中 文 信 息 学 报
第 11 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING vol. 11 No. 3

汉语文本压缩研究及其应用①

王忠效
(中国科学院软件研究所 中国科学技术大学管理学院)

【摘要】: 汉语文本压缩至今很少受到重视, 然而, 作为许多计算机应用系统的支撑技术, 其重要性毋庸置疑。本文结合汉语文本的特征对现行文本压缩技术进行评述, 指出汉语文本理论上可能获得的平均压缩比率 (> 3. 9) 及现行压缩算法所能达到的水平 (1. 6 左右)。此外, 讨论了汉语文本压缩的研究方向以及几种典型的应用。

关键词: 汉语文本压缩, 算术编码, Huffman 编码, Lempel - Ziv 算法, 熵

近年来, 多媒体与网络通信技术热使得图像、语音压缩空前活跃。相比之下, 人们给予文本压缩的重视程度远为不够。虽然文本压缩的对象与性质决定了它不可能获得如同前者那样极其可观的压缩倍率, 但是, 如本文所示, 作为通信、全文检索、电子出版物等领域的重要支撑技术, 文本压缩、尤其是汉语文本压缩同样具有广阔的应用前景。

一、文本压缩技术简介

文本压缩 (text compression) 是数据压缩 (data compression) 的一个分支, 属于无损压缩 (lossless compression) [注]。它的目标是通过对数据施加某种操作或变换使之长度变短的同时, 还必须保证原始数据能够从压缩产生的压缩码中得以精确的还原。对原始数据进行压缩的过程又称编码, 其逆过程, 即从压缩码还原原始数据的过程, 称解压缩或译码。这里, 文本主要包括用各种自然语言书写的图书文献、档案资料、电传书信, 用各种计算机语言书写的源程序, 用二进制存储的计算机可执行程序, 以及各种统计数据、文件报表等等。

反映一个压缩算法优劣的很重要的尺度是数据的压缩比。本文将压缩比定义为原始数据长度与原始数据经压缩产生的压缩码的长度的比值。我们总是很自然地希望得到尽可能大的压缩比。然而, 往往只有很复杂的算法和数据结构才能获得最理想的压缩效果, 而复杂的算法必定伴随较差的时间性能, 空间开销也可能足以使之失去实际应用价值。因此, 在实际评价一项压缩技术时, 人们总是要在压缩比、压缩/解压缩速度和空间开销三方面寻求平衡。

① 本文于 1996 年 10 月 7 日收到

说　明：在讨论汉语文本压缩比时，本文均假定汉字的计算机内码遵循 GB2312 - 80 标准按双字节 16 位编码。

二、国际、国内研究现状述评

文本压缩算法可以划分为统计方法和词典编码方法。

统计方法当以 Huffman 编码（Huffman coding）和算术编码（arithmetic coding）为代表。这种方法需要统计信源符号的概率分布情况，并根据统计结果产生压缩码。统计可以一次性完成（如静态 Huffman 编码），也可以边编码边统计（如动态 Huffman 编码）。值得指出的是，Huffman 在 50 年代初提出的以具有前缀性质的变长码为特征的 Huffman 编码是数据压缩领域的一个重要里程碑，曾经统治数据压缩领域达 30 年之久，至今研究它的文章仍旧源源不断。这种编码技术长期以来一直被许多工程技术人员视为最佳的编码方法。其实，这是一种错误。具体地讲，Huffman 编码有两个直到今天才迟迟被发现的不恰当的假设，即信源符号彼此孤立无关和每个信源符号都有 1 - 1 对应的压缩代码。实际上，以自然语言文本为例，一个信源符号（字或词）的出现总要受到句法、语义、语用乃至篇章的制约，因此，信源符号之间并非彼此孤立无关。另一方面，现代编码技术却能保证信源符号未必要有 1 - 1 对应的压缩代码，譬如，设已经为符号串 abcdefgh 编码，则信源符号串 abc 和 bcdefgh 等的压缩码可以等长。此外，由于计算机所能表示的最小信息量为 1 比特，huffman 编码还面临这样一个棘手的问题：它不能经济、精确地编码依据符号概率分布计算出来的具有浮点熵值的符号，而是依符号在 Huffman 树上的位置用整数位二进制数编码，从而产生比理论上可能的长度更长的代码。从另一个角度看，70 年代后期提出、80 年代末得以迅速流行的算术编码，无论从编码效率上还是从时间性能上，大有取代并淘汰 Huffman 编码的趋势。算术编码绕过了用一个代码一一对应一个信源符号的做法，它用一个单独的浮点输出数值给整个一条消息编码，从而每个符号对于输出代码的净效应可以是一个小数位数。尽管如此，它与 Huffman 编码本质上都假定信源中符号独立无关，即信源系零阶 Markov 信源。然而，事实上，信源所呈现的冗余，不只在个体符号的统计特征上，还表现在上下文结构以及符号之间的相关性方面。因此，任何片面强调一方面冗余而忽视其它方面冗余的算法，从理论上讲，都不可能成为所谓的最佳编码。其次，当遇到信源符号趋于等概率时，这两种方法的编码效率降低到极点。对于汉语大字符集，由于实际使用中汉字具有良好的收敛性，因此，除了高频使用的前 n 个（譬如 n〈500）汉字之外，其他大量汉字的出现概率可近似地视为等同，因此，这两种方法不适合于给前 n 个高频字以外的大量汉字编码，从而不能简单地用于汉语文本压缩。

词典编码方法则是基于数据中许多结构频繁重复再现这一事实，人们可以对相同符号串分配同一码字、通过索引、或者其他诸如此类的方法编码。不难看出，许许多多标以符号串

注：与无损压缩相对应的是有损压缩（lossy compression），它不保证被压缩的数据得到精确的还原，即还原后的对象较之压缩前存在一定的误差。这类压缩技术适用于语音、图形图像等数字化模拟信号，只需保证还原后的结果能够满足我们的视听要求。

匹配的编码技术通常都属于词典编码方法。词典可以一次性固定在压缩/解压缩程序中，这是静态方法。它对于特定的信源可能有良好的表现，但对于其他信源则很难说会有多大的价值。词典可以在对信源完成统计之后作为压缩码码流的一部分插入到压缩码码流的固定位置，这属于半自适应方法，其主要局限性在于需要事先对信源进行统计分析，因而在通信系统、以及其他对压缩过程的时间性能提出严格要求的实时系统中不宜采用。当然，词典还可以通过边编/译码边累积的方式形成，其中“词条”在这一过程中被分散插入到压缩码码流中了；或者更理想的做法是使得后来出现的新“词条”隐身在已经编/译码的信源符号构成的历史数据中，因此，当压缩过程启动之后不久，压缩码码流中不再留出空间显式地记录“词条”。这是自适应方法，其优点在于并不面向特殊的信源，在编/译码的过程中逐渐形成词典，从而可望使得信源、信宿两端同时进行的压缩与解压缩过程达到同步。这三类词典编码方法中，目前占主导地位的是自适应方法。

70 年代末提出、80 年代中期走向实用化的 LZ 压缩技术开创了现代词典编码方法，并且已经牢固地统治着通用压缩世界。LZ 的基本思想是：数据中的子串可以通过用指代先前已处理数据（即历史数据）中相同子串的方式来描述。对历史数据的存储方式可以不同，LZ77 采用滑动的缓冲区（或称窗口）记录，LZ78 则选择词典方式进行登录。应该指出的是，两者的压缩效率没有显著差异，而且都是当重复的子串越长，压缩效率越高。

国内已经有人结合汉语文本压缩对它们分别进行了实验研究。总地讲，无论如何修改，两类编码方法压缩汉语普通文本的压缩比在 1．3 到 2．0 之间，平均压缩比在 1．6 左右。在我们研究 LZ77 算法的文章中，已经初步分析了为什么 LZ 算法压缩汉语文本的压缩比远不及压缩英语文本（压缩比在 2．0 以上，平均压缩比可望接近或达到 2．5）。实际上，词频统计显示：汉语文本中单字节与双字节词占绝对多数。这表明汉语文本中重复子串平均而言不可能长，从而 LZ 类现代词典编码方法也不是很适合汉语文本。

我们已经注意到，近年来有人尝试将上述统计方法与词典编码方法有机地融合起来。按照一般的看法，如果单纯一种算法能够获得高于 1．5 的压缩比，则这一算法可称得上优秀，LZ 就是这样的算法。同时，鉴于汉语大字符集在动态使用过程中呈现良好的收敛性，人们自然想到统计方法可以用来给高频汉字编码。因此，将两类方法融为一体是研究汉语文本压缩的一种可行的方向。作者正在进行这方面的实验和研究。

值得指出的是，通行的文本压缩算法都是所谓普遍适用的（universal）。为什么它们压缩汉语文本的效率远不及压缩英语或其他印欧语言文本呢？我们认为，这些算法的提出首先是基于某一印欧语言文本的（如 LZ 算法是对六七十年代英语文本词典压缩研究的总结）；或者是基于一种抽象的信息流，而印欧语言文本碰巧较多地再现了这种信息流中被算法提取、利用的有关结构的以及统计的特征（譬如，高频再现冠词、介词、前/后缀及其前后的空格字符，等等），汉语文本却较少地呈现这些特征。

通过上述讨论，我们了解了现行两类文本压缩方法应用到汉语上存在的各种严重困难。但是，我们也看到：信息论科学、尤其是数据压缩领域半个世纪以来所取得的成就，为我们的研究工作奠定了理论基础。譬如，数据压缩领域 80 年代取得的成就之一，是将压缩过程分离为编码器（encoder）与模型（model）两个组成部分（图 1），这已经成为现代数据压缩

的典型型式。现行的理论与实践表明：按照信源中数据的概率分布，模型生成预期（或称期望）并将之提供给编码器；编码器根据预期与实际接收到的信源符号之间的差异产生信源符号的压缩码。这种方法最大的优点，是可以把各种专门的压缩方法纳入少数几类模型中，而与独立的模型相应的又只有少数几种类型的编码器。因此，这种编码器与模型相分离的方法为不同类型的信源数据选择不同的压缩技术提供了基础。这里，压缩算法的关键是构造恰当的模型。

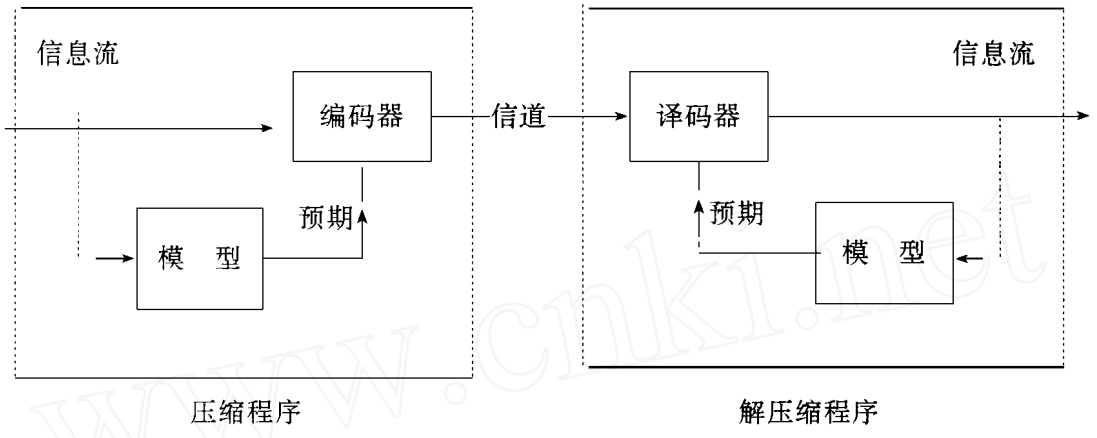

图 1 现代数据压缩的典型型式

我们认为，进行汉语文本压缩研究，有必要对模型这一概念作出拓展：模型在形成预期时，不仅仅只根据信源符号的概率分布情形，还将根据有利于形成准确预期的有关信源中符号的构造规律及信源符号之间可能存在的相互制约关系。这样一来，毫无疑问，我们的模型将有可能吸收有关现代汉语短语结构及字词频统计的研究成果，特别是，能够从认知心理学、计算语言学以及人工智能的角度探讨自然语言的文本压缩。实际上，这种研究方向代表了自然语言文本压缩研究的发展趋势，因为自然语言文本中的冗余毕竟不只单纯表现在言语外观表面的数理统计特征上。

现在，我们来看一看国内的研究状况。我们不得不承认，我们在中文（包括汉语）文本压缩方面，尚处于起步阶段。实际上，直到现在，人们仍未将中文文本压缩列为中文信息处理的一个研究领域。在其他使用汉语的国家和地区，我们也还没有发现有关开展汉语文本压缩研究的报道。然而，一个领域未被重视决不意味着它不重要。拿汉语文本压缩来讲，单从本文第四部分的数据，我们不难看出这是一个十分紧迫的课题。从另一方面看，尽管我们在正面开展的工作很少，我们在中文信息处理的其他领域进行了许许多多的工作，如字词频统计、汉字编码、汉字熵值估算、汉语词库分级组织以及汉语分析与理解，等等，这些方面的成果都将给汉语文本压缩研究提供直接的支持或有益的借鉴。

三、汉语文本的可压缩性

汉语文本的可压缩性究竟如何？

国内至今已完成了多项工程浩大的汉语字、词频计算机统计工作，取得了可喜的成就。从这些统计结果，人们惊喜地发现汉字的使用比之英语单词要拥有好得多的收敛性，譬如3072个汉字的累计出现频率高达99．7％（尽管单是《汉语大字典》所收集的汉字就超过56，000个），其中116个汉字的累计出现频率高达50．1％，前15个高频字在现代汉语文献中的累计出现频率更是高达19．8％。此外，汉语词汇也具有极好的收敛性。譬如，北京语言学院的统计结果显示：前15个高频词的累计出现频率为21．3％，前190个高频词覆盖了文献的50．0％，前5，000个高频词的覆盖率达到91．7％。

根据类似的统计资料，有人计算出汉字的熵为9．65比特。其计算依据是信息论中关于平均信息量（即熵）的定义。

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i$$

在这里，P_i是某汉字在汉语中出现的概率，可用它在被抽样统计的汉语语料中的频率来近似；n为抽样中不同汉字的数量。显然，抽样的文本量的大小及文本类型、风格等的差异将影响到H的值。实际上，在这个基础上推算出的汉字熵值只可能是一种估计。这里显然存在类似Huffman编码的假定，忽视了一个汉字在文句中既受到上下文环境的制约、又作为上下文环境的成员对别的汉字形成制约，忽视了汉语由字组词、由词构句存在的语法上的以及认知上的规律。换个角度看，如汉字的熵为9．65比特，则汉语文本平均压缩比为16/9．65〈1．66，这与国外流行的LHA、ARJ等通用压缩软件压缩汉语文本的结果基本一致，而这些软件根本就没有去发掘属于汉语文本固有的、统计学以及语言认知心理学意义上的冗余现象。因此，汉字熵为9．65比特的结论是不科学的。当然，还有其他有关研究汉字熵值的报告。譬如，有人根据对数十万字的语料进行统计获得的结果，推断汉字熵为8．8比特、9．5比特或类似数值。由于同样的原因，它们同样是既不准确也不科学的。实际上，将汉字纳入具体的上下文语境（单句、乃至篇章）进行研究，通过猜字实验，根据Cover方法，有人估算汉字熵不大于4．1比特。这一结论与根据Shannon方法估算汉字多维熵得到的结论非常一致，后者本质上将汉语语流视作多阶马尔柯夫信源，而不是假定语流中的汉字为前后独立无关的事件。我们认为将字、词纳入言语环境中探讨汉字信息熵是科学的做法，其结论是正确可信的。

于是，汉语文本的可压缩性就体现在其理论上的压缩比16/4．1〉3．9上。从技术实现和实时系统的要求两方面考虑，我们有理由相信，要达到平均压缩比2．5或更好是可以做到的，而这又是现行所有编码方法所远远达不到的。

四、汉语文本压缩的意义

我们透过汉语文本压缩在以下三个方面的应用来看它的意义。

1、以扩大存储介质的信息存储容量为主要目的

①独立的软件产品

以这种形式流行于市场的产品不少，常见的有UNIX系统的compress、DOS系统的ARJ、LHA、PAK、ZIP与UNZIP以及DOS6．0的磁盘容量倍增工具，等等。这类工具软

件主要用来进行文档（压缩）管理，日益受到广大计算机用户的青睐和赞赏。

②作为大中型软硬件系统的重要组成部分

压缩还原算法可以嵌入象全文检索系统、汉语电子词典（辞书）系统、文献档案系统、编辑排版系统、智能文本（intelligent text）系统、以及个人数据助理（PDA）等大中型软硬件系统，使得这些系统更高效、经济、方便使用。以电子出版物为例，它的出现被誉为是对传统图书出版业的伟大革命，电子出版业作为一门新兴的产业正蓬勃发展。统计表明，我国每年由杂志传递 100 亿字的信息，由图书传递 200 亿字的信息，由报纸传递的信息更是多达 400 亿字。由于这 700 亿汉字的信息大都使用计算机排版，所以，只需进行不多的加工即可转换成电子出版物（实际上，1996 年初，北京已经出现专门的组织机构，协调按照统一制作光盘的要求制作学术期刊）。拥有如此丰富的信息资源，我们将可能成为电子出版物的生产、消费大国。可以预见，如果存储的是压缩码，光盘上就可以存储较之不压缩多数倍的书目内容，这不仅节省了存储媒体，大型出版物也可能不再需要分“卷”出版了。其结果，不仅产品有条件达到更理想的性能价格比，自然更可以推动我们的电子出版业迅速成长、发达起来。

2、以提高通信效能为主要目的

随着计算机近、远程网络的建立和普及，尤其是随着卫星通信业的发展，用于数据通信的费用越来越高。以往，人们总是在编/译码的计算费用与数据通信的费用之间寻求一种平衡，甚至不得不放弃某些数据压缩比可观、然而计算费用稍高的算法。但是，现在不一样了。微型计算机在性能上获得飞速发展的同时得到了很好的普及，专用芯片也取得了广泛的应用。这些进步使得计算费用逐渐远远低于数据通信的费用。今天，借助微机或者专用芯片进行数据压缩正日益成为通信过程必不可少的一个环节。譬如，专门的编/译码芯片置于信道的端口，与通信同步进行的编/译码过程可以大大减少数据通信的费用。图 2 描述了现代数据通信过程的典型型式，其中，数据被传送到信道之前在发送端被压缩，而在接收端再根据接收到的压缩码进行译码还原。

让我们看一个例子。据悉，新华社《每日电讯》报租用卫星线路将每日 8 个版面（除去图片外，每版约八九千字的汉语文本）传送给海内外 14 个地区的印刷机构，使得该报能在这些地区同时开印以加速媒体的传播。如果采用数据压缩技术而且假定汉语文本的压缩比达到 2．0，对文本进行压缩能够使得相同时间内现行线路的通信容量得到翻倍。换言之，花费相同的通信费用，能够传送两倍于现在的数据量；或者，只租用一半的线路或一半的时间就能够满足需要。当然，我们不只是《每日电讯》，其他全国性报纸，如《人民日报》、《光明日报》等，也都面临同样的出版发行程序。注意，这里未涉及调制解调器的性能 ——为了追求现有通信线路的高效率，人们往往只是片面地强调调制解调器的作用，没有看到数据压缩的重要性。实际上，调制解调技术与数据压缩技术提供了两种提高通信能力的不同途径，二者并不矛盾，相反，可以结合在一起从两个不同角度大大提高数据通信能力。

我国还是一个发展中国家，电讯业还远远不能满足现代社会由于信息急剧增长对通信的日益迫切的要求。这种形势下，通信资源短缺，通信将成为严重制约国民经济发展的滞后因

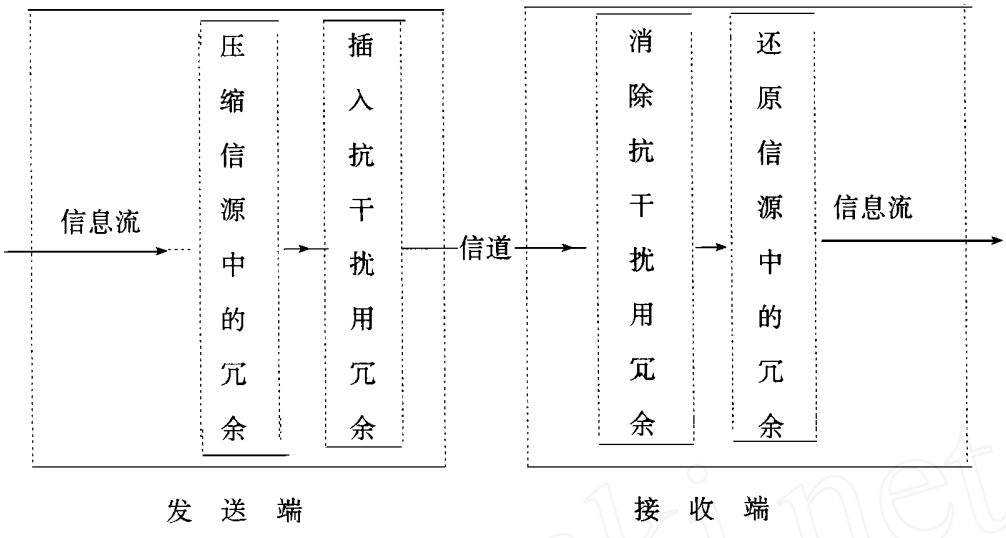

图 2 现代数据通信系统的典型型式

素。正是基于这种认识，我国政府在资金紧缺的情况下仍采取优先发展通信事业的政策。从上面的分析，我们不难看出数据压缩技术有利于缓和通信能力落后于社会需求这一紧张矛盾。

3、提供一类新的汉语文句录入方法

将汉语文本压缩技术应用到汉语文句（非孤立的字、词）的录入系统，是一个全新的、很自然的设想。假定文本的压缩比为 2．5，则意味着其中汉字的熵接近 16/2．5 = 6．4 比特。如果计算机键盘上每键的信息量按 5 比特计算，显然，汉字的平均击键次数为 6．4/5 = 1．28。当然，这里不排斥将压缩技术与现行汉字输入的编码方法有效地结合起来，以便进一步减少平均击键次数。

汉语文本压缩技术除可以在上述三方面得到广泛的应用之外，还可以用于对公用计算机网络上传送的重要政府文件、其他机密材料等进行加密。由于压缩/解压缩算法通常比较复杂，依赖的参数众多，因此，破译密码（即压缩码）十分困难。通过在发信方和收信方内部发行压缩/解压缩程序的专门版本，数据的保密要求便得以顺利实现。在这种意义上讲，压缩技术可谓一举两得。

五、结　　语

研究汉语文本压缩技术不仅具有理论意义，以此为重要支撑技术的相应软、硬件计算机系统更具备广阔的应用前景。成功的汉语文本压缩技术必将产生巨大的社会经济效益。我们期盼更多的人关注并投身这一领域。

参考文献

[1] 王世宁，贵青，依香浓方法估求汉字多维熵，中国电子学会信息论会议文集，1983
[2] 石贵青，徐秉铮，汉字字频分布、最佳编码与输入问题，电子学报，1984 年第 4 期
[3] 郭平欣，张淞艾，汉字信息处理技术，国防工业出版社，1985
[4] 王忠效，基于期望的汉语句子分析，中文信息处理国际会议论文集，1987
[5] 冯志伟，现代汉字和计算机，北京大学出版社，1989
[6] 徐秉铮，吴立中，Victor K. Wei，中文文本压缩的 LZW 算法，华南理工大学学报（自然科学版），1989 年第 3 期。
[7] 贺前华，徐秉铮，彭磊，中文文本压缩的自适应算法，中文信息学报，1993 年第 3 期
[8] 王忠效，基于字符串匹配的通用数据压缩算法，计算机应用，1995 年第 1 期
[9] 王忠效，姜丹，关于 Lempel - Ziv 77 压缩算法及其实现的研究，计算机研究与发展，1996 年第 5 期
[10] 北京语言学院语言教学研究所编，现代汉语频率词典，北京语言学院出版社，1986
[11] Bell T. C., Cleary J. G., Witten I. H., Text Compression. Prentice Hall, Inc., 1990
[12] Williams R. N., Adaptive Data Compression, Kluwer Academic Publishers, 1991

Chinese Text Compression And Its Applications

Wang Zhongxiao

(Institute of Software, CAS　　School of Management, USTC)

Abstract

Chinese text compression has got little attention, but its importance as one supporting technique for many computer applications is beyond any doubt. This paper has investigated current theories and methods of text compression in accordance with the characteristics of Chinese text. It shows that Chinese text compression can even reach an average compression ratio as high as 3.9 theoretically, while it merely stays at around 1.6 with current compression algorithms. Besides, some research directions of Chinese text compression as well as its major applications are also discussed..

Key Words: Arithmetic coding, Chinese text compression, Entropy, Huffman coding, Lempel - Ziv compression algorithm.