L_{max} 为止。

A4：若实施 A3 查不到词，去掉语句中最后一个字，再实施 A3，直至整个语句只剩下 L_{min} 为止。

3）最少切分（使每一句中切出的词数最小）

还可以将上述各种方法相互组合，例如，可以将正向最大匹配方法和逆向最大匹配方法结合起来构成双向匹配法。由于汉语单字成词的特点，正向最小匹配和逆向最小匹配一般很少使用。一般说来，逆向匹配的切分精度略高于正向匹配，遇到的歧义现象也较少。单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245。（这可能是因为汉语的中心语靠后的特点。）但这种精度还远远不能满足实际的需要。由于分词是一个智能决策过程，机械分词方法无法解决分词阶段的两大基本问题：歧义切分问题和未登录词识别问题。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

一种方法是改进扫描方式，称为特征扫描或标志切分，优先在待分析字符串中识别和切分出一些带有明显特征的词，以这些词作为断点，可将原字符串分为较小的串再来进行机械分词，从而减少匹配的错误率。

另一种方法是将分词和词类标注结合起来，利用丰富的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整，从而极大地提高切分的准确率。

对于机械分词方法，可以建立一个一般的模型，形式地表示为 ASM(d,a,m)，即 Automatic Segmentation Model。其中：

d：匹配方向，+1 表示正向，-1 表示逆向；

a：每次匹配失败后增加/减少字串长度（字符数），+1 为增字，-1 为减字；

m：最大/最小匹配标志，+1 为最大匹配，-1 为最小匹配。例如，ASM(+, -, +)就是正向减字最大匹配法（即 MM 方法），ASM(-, -, +)就是逆向减字。

最大匹配法(即 RMM 方法)。对于现代汉语来说，只有 m=+1 是实用的方法。用这种模型可以对各种方法的复杂度进行比较，假设在词典的匹配过程都使用顺序查找和相同的计首字索引查找方法，则在不记首

字索引查找次数（最小为 log<汉字总数> =12 - 14）和词典读入内存时间的情况下，对于典型的词频分布，减字匹配 ASM(d,-,m)的复杂度约为 12.3 次，增字匹配 ASM(d,+,m)的复杂度约为 10.6。

2、基于理解的分词方法

通常的分析系统，都力图在分词阶段消除所有歧义切分现象。而有些系统则在后续过程中来处理歧义切分问题，其分词过程只是整个语言理解过程的一小部分。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

3、基于统计的分词方法

从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息为： $M(X,Y)=\log P(X,Y)/P(X).P(Y)$，其中 P(X,Y)是汉字 X、Y 的相邻共现概率，P(X) 、P(Y)分别是 X、Y 在语料中出现的概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计，不需要切分词典，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，并且对常用词的识别精度差，时空开销大。实际应用的统计分词系统都要使用一部基本的分词词典（常用词词典）进行串匹配分词，同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

§3.1.2 自动分词歧义问题及初步处理

自动分词过程中歧义产生一般有三种情况：

(1)由自然语言的二义性产生的歧义。例如：“在日本保留和尚使用的古典乐器很多”。这句若没有上下文辅助，连人也难理解其真实含义，

计算机程序肯定出现两种分词情况：

在/日本/保留/和/尚使用/的/古典/乐器/很多；

在/日本/保留/和尚/使用/的/古典/乐器/很多。

(2)由计算机程序分词产生的歧义。这种情况虽然人可以正确分词，但计算机毕竟不是人，出现歧义难免。计算机程序分词产生的歧义一般有两种：组合型歧义。即，对于字串 AB，可以分成 AB，也可以分成 A / B；交集型歧义。即，对于字串 ABC，可以分成 AB/C，也可以分成 A/BC。

(3)由词典大小产生的歧义自动分词必须借助词典，若词典中没有的词，就不可能正确分词。

三种歧义的解决第一种最难，可以说目前还没有好的方法，好在统计表明这类歧义只占歧义总数的 5%左右；第二，三种怎样解决是研究的热点。我们给出一种初步研究的方法：

步骤 1：待分词的句子用正向最大匹配法和逆向最大匹配法初步自动分词。

步骤 2：比较两个分词结果，若结果一致，正确而分词结束；否则，继续步骤 3。

步骤 3：比较词数，若不等，选词数较少的一个作为分词结果；相等，继续步骤 4。

步骤 4：比较未登录词词数，若不等，选词数较少的一个作为分词结果；相等，继续步骤 5。

步骤 5：查找规则库，用规则进一步确定分词结果。

说明：步骤 2 中两个分词结果一致一般就是正确的；步骤 3 的情况是根据组成长词的情况可能性比例很高作为依据；步骤 4 的解释雷同于步骤 3；步骤 5 的情况比较复杂，目前规则主要考虑具体词之间邻接的可能性、词类之间的邻接概率，还需要进一步研究。

§3.2 知网

自然语言处理系统最终需要更强大的知识库的支持。知识库是机器获得理解的知识来源。知网（英文名称为 HowNet）[36]是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。它是近期用以解决自然语言理解中语义问题的一种常用途径和方法。

§3.2.1 知网的特色

计算机化是知网的重要特色。知网是面向计算机的，是借助于计算机建立的，将来可能是计算机的智能构件。

知网作为一个知识系统，实副其名是一个网而不是树。它所着力要反映的是概念的共性和个性，例如：对于“医生”和“患者”，“人”是它们的共性。知网在主要特性文件中描述了“人”所具有的共性，那么“医生”的个性是他是“医治”的施事，而“患者”的个性是他是“患病”的经验者。对于“富翁”和“穷人”，“美女”和“丑八怪”而言，“人”是它们的共性。而它们的个性，即：“贫”、“富”与“美”、“丑”等不同的属性值，则是它们的个性。

同时知网还着力要反映概念之间和概念的属性之间的各种关系。知网把下面的一种知识网络体系明确的教给了计算机进而使知识对计算机而言是可操作的[37-44]。

总的来说，知网描述了下列各种关系：

(a) 上下位关系 (由概念的主要特征体现)

(b) 同义关系

(c) 反义关系

(d) 对义关系

(e) 部件-整体关系（由在整体前标注 % 体现，如“心”，“CPU”等）

(f) 属性-宿主关系（由在宿主前标注 & 体现，如“颜色”，“速度”等）

(g) 材料-成品关系（由在成品前标注 ? 体现，如“布”，“面粉”等）

(h) 施事/经验者/关系主体-事件关系（由在事件前标注 * 体现，如“医生”，“雇主”等）

(i) 受事/内容/领属物等-事件关系（由在事件前标注 $ 体现，如“患者”，“雇员”等）

(j) 工具-事件关系（由在事件前标注 * 体现，如“手表”，“计算机”等）

(k) 场所-事件关系（由在事件前标注 @ 体现，如“银行”，“医院”等）

(l) 时间-事件关系（由在事件前标注 @ 体现，如“假日”，“孕期”等）

(m) 值-属性关系（直接标注无须借助标识符，如“蓝”，“慢”等）

(n) 实体-值关系（直接标注无须借助标识符，如“矮子”，“傻瓜”等）

(o) 事件-角色关系（由加角色名体现，如“购物”，“盗墓”等）

(p) 相关关系（由在相关概念前标注 # 体现，如“谷物”，“煤田”等）

知网的一个重要特点是：类似于同义、反义、对义等种种关系是借助于“同义、反义以及对义组的形成”，由用户自行形成而不是逐一地、显性地标注在各个概念之上的。

知网是一个知识系统，而不是一部语义词典。尽管被我们称为知识词典的常识性知识库是知网的最基本的数据库。知网的全部的主要文件包括知识词典构成了一个有机结合的知识系统。例如，主要特征文件、次要特征文件、同义、反义以及对义组的形成，以及事件关系和角色转换等都是系统的重要组成部分，而不仅仅是标注的规格文件。我们预计用户将来把它们与知识词典一起加以利用。

§3.2.2 义原的提取与考核

知网是一个以上述各类概念为描述对象的知识系统。知网不是一部义类词典。知网是把概念与概念之间的关系以及概念的属性与属性之间的关系形成一个网状的知识系统。这是它与其他的树状的词汇数据库的本质不同。知网的哲学和它的根本特性决定了它的特别的建设方法。

我们首先要理解“知网”中两个主要的概念：“概念”与“义原”。“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。

什么是义原，跟什么是词一样的难以定义。但是也跟词一样并不因为它难于定义人们就无法把握和利用它们。大体上说，义原是最基本的、不易于再分割的意义的最小单位。例如：“人”虽然是一个非常复杂的概念，它可以是多种属性的集合体，但我们也可以把它看作为一个义原。我们设想所有的概念都可以分解成各种各样的义原。同时我们也设想应该有一个有限的义原集合，其中的义原组合成一个无限的概念集合。如果我们能够把握这一有限的义原集合，并利用它来描述概念之间的关系以及属性与属性之间的关系，我们就有可能建立我们设想的知识系统。利用中文来寻求这个有限的集合，应该说是个捷径。中文中的字（包括单纯词）是有限的，并且它可以被用来表达各种各样的单纯的或复杂的概念，以及表达概念与概念之间、概念的属性与属性之间的关系。

在初步确定了一批义原并形成了一个基本的标注集之后，如何加以考核和确定？

第一、在扩大标注中观察该义原的覆盖面。我们有一条原则：我们已有的义原一定要能够描述全部的概念。这里有一个比较硬性的规定，即当我们发现一个具有多个概念的词语，例如八个，而我们以有的义原不能够把这八个概念区别开来时，我们就必须对我们的标注集加以调整，这是绝大多数的情况。在很个别的情况下我们不排除怀疑其中某个概念是否存在，以决定取舍。

第二、观察某一个义原在概念之间关系中的地位。如果一个义原在同类别的许多概念中出现或者不同类别的概念中出现，那么这样的义原就是稳定的义原是一个必须确定的义原。以事件类“医治”这个义原为例，它不仅出现在“医”、“治”、“治疗”、“医疗”、“治病”、“求医”、“看病”等概念中，并且还出现在“医生”、“医院”、“医药”、“诊所”、“不治之症”、“有病乱投医”。因此，“医治”这个义原是稳定的、是必须确定的。

§3.2.3 知网系统的概貌

1、知网系统包括下列数据文件和程序：

(01) 知网管理系统

(02) 中英双语知识词典

知网的规模主要取决于双语知识词典数据文件的大小。由于它是在线的，修改和增删都很方便，因此它的规模是动态的。它的规模通常以词语的条数以及由词语所表述的概念的条数计算。

2、知识词典的记录样式

知识词典是知网系统的基础文件。在这个文件中每一个词语的概念及其描述形成一个记录。每一种语言的每一个记录都主要包含 4 项内容。其中每一项都由两部分组成，中间以“=”分隔。每一个“=”的左侧是数据的域名，右侧是数据的值。它们排列如下：

W_X= 词语

E_X= 词语例子

G_X= 词语词性

DEF= 概念定义

3、关于词语的例子

迄今为止，我们主要是为那些具有多个义项提供例子。这些例子的要求是：强调例子的区别能力而不是它们的释义能力。它们的用途在于为消除歧义提供可靠的帮助。这里试以“打”的两个义项为例，一个义

项是“buy/买”，另一个是“weave/辫编”。

NO.=000001
W_C=打
G_C=V
E_C=~酱油，~张票，~饭，去~瓶酒，醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy/买

NO.=015492
W_C=打
G_C=V
E_C=~毛衣，~毛裤，~双毛袜子，~草鞋，~一条围巾，~麻绳，~条辫子
W_E=knit
G_E=V
E_E=
DEF=weave/辫编

其中 NO.为概念编号，W_C，G_C，E_C 分别是汉语的词语、词性和例子，W_E、G_E、E_E 分别是英语的词语、词性和例子，DEF 是知网对于该概念的定义，我们称之为一个语义表达式。其中 DEF 是知网的核心。我们这里所说的知识描述语言也就是 DEF 的描述语言。

设我们要判定的歧义语境是“我女儿给我打的那副手套哪去了”。我们通过对“手套”与“酱油”等的语义距离的计算以及跟“毛衣”等的语义距离的计算的比较，我们将会得到一个正确的歧义判定结果。这种方法的好处有二：第一，多数的判定可以避免采用规则；第二，多数的情况基本的算法可以是不依赖特定语言的。

4、HowNet 知识描述语言

运用 HowNet 里面所带的知识库作为对词语意义赋值的重要资源，通过处理可以得到 HowNet 里面的一些有用信息。其格式(将此格式定义为 HowNettool)描述如下：

W_X=词语
G_X=词语的词性
DEF=词语的定义
我们看几个例子：

表 3.1“知网”知识描述语言实例

打	017144	exercise|锻练,sport|体育
男人	059349	human|人,family|家,male|男
高兴	029542	aValue|属性值,circumstances|境况,happy|福,desired|良
生日	072280	time|时间,day|日,@ComeToWorld|问世,$congratulate|祝贺
写信	089834	write|写,ContentProduct=letter|信件
北京	003815	place|地方,capital|国都,ProperName|专,(China|中国)
爱好者	000363	human|人,*FondOf|喜欢,#WhileAway|消闲
必须	004932	{modality|语气}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|钥匙)
从良	016251	cease|停做,content=(prostitution|卖淫)
打对折	017317	subtract|削减,patient=price|价格,commercial|商,(range|幅度=50%)
儿童基金会	024083	part|部件,%institution|机构,politics|政,#young|幼,#fund|资金,(institution|机构=UN|联合国)

从这些例子我们可以看到，“知网”的知识描述语言是比较复杂的。我们将这种知识描述语言归纳为以下几条：

1) “知网”收入的词语主要归为两类，一类是实词，一类是虚词；
2) 虚词的描述比较简单，用“{句法义原}”或“{关系义原}”进行描述；
3) 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下三种形式：
   独立义原描述式：用“基本义原”，或者“(具体词)”进行描述；
   关系义原描述式：用“关系义原=基本义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”来描述；
   符号义原描述式：用“关系符号 基本义原”或者“关系符号(具体词)”加以描述；
4) 在实词的描述中，第一个描述式总是一个基本义原，这也是对该实词最重要的一个描述式，这个基本义原描述了该实词的最基本的语义特征。

除了义原以外，”知网”中还用了一些符号来对概念的语义进行描述，如下表所示：

表 3.2“知网”知识描述语言中的符号及其含义

,	多个属性之间，表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
$	表示“可以被该‘V’处置，或是该“V”的受事，对象，领有物，或者内容
*	表示“会‘V’或主要用于‘V’，即施事或工具
+	对V类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，如对于布匹，我们标以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 对于V类，置于［］中的是该类V所有的“必备角色”。如对于“购买”类，一旦它发生了，必然会在实际上有如下角色参与：施事，占有物，来源，工具。尽管在多数情况下，一个句子并不把全部的角色都交代出来 (2) 表示动态角色，如介词的定义
()	置于其中的应该是一个词表记，例如，(China|中国)
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，例如：“味道”对于“食物”，“高度”对于“山脉”，“温度” 对于“天象”等
[]	标识概念的共性属性

我们把这些符号又分为几类，一类是用来表示语义描述式之间的逻辑关系，包括以下几个符号：, ~ ^，另一类用来表示概念之间的关系，包括以下几个符号： # % $ * + & @ ? !，第三类包括几个无法归入以上两类的特殊符号：{} () []。

我们看到，概念之间的关系有两种表示方式：一种是用“关系义原”来表示，一种是用表示概念关系的符号来表示。按照我们的理解，前者类似于一种格关系，后者大部分是一种格关系的“反关系”，例如“$”

我们就可以理解为“施事、对象、领有、内容”的反关系，也就是说，该词可以充当另一个词的“施事、对象、领有、内容”。

§3.2.4 基于“知网”的语义相似度计算方法[45]

（一）词语相似度与度量词语关系

什么是词语相似度？

我们认为，词语相似度是一个主观性相当强的概念。脱离具体的应用去谈论词语相似度，很难得到一个统一的定义。因为词语之间的关系非常复杂，其相似或差异之处很难用一个简单的数值来进行度量。从某一角度看非常相似的词语，从另一个角度看，很可能差异非常大。

不过，在具体的应用中，词语相似度的含义可能就比较明确了。例如，在基于实例的机器翻译中，词语相似度主要用于衡量文本中词语的可替换程度；而在信息检索中，相似度更多的要反映文本或者用户查询在意义上的符合程度。

本文的研究主要以基于实例的机器翻译为背景，因此在本文中我们所理解的词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。

相似度是一个数值，一般取值范围在[0,1]之间。一个词语与其本身的语义相似度为 1。如果两个词语在任何上下文中都不可替换，那么其相似度为 0。

相似度这个概念，涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中，对词语相似度影响最大的应该是词的语义。

度量两个词语关系的另一个重要指标是词语的距离。

一般而言，词语距离是一个[0,∞)之间的实数。

一个词语与其本身的距离为 0。

词语距离与词语相似度之间有着密切的关系。

两个词语的距离越大，其相似度越低；反之，两个词语的距离越小，其相似度越大。二者之间可以建立一种简单的对应关系。这种对应关系需要满足以下几个条件：

1) 两个词语距离为 0 时，其相似度为 1；
2) 两个词语距离为无穷大时，其相似度为 0；
3) 两个词语的距离越大，其相似度越小（单调下降）。

对于两个词语 W_1 和 W_2，我们记其相似度为 $Sim(W_1,W_2)$，其词语距

离为 $Dis(W_1,W_2)$，那么我们可以定义一个满足以上条件的简单的转换关系：

$$Sim(W_1,W_2)=\frac{\alpha}{Dis(W_1,W)+\alpha} \tag{3.1}$$

其中 α 是一个可调节的参数。α 的含义是：当相似度为 0.5 时的词语距离值。

这种转换关系并不是唯一的，我们这里只是给出了其中的一种可能。

在很多情况下，直接计算词语的相似度比较困难，通常可以先计算词语的距离，然后再转换成词语的相似度。所以在本文后面的有些章节，我们只谈论词语的距离，而没有提及词语的相似度，读者应该知道这二者是可以互相转换的。

度量两个词语关系的另一个重要指标是词语的相关性。

词语相关性反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性来衡量。

词语相关性也是一个[0,1]之间的实数。

词语相关性和词语相似性是两个不同的概念。例如“医生”和“疾病”两个词语，其相似性非常低，而相关性却很高。可以这么认为，词语相似性反映的是词语之间的聚合特点，而词语相关性反映的是词语之间的组合特点。

同时，词语相关性和词语相似性又有着密切的联系。如果两个词语非常相似，那么这两个词语与其他词语的相关性也会非常接近。反之，如果两个词语与其他词语的相关性特点很接近，那么这两个词一般相似程度也很高。

（二）语义相似度计算方法

从上面的介绍我们看到，与传统的语义词典不同，在“知网”中，并不是将每一个概念对应于一个树状概念层次体系中的一个结点，而是通过用一系列的义原，利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。我们的目标是要找到一种方法，对用这种知识描述语言表示的两个语义表达式进行相似度计算。

利用“知网”计算语义相似度一个最简单的方法就是直接使用词语语义表达式中的第一独立义原，把词语相似度等价于第一独立义原的相似度。这种方法好处是计算简单，但没有利用知网语义表达式中其他部分丰富的语义信息。

1）词语相似度计算

对于两个汉语词语 W_1 和 W_2，如果 W_1 有 n 个义项（概念）：S_{11}，S_{12}，……，S_{1n}，W_2 有 m 个义项（概念）：S_{21}，S_{22}，……，S_{2m}，我们规定，W_1 和 W_2 的相似度各个概念的相似度之最大值，也就是说：

$$Sim(W_1,W_2)=\max_{i=1\ n,j=1\ m} Sim(S_{1i},S_{2j}) \tag{3.2}$$

这样，我们就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。当然，我们这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语，最好是先进行词义排岐，将词语标注为概念，然后再对概念计算相似度。

2）义原相似度计算

由于所有的概念都最终归结于用义原（个别地方用具体词）来表示，所以义原的相似度计算是概念相似度计算的基础。

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，我们这里采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d，根据公式(3.1)，我们可以得到这两个义原之间的语义距离：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{3.3}$$

其中 p_1 和 p_2 表示两个义原（primitive），d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数。

用这种方法计算义原相似度的时候，我们只利用了义原的上下位关系。实际上，在“知网”中，义原之间除了上下位关系外，还有很多种其他的关系，如果在计算时考虑进来，可能会得到更精细的义原相似度度量，例如，我们可以认为，具有反义或者对义关系的两个义原比较相似，因为它们在实际的语料中互相可以互相替换的可能性很大。对于这个问题这里我们不展开讨论，留给以后的研究工作来处理。

另外，在知网的知识描述语言中，在一些义原出现的位置都可能出现一个具体词（概念），并用圆括号()括起来。所以我们在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的相似度计算。理想的做法应该是先把具体词还原成“知网”的语义表达式，然后再计算相似度。这样做将导入函数的递归调用，甚至可能导致死循环，这会使算法会变得很复杂。由于具体词在“知网”的语义表达式中只占很小的比例，因此，在我们的实验中，为了简化起见，我们做如下规定：

- 具体词与义原的相似度一律处理为一个比较小的常数（γ）；
- 具体词和具体词的相似度，如果两个词相同，则为 1，否则为 0。

3）虚词概念的相似度的计算

我们认为，在实际的文本中，虚词和实词总是不能互相替换的，因此，虚词概念和实词概念的相似度总是为零。

由于虚词概念总是用“{句法义原}”或“{关系义原}”这两种方式进行描述，所以，虚词概念的相似度计算非常简单，只需要计算其对应的句法义原或关系义原之间的相似度即可。

但由于虚词对于文章内容的表达不是很重要，本文可通过停用此表把那些与文章中心内容无关的虚词过滤，再对分词结果进行概念划分。

4）实词概念的相似度的计算

由于实词概念是用一个或几个语义表达式来描述的，在计算两个实词的相似度时，要分别比较每个实词的语义表达式，综合权衡这两个实词概念的相似度，因此其相似度计算变得非常复杂。

如何计算两个语义表达式的相似度呢？

我们的基本设想是：整体相似要建立在部分相似的基础上。把一个复杂的整体分解成部分，通过计算部分之间的相似度得到整体的相似度。

假设两个整体 A 和 B 都可以分解成以下部分：A 分解成 A_1，A_2，……，A_n，B 分解成 B_1，B_2，……，B_m，那么这些部分之间的对应关系就有 $m \times n$ 种。问题是：这些部分之间的相似度是否都对整体的相似度发生影响？如果不是全部都发生影响，那么我们应该如何选择那些发生影响的那些部分之间的相似度？选择出来以后，我们又如何得到整体的相似度？

我们认为：一个整体的各个不同部分在整体中的作用是不同的，只有在整体中起相同作用的部分互相比较才有效。例如比较两个人长相是否相似，我们总是比较它们的脸型、轮廓、眼睛、鼻子等相同部分是否相似，而不会拿眼睛去和鼻子做比较。

因此，在比较两个整体的相似性时，我们首先要做的工作是对这两个整体的各个部分之间建立起一一对应的关系，然后在这些对应的部分之间进行比较。我们把这种做法比喻成古代的战场的两军对垒：兵对兵、将对将，捉对厮杀。

还有一个问题：如果某一部分的对应物为空，如何计算其相似度？我们的处理方法是：

- 将任何义原（或具体词）与空值的相似度定义为一个比较小的常数（δ）；

整体的相似度通过部分的相似度加权平均得到。

对于实词概念的语义表达式，我们将其分成四个部分：

1) 第一独立义原描述式：我们将两个概念的这一部分的相似度记为 $Sim_1(S_1,S_2)$；

2) 其他独立义原描述式：语义表达式中除第一独立义原以外的所有其他独立义原（或具体词），我们将两个概念的这一部分的相似度记为 $Sim_2(S_1,S_2)$；

3) 关系义原描述式：语义表达式中所有的用关系义原描述式，我们将两个概念的这一部分的相似度记为 $Sim_3(S_1,S_2)$；

符号义原描述式：语义表达式中所有的用符号义原描述式，我们将两个概念的这一部分的相似度记为 $Sim_4(S_1,S_2)$。

于是，两个概念语义表达式的整体相似度记为：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \tag{3.4}$$

其中，β_i（$1\leqslant i\leqslant 4$）是可调节的参数，且有：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。后者反映了 Sim_1 到 Sim_4 对于总体相似度所起到的作用依次递减。由于第一独立义原描述式反映了一个概念最主要的特征，所以我们应该将其权值定义得比较大，一般应在 0.5 以上。

在实验中我们发现，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比较大，将导致整体的相似度仍然比较大的不合理现象。因此我们对公式(4)进行了修改，得到公式如下：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \tag{3.5}$$

其意义在于，主要部分的相似度值对于次要部分的相似度值起到制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。

下面我们再分别讨论每一部分的相似度。

1) 第一独立义原描述式：就是两个义原的相似度，按照公式(3.3)计算即可；

2) 其他独立义原描述式：由于其他独立义原描述式不止一个，所以计算较为复杂。我们还是按照上面的思想，把整体相似度还原为部分相似度的加权平均。困难在于，各个独立义原描述式之间没有分工，所以很难找到对应关系。我们按照如下步骤对这些独立义原描述式分组：

先把两个表达式的所有独立义原（第一个除外）任意配对，计算出所有可能的配对的义原相似度；

取相似度最大的一对，并将它们归为一组；

在剩下的独立义原的配对相似度中，取最大的一对，并归为一组，如此反复，直到所有独立义原都完成分组。

3) 关系义原描述式：关系义原描述式的配对分组较为简单，我们把关系义原相同的描述式分为一组，并计算其相似度；
4) 符号义原描述式：符号义原描述式的配对分组与关系义原描述式类似，我们把关系符号相同的描述式分为一组，并计算其相似度。
5) 在以上 2)、3)、4)的计算中，最后求加权平均时，各部分取相等的权值。

§3.3 关键技术

§3.3.1 基于 HowNet 概念获取

1、预处理

对文本进行分词处理后，需对每个已经切分的词语进行词性的标注，这样可以在预处理阶段就排除那些对文本文摘作用不大的介词、虚词、数词等词语，只对一些关键的名词，形容词等重要词语进行处理，这样可以大大提高程序运行的速度。

2、概念获取

本文概念指的是在文章中词义相关的基本语义单元。一个概念可以对应文中的一个词，也可以对应文中的多个词义相近的词。

下面是对 HowNet 处理后得到有用信息的一部分。

表 3.3 HowNet 有用信息

W_X	G_X	DEF
工作	N	affairs|事务，$undertake|担任
工作	N	fact|事情，do|做
工作	N	affairs|事务，#occupation|职位，earn|赚，alive|活着
饭碗	N	affairs|事务，#occupation|职位，earn|赚，alive|活着
职业	N	affairs|事务，#occupation|职位，earn|赚，alive|活着
差事	N	affairs|事务，#occupation|职位，earn|赚，alive|活着

例如上表 3.3 中“饭碗”、“职业”、“差事”就是多个词义相近的词，本文认为它们代表的是同一个概念。

本文采用基于 HowNet 的知识库来获取各个词语所代表的概念。

HowNet 对该词典中存在的词语都给出了一个定义，即 HowNet 中的 DEF 项，本文假定 DEF 项为该词语所表达的概念。对于文章中的单义词并且在 HowNet 词典中存在的词语来说，只需要给该词语一个相应的概念就可以了；而如何获得多义词和未登录词语所表达的概念[46, 47]？

本文采用一种最大匹配的方法来获得多义词所表达的概念。试验表明，即便是多义词，该词语某个义项的同义词出现的频率也会大大增加。因此，该词语的某个义项也会通过不同的词语不断被运用出来。基于此原因，在得到该词语后对每个词语的词义进行判断，查找与该词义相同的词语的个数，然后通过对个数的排序来决定到底应该选择何种义项作为多义语的实际义项。例如：在表 3.3 中，“工作”是一个多义词，系统将对“工作”的每个义项(即 DEF)进行查找，记录下每个义项在文本所出现的次数，然后根据每个义项含有同义词的个数来判断“工作”一词在本文中的具体义项。如果在某篇文章中“affairs|事务，$undertake|担任”义项同义词出现 1 次，“fact|事情，do|做”义项同义词出现 2 次，而“affairs|事务，#occupation|职位，earn|赚，alive|活着”义项同义词出现 5 次的话，就认为在该篇文章中“工作”一词的具体义项为“affairs|事务，#occupation|职位，earn|赚，alive|活着”。通过此方法在一定程度上消除了词语歧义。

对于未登录词的处理是，计算每个未登录词在文章中出现的次数是否大于预先设定的阈值，如果是，则标注为一种不存在于概念获取中已经发现的概念；否则，删除那些在文章中出现次数小于阈值的未登录词语。

以国家语委语料库的(分类号 ba10000101)文章“我国哲学研究的新趋势”为例进行测试，分别用基于概念的方法和基于词频统计的方法进行义项数目统计。结果表明，同词频统计方法相比，基于概念统计算法的义项数目大大减少，只考虑文章中出现频率大于一次的词语和概念，该文章中共 139 个词语和 98 个概念(并不是说 139 个词语包含在 98 个概念中)。考虑到概念中都包含多个词语的情况，我们统计了这 98 个概念共包含了 223 个互不相同的词语，说明我们使用了概念统计的方式，使得更多的词语包含于更少的概念之中，这样就能够不漏掉那些词频统计中出现次数很少的而表达的是文章一个重要概念的词语，从而能够提高文章的召回率。

3、构造概念向量空间模型

得到全文中每个词语所表达的概念后，在构造向量空间模型时，使用概念向量空间模型，而不是单独以词形为基础的向量空间模型；即 VSM 由以前的 $S_j(W_1, F_{1j}; W_2, F_{2j}; \ldots W_n, F_{nj})$ 变为现在的 $S_j(C_1, F_{1j}; C_2, F_{2j}; \ldots\ldots C_k, F_{kj})$(其中 $k \leqslant n$)，C_i 为相互独立的一个个概念，是一些同

义词的集合，F_{ij}为每个概念在S_j中出现的频度。

基于 HowNet 概念向量空间模型构造的算法可描述为:

(1)将 Text（分词处理后的词语集合）中一个词语 Wi 送入 HowNettool 中，得到 Wi 的概念集合 SMi；若 TEXT 为空，则结束；

(2)SMi 不为空时，找出 Text 中概念为 SM1 的所有词语，并记录在 A1 中，计算出词语数量 N1，将概念 SM1 从 SMi 中删除，处理过的词语从 Text 中删除；若 SMi 为空，转到（1）；

(3)SMi 不为空，从 SMi 中找出概念为 SM2 的所有词语，并记录在 A2 中，计算出词语数量 N2，将概念 SM2 从 SMi 中删除，处理过的词语从 Text 中删除；若概念集合 SMi 为空，则将 A1 送入到 Ti 中，并删除 A1 在 TEXT 中包含的词语，转到(1)；

(4)比较 N1 与 N2 大小，将大的 Ni 送入 N1 中，Ai 送入到 A1 中，SMi 送入到 SM1 中；转到(3)(其中 i=1,2)。

§3.3.2 概念重要度计算

基于概念向量空间模型建立的算法，完成对所有待分词语进行概念归类合并后，将得到一个概念集合$\{C_1, C_2\cdots, C_k\}$，其中C_k是词语的集合$\{W_1, W_2, \ldots\ldots W_h\}$，本文将各个概念在待处理的文本中直接出现的次数定义为概念出现的频度。则概念C_i的概念出现频度$F(C_i)$为:

$$F(C_i)=\sum_{n=1}^{h}F(W_n) \tag{3.6}$$

其中$F(W_n)$为词语W_n在文本中出现的频度。概念出现频度通过词义相同而词形不同的词语出现的频度反映出各个概念在文本中的重要程度。

此时，通过概念向量空间模型的方法对待处理的文本建立起的句子 S 对应的向量变为$S_j(C_1, F_{1j}; C_2, F_{2j}; \ldots\ldots C_n, F_{nj})$，其中$C_i$为句子所含有的概念，$F_{ij}$为$C_i$对应的频度。由公式 3.6 可知:

$F_{ij}=\sum_{n=1}^{h}F(W_n)$，其中$\{W_1, W_2, \ldots W_k\}$为概念$C_i$所包含的词语。

[定义 1]对已得到一个概念集合$\{C_1, C_2\ldots C_n\}$的文本，定义其概念向量空间模型如下:

$$C=(F_{ij})_{n\times m}\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1m} \\ F_{21} & F_{22} & \cdots & F_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ F_{n1} & F_{n2} & \cdots & F_{nm} \end{bmatrix}$$

其中 n 为文本中的概念个数，m 为文本中句子的个数，第 i 行第 j 列元素 F_{ij} 表示概 C_i 在句子 S_j 中出现的频度。

为了计算每个句子的重要度，系统还计算出各概念 C_i 的重要度 $W(C_i)$。其中 F_{ij} 为概念出现频度 C_i，则 $W(C_i)$ 的计算公式如下：

$$W(C_i)=\lambda\times\log\sum_{j=1}^{m}F_{ij} \tag{3.7}$$

其中 F_{ij} 为 C_i 的概念出现频度，λ 是当 C_i 为标题词的加权系数，在本系统中是以 1.1 作为计算的。

这里仍以文章“我国哲学研究的新趋势”为例进行测试，我们统计了概念方法和词语统计两种方法建立的 VSM 的义项频度分布如图 3.4 。

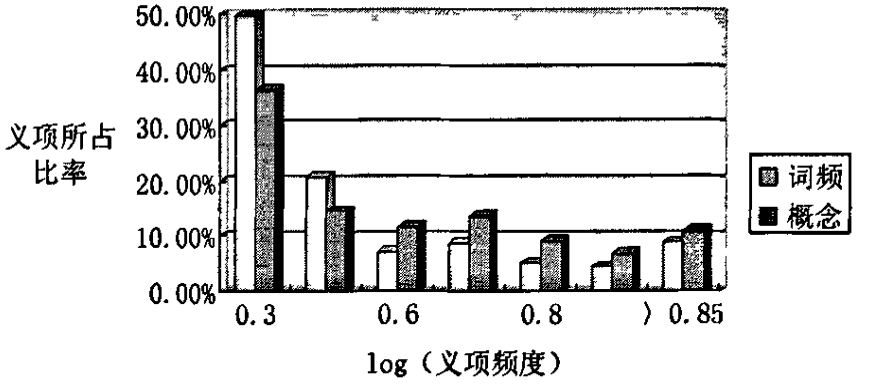

图 3.4 义项频度分布图

通过对图 3.4 的观察发现：词频统计的高频义项词只占有 31%，而经过概念归纳后，高频义项集合的占有率超过了 50%(高频义项是指 log(义项频度)>0.5)。这说明基于概念的方法对待处理文本包含的语义概念进行了有效的融合和归纳，高频概念里都包含了一定的语义相关的词语，频度得到加强，更能体现文章主题。同时，对基于词频和基于概念的方法统计后，发现基于概念的方法所获得的高频词语数量比基于

词频统计方法所得到的高频词语数量增加了一倍多，一部分低频词语通过语义关系归纳形成了一些新的高频义项，有利于表达用词形统计无法找到的文章深层主题。

§3.3.3 基于概念向量空间模型的自动文摘生成

1、句子重要度计算

句子重要度计算是基于概念对待处理的文本建立起句子的向量空间模型 $S_j(C_1, F_{1j}; C_2, F_{2j}; \ldots\ldots C_n, F_{nj})$ 来计算进行句子重要度。对大量文本试验结果分析发现，句子权重主要与句子所包含的概念、未登录词、句子本身所处段落的位置以及段落本身的重要度等因素密切相关。句子权重的计算函数为：

$$W'(S_j) = \lambda_1 \lambda_2 \frac{\sum_{i=1}^{n} F_{ij} \times W(C_i)}{MM'} \qquad (3.8)$$

其中 $W(C_i)$ 为 Ci 的重要度，F_{ij} 为 Ci 在句子 S_j 中出现的频度，M 为句子 S_j 包含的所有词语；M' 表示句子 S_j 中包含的分句数。λ_1 为当句子是段落的句首或者结尾是的加权值，本系统设为 1.5 。λ_2 为句子所处段落重要度，计算方法如下：

对段落采用自动聚类的方法，聚类算法的目标是将一组对象划分成若干组或类别，简单地说就是相似元素同组、相异元素不同组的划分过程。

这里引入 K-means 算法，此算法以每个类中的一个伪对象作为该类的质心。这个伪对象往往并非存在于该类中的实际对象而是该类包含的所有对象的均值对象。K-means 算法依据对象与类质心之间距离的最小原则将每个对象分配进与某类质心距离最近的那个类中，进而重新计算各个类的质心以便进行下一轮的对象分配，这种分配过程直到各类质心不再变化或满足某种特定的结束条件为止。

段落聚类的步骤如下（其中 K 为聚类数）：

第一步：随机选择 K 个段落特征向量作为初始 K 个类的均值向量；

第二步：通过计算剩下的段落特征向量到 K 个类的均值向量的欧拉距离来分派它们进与之距离最近的 K 个类当中的一个；

第三步：重新计算现在 K 个类的均值向量；

第四步：重复第二步到第二步，直至各个类的均值向量不再变化为止。

该聚类分析方法的基本思想是：如果聚类数 K 能够被正确地确定，那么相应的聚类结果就能够有效地区分文本中隐含的不同主题，因此相应的所有 K 个类下的每个类代表句和其对应类的语义相似度的平均值将趋向最大化。换句话说，所有 K 个类下的每个类代表句和其对应类的中心的平均距离将趋向最小化。

实现过程如下：

把文章的每个段落映射为 n 维向量空间中的一个节点 $P(C_1, W_1; C_2, W_2; \ldots C_n, W_n)$，其中 C_k（$1 \leqslant k \leqslant n$）为已经获得的文章的概念，n 为获得的文章概念总数，W_k 为概念 C_k 在段落 P 中出现的次数。通过计算两个段落 P_i 和 P_j 的语义相似度来进行聚类。

用顺序排列的 n 个节点表示段落，节点号表示段落号，计算出文章中所有段落间的段落相似度 SP_{ij}，构造任意两段之间的段落相似度矩阵 P_{mn}，

其中 SP_{ij} 表示第 i 段和第 j 段间的段落相似度。这里用运信息熵公式计算段落重要度，信息熵表示矩阵 P_{mn} 中每行在矩阵中的信息量大小。则计算公式定义为：

$PW_i = \sum_{j=1}^{n} \left[SP_{ij} \times \log(SP_{ij} + 1) \right]$，对数里面用（$SP_{ij}$ +1）代替 SP_{ij} 是为了避免出现 SP_{ij} 为 0 时，对数没有意义的情况。$\lambda_2 = PW_i$，为句子 S_j 所属段落。

所有的句子重要度都进行计算后，按句子重要度的大小排序，选择重要度高的句子作为文摘句，并且按照这些文摘句在原始文章中的顺序进行排序，就可以得到一个粗略的文本文摘。

2、减小文摘冗余度

通过上面几个步骤可以得到一个粗略的文摘，但是这个文摘中往往会出现文摘句的冗余度较大的问题，我们通过句子相似度计算减少文摘中这样的句子[48]。

[定义 2] 存在句子 S1, S2，用 SameWC(S1, S2)表示 S1 和 S2 中相同概念的个数。则句子 S1 , S2 的语义相似度为：

$$Sim(S1, S2) = 2 \times \frac{SameWC(S1, S2)}{Len(S1) + Len(S2)} \qquad (3.9)$$

其中 Len(S1)和 Len(S2)分别是句子 S1, S2 中兵有的概念个数。

系统会对每个抽取出的句子之间的相似度进行计算，如果相似度的

值大于 0.7，则认为 S1，S2 描述的是同一个主题，系统将提取重要度大的一句而删除重要度小的句子，直到抽取出足够的文摘句。降低冗余度后，可提高文摘的准确度，形成一个较好的文摘，然后通过进一步的加工形成最后的文摘。

第四章　系统实现与试验分析

§4.1 系统实现

§4.1.1 系统的主要功能

本系统的主要功能是，对一篇文章，通过文本的词语自动分词、应用停用词词表进行过虑停用词、切分后词语所表达的概念获取、计算概念和句子重要度等环节，得到该文本的文摘，并且通过术语抽取、概念重要度比较等环节，确定该文本的主要概念。

§4.1.2 系统的主要模块设计

系统由六个主要模块组成，各个模块的总体关系如图 4.1 所示。

1、自动分词

自动分词模块主要对欲进行文摘的文章进行词语切分预处理。众所周知，与英文文本相比中文文本没有特殊的词语之间或者词组间的明确分隔符，因此在进行文本文摘前的第一步就是要对欲进行文摘的文本进行一个词语切分，本系统采用一个分词软件对文本进行词语切分。

2、过滤停用词

能表达一篇文章的词语是那些有实际意义的词语，即那些能表达实际含义的实词而不是虚词。本系统应用一个停用词词表将这些停用词剔出掉。这样，一方面可以提高程序的运行速度，另一方面还可以在对文本进行关键词或者关键概念的提取中，有效的提高抽取的准确率。

3、提取实词，获取概念

本模块引入 HowNet，对分词后的词语进行处理，得到各个词语词性和各词性的概念，应用 HowNet 中包含的词语概念的信息提取各实词的概念。

4、计算概念重要度

获得文本所有实词概念后，建立词语的概念向量空间模型，对各个实词进行概念归类，得到概念集合，计算概念重要度。

5、计算句子重要度

通过建立概念空间向量模型，来计算出各个概念在文本中的重要度，并进一步计算出句子在文本中的重要度。对句子重要度进行排序，

提取出重要度的权重大的那些句子，作为一个粗略的文摘句集。

6、计算句子相似度

对得到的粗略文摘集进行处理。为避免各个文摘句存在语义重复的现象，提出通过计算每个文摘句的相似度的方法，在相似度大的文摘句中，取其重要度大的语句留为本文章的文摘句，并舍去其他文摘句，可以提高文摘的召回率。

7、文摘输出

生成一篇文章的文摘并输出。

下面给出了系统模块关系图（图 4.1），详细介绍了此系统各个模块之间的关系及整合顺序。为了更好的说明此系统各个模块具体功能的实现，在系统模块关系图之后，给出了一个具体实例，并大体叙述了实现步骤，更深入解说此系统如何实现，最后给出文摘。

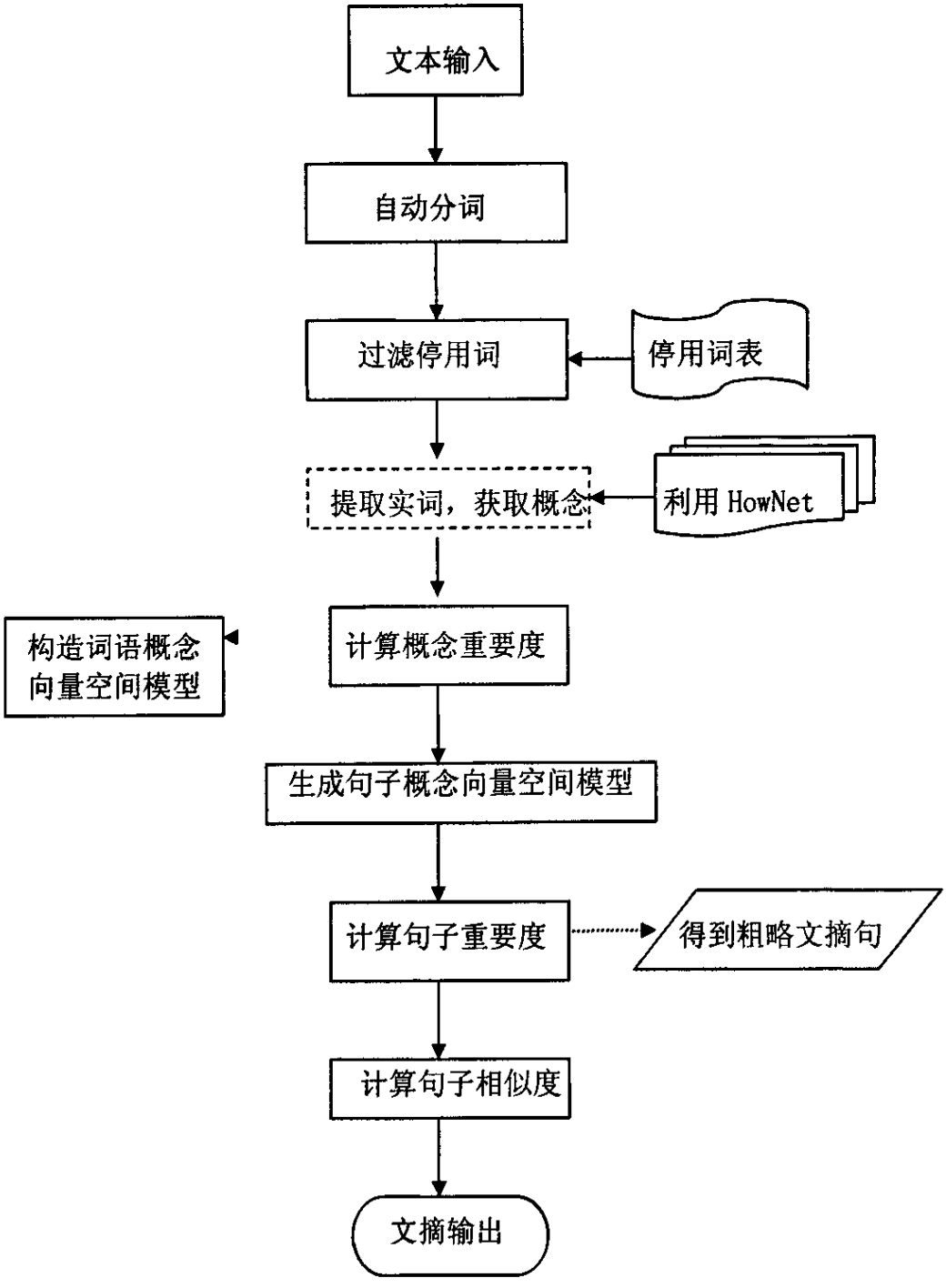
文本输入
自动分词
过滤停用词
停用词表
提取实词，获取概念
利用 HowNet
构造词语概念向量空间模型
计算概念重要度
生成句子概念向量空间模型
计算句子重要度
得到粗略文摘句
计算句子相似度
文摘输出

图 4.1 系统模块

应用该系统对国家语委语料库中的“冬天吃什么水果好”文章为例生成的文摘结果如下所示。

表 4.1 实例

冬天吃什么水果好

冬季气候干燥，常常使人感到鼻、咽干燥不适，这时如果能吃些生津止渴，润喉去燥的水果，会使人顿觉清爽舒适。

*冬季有保健医疗性质的水果，要数梨和甘蔗了。*中医认为，梨有生津止渴、止咳化痰、清热降火、养血生肌、润肺去燥等功能，最适宜于冬季发热和内热的病人食用。尤其对肺热咳嗽、小儿风热、咽干喉疼、大便燥结症较为适宜。

*梨还有降低血压、清热镇定的作用。*梨含有丰富的糖分和维生素，有保肝和帮助消化的作用。但是，因为梨性寒冷，若孩子脾胃虚寒、消化不良，则不宜多吃。

*甘蔗有滋补清热的作用，含有丰富的营养成分。*作为清凉的补剂，对于低血糖、大便干结、小便不利、反胃呕吐、虚热咳嗽和高热烦渴等病症有一定的疗效。劳累过度或饥饿头晕的人，只要吃上两节甘蔗就会使精神重新振作起来。但甘蔗性寒，脾胃虚寒和胃腹疼痛的人不宜食用。

*此外，适于冬季吃的水果还有苹果、橘子、香蕉、山楂等。*苹果可生津止渴，和脾止泻；橘子可理气开胃、消食化痰；香蕉清热润肠、降压防痔；山楂可扩张血管、降低血脂、增强和调解心肌功能，有防治冠状动脉硬化的作用。

具体实现步骤如下：

（1）对上述的文章进行词语切分处理，我们利用上述的自动分词技术，把文章分为单个词，再利用停用词表，把对文章内容没有太大影响的虚词过滤。

（2）应用 HowNet 得到各个实词的概念后，就可以构造基于概念的向量空间模型。通过实词概念相似度公式，计算出各实词之间相似程度，对各个实词进行概念归类，得到概念集合{C_1，C_2…，C_k}，其中 C_k 是词语的集合{ W_1, W_2. ..... W_n}。

（3）计算出词语 W_n 在文本中出现的频度 $F(W_n)$，通过公式（3.6）

计算出概念 C_i（其中 i=1, 2, …, k）出现的频度 $F(C_i)$。概念出现频度通过词义相同而词形不同的词语出现的频度反映出各个概念在文本中的重要程度。

（4）通过概念向量空间模型的方法对待处理的文本建立起的句子 S 对应的向量变为 $S_j(C_1, F_{1j}; C_2, F_{2j}; \ldots\ldots C_n, F_{nj})$，其中 C_i 为句子所含有的概念，F_{ij} 为 C_i 对应的频度。即 $F(C_i)= F_{ij}$。如上 3.3.2 节中定义 1 所述，得到概念集合的概念向量空间模型。再算出各概念 C_i 的重要度 $W(C_i)$（利用公式 3.7）。

（5）句子重要度计算是基于概念对待处理的文本建立起句子的向量空间模型 $S_j(C_1, F_{1j}; C_2, F_{2j}; \ldots\ldots C_n, F_{nj})$。要计算句子权重 $W'(S_j)$（利用公式 3.8）先计算句子在段落中的重要度 λ_2。通过计算两个段落 P_i 和 P_j 的语义相似度来进行聚类。其中段落 P_i 和 P_j 的段落相似度定义为公式（2.3）。计算出各个段落的相似度后再计算各个段落的重要程度。最后计算出各个句子的重要度，可以得到一个粗略的文本文摘。

（6）通过计算各个文摘句的相似度（利用公式 3.9），减少文摘冗余，得到最终的文摘。

其上斜黑体为系统抽取出的文本摘要，抽取文摘比率为占文本长度的 25%。

§4.2 试验分析

从广义的角度可将自动文摘的评价方法大致分为两类：一种称作内部评价（Intrinsic）方法，与系统的目的相关，它通过直接分析摘要的质量来评价文摘系统。第二种称作外部评价(Extrinsic)方法，它是一种间接的评价方法，与系统的功能相应，将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。文摘的评价较索引的评价要复杂的多，它需要更多的知识(语言学、领域知识和上下文关系)做支持。

§4.2.1 内部评测

内部评价方法按信息的覆盖面和正确率来评价文摘的质量，一般采用与“理想摘要”相比较的方法。这种评价方法源于信息抽取技术。信息抽取是将原文的关键要点抽取出来，并与人工抽取的内容在召回率

(recall)和准确率(precision)。内部评测就是测试文摘本身是否与文章要点一致，以及是否包含文章的基本要点。

这其中“理想摘要”的获得是个问题。通常可直接用原文作者摘要作为“理想摘要”，但也有人认为原文作者对文章所作的摘要具有一些不足之处，如：作者撰写的摘要可能不够规范；编写摘要时可能不够客观等缺点。所以很多人主张采用专家对文章直接抽取句子进行摘要。不同的专家所撰写的摘要是不同的，有时在内容上很少交叠。特别是评论或解释性文摘，内容更难达到一致。许多情况下，只能判断文摘的合理性。

这里采用召回率、准确率以及 F_measure 三个参数对文摘系统进行评估。其中召回率指系统正确识别的比率，准确率指系统准确识别的比率。

具体表现为：

$$召回率(Recall)=\frac{同时被文摘系统和专家文摘抽取的句子数目}{专家文摘抽取的句子数目}$$

$$准确率(Precision)=\frac{同时被文摘系统和专家文摘抽取的句子数目}{文摘系统抽取的句子数目}$$

为了综合衡量文摘质量，采用一种综合评估标准，即 F_measure 测试值。

$$F\_measure=\frac{2\times P\times R}{P+R} \qquad (4.1)$$

例如：某篇文章在文摘长度占文章比例 15%时，系统抽取出文摘句子数为 9 句，该文章的专家文摘抽取的句子数量为 13 句，同时存在于文摘系统和专家文摘句中的句子数量为 6 句，则系统在该文章的文摘长度为 15%时：

召回率=6/13=0.462

准确率=6/9=0.667

F_measure =(0.462*0.667*2)/(0.462+0.667) =0.546

测试试验：

采用国家语委语料库在对已经做了专家文摘的 50 篇文本(包含报刊、经济、新闻报道、文学几个方面)进行自动文摘后，每类文本的平均 F_measure 评测参数情况如表 4.2 所示。

表 4.2 F_measure 测评结果

文摘体裁 \ 文摘长度	5%	10%	15%	20%	25%	30%
经济	0.373	0.485	0.556	0.597	0.653	0.640
新闻报道	0.352	0.478	0.552	0.620	0.640	0.642
科技	0.319	0.410	0.495	0.498	0.486	0.467
文学	0.272	0.379	0.406	0.418	0.393	0.363

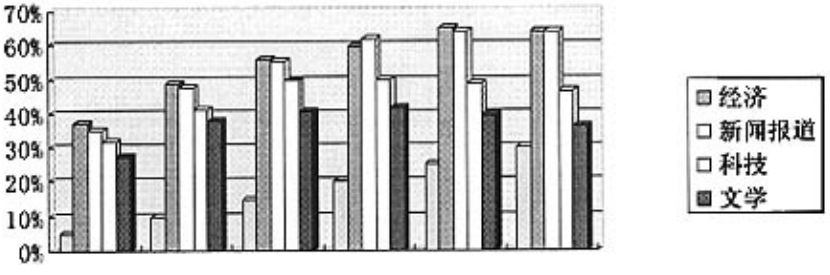

图 4.2 F_measure 测评图

在对 4 类文摘的 F_measure 值评估时发现整体情况还比较理想。对经济和新闻报道类的文章采用该方法进行文摘得到了比较好的效果，对科技文摘效果一般，而对文学类的文摘还需要进行进一步的修改。出现这种情况是因为对于经济和新闻报道类的文章描述重点比较突出，文章结构也有一定规律可寻；而文学类的文章结构相对复杂，文章重点反应也不是特别突出，所以该类文章的文摘结构也相对较差。

通过对比测试数据发现，在文摘长度为 15%之前的两种不同方法的平均召回率和准确率基本相同；但是在文摘长度在 15%后，基于词频方法的文摘结果就不如基于概念方法的文摘了。可以看到，在文摘长度大于 15%后，无论是平均召回率还是准确率，基于词频的方法都比基于概念方法要低，并且随着文摘长度的增加到 30%时差距还在加大(在文摘长度为 20%时，基于概念的方法比基于词频方法召回率和准确率分别要低 1.5%和 2%，而在文摘长度为 30%时，这一差距分别变为 3.6%和 3%)。

§4.2.2 外部评测

外部评测就是通过文摘与文章的相似度计算或者文摘在信息检索中所起作用的大小来评估文摘好坏的方法。

本文通过计算两种不同文摘方法对文本分类准确率的影响来评测文摘结果。

我们从语料库中另外选择30篇文章(包含报刊、经济、新闻报道、文学几个方面)分别用基于词频的方法和基于概念的方法进行文摘，运用一个事先已经做好的分类器进行分类。分类结果如表4.3所示。

表4.3 分类结果表

分类语料	原文本	基于词频文摘	基于概念文摘
准确率	60.4%	66.9%	68.7%

通过比较发现，文章所使用的两种文本摘要方法所得到的文本分类效果比单纯使用文本效果要好，主要是由于文摘提取出了文本中的主要内容，过虑掉了很多噪音，所以用文摘来进行分类比仅仅用原文来进行分类效果要好；在两种不同的文摘方法中，基于概念的文摘方法比单纯基于词频的文摘方法效果要好，说明基于概念的文摘更能够反应出文章的主旨。

§4.3 中文自动文摘研究存在的问题及提出的相关技术

§4.3.1 存在的问题

目前，中文自动文摘的研究主要还集中在中文单文本自动文摘上面，关于中文多文本自动文摘的研究还刚刚起步，仅有很少量的研究成果发表。

单文本自动文摘根据摘要对象的范围不同又可分为通用型和专用型两种。由于各种专用型的自动文摘技术所产生的摘要，其应用领域往往严格受限，系统的实现过于依赖深层自然语言处理技术和知识工程技术，因而并不能很好地适应目前国际上自然语言处理的主流发展趋势，即对大规模真实的非受限文本的高效处理的需求。因而，国内学术界、工业界对自动文摘的研究主要还是集中在对通用型单文本自动文摘的

研究上面。在研究的过程中，我们发现目前国内通用型单文本自动文摘存在的主要问题是：

1、过于依赖并沿袭早期自动文摘研究中的一元化处理模式，没有充分考虑到不同题材的文本其潜在的主题分布对摘要方式以及摘要结果的影响。

2、在传统的统计学句子打分方法中，由于它未做文本的主题结构分析，因此往往导致抽取出来的文摘句至多只能覆盖文本中最重要的那些主题而忽视掉其他次重要的主题。

究其原因，这主要是由于反映文本中最重要主题的句子在现有的打分方法上往往得分偏高，因而导致产生的文摘主题覆盖不全且冗余偏大。

国内已有部分学者注意到上述问题，并着手开展了针对性的研究工作，如南京大学的王继成等研究人员提出的基于篇章结构指导的中文Web文档自动摘要便是此类工作的代表。然而，正如我们在2.2节中所提到的那样，他们仅仅计算了文本中相邻段落之间的语义相似性，而忽视了对那些可能会跨段落分布的主题的处理，尤其是当针对一些非说明文、议论文体裁的文本做自动摘要时，采用这种方法所提出来的处理方式其摘要效果是不理想的。而且必须由人工来主观设定段落之间的语义相似度的阈值也是该方法的一大软肋。因为阈值的设定往往和不同的因素息息相关，如文本的题裁、体裁等，所以仅仅通过人工来强制设定其相似度的阈值往往并不合适。

§4.3.2 提出的相关技术

针对上述问题，我们尝试性地提出了以下四种关键技术，以便能够在一定程度上利用这些关键技术去致力于这些问题的解决。

1、基于无监督特征抽取的文本各级语言单元的特征向量表达

该技术的提出是为了在一定程度上解决目前文本各级语言单元在采用向量空间模型表达时所反映出来的维数过高、特征之间语义重复等问题，以便能有效地提高后续聚类算法的效率。

2、基于自适应段落聚类的文本潜在主题的自动发现

该技术的提出是为了在一定程度上克服传统的中文自动文摘方法所面临的难以适应对不同题材的多样主题文本的有效摘要问题，从而使得产生的摘要能平衡于主题覆盖度和冗余之间。同时，该技术还尝试解决了现有的主流聚类方法(如K-means)所面临的关键参数K需要人为主观设定的问题。

3、基于主题语义相似度计算的文本主题代表句的自动选取

该技术的提出是为了在一定程度上抽取出各个主题下最具有主题代表性的句子，而非传统方法中的全文权值最大的句子。

4、基于表达熵的文摘冗余的量化评价

该技术的提出是为了在一定程度上克服大多数基于内部（Intrinsic）式的评价策略在评价文摘冗余的过程中由大量人工干预所导致的结果不一致性问题。

第五章　总　结

在信息时代，人们迫切希望能借助一些有效的工具或手段来方便快捷地找到满足自己特定需求的信息。而自动文摘技术作为解决当前信息过载问题的一种辅助手段，必将发挥着越来越突出的作用。

本文主要描述了我们提出的基于概念向量空间模型的方法，针对当前自动文摘技术研究中存在的若干问题，我们也给出了相关技术加以研究并解决一些问题。本文回顾了当前国内外对自动文摘技术研究的几种方法及研究成果，介绍了几种自动文摘的研究技术。当然，本文是对基于概念统计的自动文摘获取系统的研究，与基于词频统计的自动文摘系统做了比较，其基于概念统计的方法得到的文摘很大程度上要优于基于词频统计的方法。

综上全文，本文的研究成果主要包括以下几个方面：

1)提出知网，并且利用知网获得了某个词语所表达的概念；

2)建立词语基于概念的向量空间模型，设计和实现了一个中文文本自动摘要系统；

3)采用一系列内部和外部的评测手段对该文摘系统进行了评测试验。

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)本系统采用知网以外的词典工具获得某个词的概念，如采用“词林”等词典，比较不同词典得到的概念，哪个更能满足系统需要；

2)对于聚类算法的研究，在 K-means 算法中，对于 k 的确定问题始终是其存在的一个突出却没能得到有效解决的问题。如何不通过人工来主观地提供 K 值，转而由机器自动地根据不同的应用环境自适应的定 K 将是一个值得我们去深入探究的问题。

3)探讨特征项在文本中出现的位置对文本自动摘要的影响，给标题、段首和段尾的特征项一个合适的权重，提高文摘的准确率和召回率；

4)按照文本在文档中的位置不同，一般分为标题、摘要、关键词、正文、结论和超链接等 6 个位置，可分别赋予不同的加权系数。

致　　谢:

两年半的硕士研究生生活就要结束了，首先我要感谢两年多教育我的老师，给予我帮助的同学。

衷心地感谢我的导师崔广才教授。崔老师平易近人的高尚人品、一丝不苟的治学态度、敏捷的思维、力求完美的工作作风以及孜孜不倦的工作精神将是我终身学习的楷模。在崔老师的悉心教导和关怀下，我不仅学会了做学问的道，更懂得了做人的理。能在崔老师这样优秀的导师门下学习和工作将是我一生中最值得庆幸和骄傲的事。

我也要感谢在研究生学习中教导过我的老师，他们平时的精心指导，在我完成毕业设计的过程中，能够运用自己的知识，发挥自己的才能，积累了一定的经验，也使能力进一步提高。

最后，祝愿我们的校园建设更美丽，教学质量更上一层楼，培养出更多的杰出人才，投身到我国的现在化建设中。

参 考 文 献:

1 刘挺，吴岩，王开铸。自动文摘综述。情报科学，1998(1)：P63～69

2 Luhn H P. The Automatic Creation of Literature Abstract. IBM Journal of Research and Development, 1958, 2(2):P159-165

3 王永成，苏海菊，中文科技文献文摘的自动编写。情报学报，1989(6)

4 Udo Hahn. Automatic Text Summarization:Methods, Systems, Evaluation. 2001, http://www.coling.uni-freiburg.de/-hahn

5 Inderjeet MANI. Summarization Evaluation: An Overview. In Proceedings of the NTCIR Wordshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo: National Institute of Informatics, 2001

6 H.P. Edmundson. New Methods in Automatic Abstracting. Journal of the Association for Computing Machinery, 1969. 16(2):264-285

7 王永成，许慧敏。OA 中文文献自动摘要系统。情报学报，1997. 16(2)

8 刘廷，王开铸。自动文摘的四种主要方法。情报学报，1999. 18(1)

9 刘开瑛，郭炳炎。自然语言处理，科学出版社，1991

10 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

11 姚天顺等。自然语言理解——一种让机器懂得人类语言的研究。清华大学出版社，广西科学技术出版社，1995

12 L.F. Rau, P.S. Jacobs, Uri Zernik. Information Extracting and Text Summarization Using Lingusitic Knowledge Acquisition. Information Processing and Management, 1998. 25(4):419-428

13 王建波。面向议论文理解的自动文摘系统研究与探讨。哈尔滨工业大学博士学位论文，1992

14 刘挺，吴岩，王开铸。基于信息抽取和文本生成的自动文摘系统设计。情报学报，1997. 16(增刊):24-24

15 杨晓兰，钟义信。基于文本理解的自动文摘系统的研究与实现。电子学报，1998. 26(7)

16 史磊。中英文自动文摘系统及其若干相关技术研究。上海交通大学博士学位论文，2000

17 G. Salton, J. Allan, C. Buckkeym, Amit Singhal. Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts. Science, 1994. 264(3):1421-1426

18 王继成，武港山等。一种篇章结构指导的中文 Web 文档自动摘要方法。计算机研究与发展，2003. 40(3):398-405

19 Kupiec, Julian, Jan O. Pedersen, Francine Chen. A trainable document summarizer. Research and Development in Information Retrieval, 1995:68-73

20 Dragomir R. Radev, Eduard Hovy, Katheen McKeown. Introduction to the Special Issue on Summarization, 2002. 28(4):399-408

21 Lin, C, E. Hovy. Identifying topics by position. In Fifth Conference on Applied Natural Language Processing, Association for Computational Linguisitics, 31March-3April, 1997:283-290

22 Jaime Carbonell, Jade Goldstein. The Use of MMR, Diversity-Based Rcranking for Rcordering Documents and Producing Summaries. In Proceedings of ACM SIGIR'98, 1998:335-336

23 刘挺，王开铸。基于篇章多级依存结构的自动文摘研究。计算机研究与发展，1999. 36(4)

24 Yihong Gong, Xin Liu. Generic Text Summarization Using Relevance Measure and Latent Senmantic Analysis. In Proceedings of ACM SIGIR'01, 2001:19-25

25 Conroy, John. Dianne O' lxary. Text Summarization Via Hidden Markov Models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001:406-407
26 沈洲，王永成，韩客松。一种基于主题敏感辞分析的新闻文献自动摘要系统的研究与实践。高技术通讯，2001(9)
27 宋今，赵东岩。基于语料库与层次词典的自动文摘研究。软件学报，2000.11(3)
28 孙春葵，钟义信。关于自动文摘系统中文摘句式的一种机器学习方法。计算机工程与应用，2000(5):18-23
29 沈玮杰。基于文献结构的自动文摘初探。现代图书情报技术，2002(3):23-34
30 Simone Teufel, Marc Moens. Summarising Scientific Articles-Experiments with Relevance and Rhetorical Status. Computational Linguistics, 2002.28(4):409-445
31 Knight, Kevin, Daniel Marcu. Statistics-based Summarization-Step One:Sentence Compression. In Proceedings of the 17th National Conference of the American Association for Artificial Intelligence, 2000 703-710
32 Tadashi Nomoto, Yuji Matsumoto. A New Approach to Unsupervised Text Summarization. In Proceedings of ACM SIGIR'01, 2001:26-34
33 李蕾，钟义信，郭祥昊。面向特定领域的理解型中文自动文摘系统，计算机研究与发展，2000.37(4)
34 Salton G,MeGill M J. Introduction to modern Information Retrieval [M].New York:McGraw-Hill Book Company,1983, P400.
35 Miller, G.A., Charles, W. Contextual Correlates of Semantic Similarity Language and Cognitive Processes, 1991,6(1):P1-28.
36 http://www.keenage.com/zhiwang/c_zhiwang.html
37 现代汉语通用字典，中国人民大学语言文字研究所，外语教学与研究出版社，1987.
38 现代汉语词典(修订本)，中国社科院语言研究所词典编辑室，商务印书馆，1996.
39 汉英词典(修订本)，北京外国语大学英语系词典组，外语教学与研究出版社，1995.
40 WordNet 1.6，普林斯敦大学，1999.
41 SenseWeb，原新加坡系统科学研究院，1996.
42 牛津一杜登英汉图解词典，卜纯英译，轻工业出版社，1988.
43 LONGMAN English-Chinese Dictionary Of Contemporary English, Longman Group UK Limited, 1988.
44 现代汉语语法信息词典详解，俞士汶等，清华大学出版社，1998.
45 孙春葵。自动文摘及其知识获取技术研究。北京邮电大学博士学位论 2000
46 陈浩，何婷婷，姬东鸿。基于 HowNet 的无导词义消歧[C]。第五届汉语词汇语义研讨会，P326-332
47 Mitra, P.,Murthy, A. C.,&Pal, K. S.: Unsupervised Feature Selection Using Feature Similarity [J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. 24:4, P301-312.
48 吕学强，任飞亮，黄志丹，姚天顺.句子相似模型和最相似句子查找算法[J]。东北大学学报。2003.6(24), P531-534.