“十三五”国家重点出版物出版规划项目

藏文信息处理技术
藏语语言资源建设研究中心资助

藏文文本分析与挖掘技术研究

艾金勇　陈小莹　**著**

西南交通大学出版社
·成　都·

图书在版编目（CIP）数据

藏文文本分析与挖掘技术研究 / 艾金勇，陈小莹著
. —成都：西南交通大学出版社，2020.8
（藏文信息处理技术）
“十三五”国家重点出版物出版规划项目
ISBN 978-7-5643-7590-4

Ⅰ. ①藏… Ⅱ. ①艾… ②陈… Ⅲ. ①藏文—语言信息处理学—研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2020）第 166837 号

“十三五”国家重点出版物出版规划项目
藏文信息处理技术
Zangwen Wenben Fenxi yu Wajue Jishu Yanjiu
藏文文本分析与挖掘技术研究

艾金勇　陈小莹／著
责任编辑／李芳芳
封面设计／墨创文化

西南交通大学出版社出版发行
（四川省成都市金牛区二环路北一段 111 号西南交通大学创新大厦 21 楼　610031）
发行部电话：028-87600564　　028-87600533
网址：http://www.xnjdcbs.com
印刷：四川森林印务有限责任公司

成品尺寸　210 mm×285 mm
印张　11.5　　字数　286 千
版次　2020 年 8 月第 1 版　　印次　2020 年 8 月第 1 次

书号　ISBN 978-7-5643-7590-4
定价　88.00 元

前　言

随着大数据时代的到来，用户可获得的信息也越来越丰富和多样化，但是这些信息中超过 80%是非结构化的，基本上都是以自然语言文本的形式表现出来的，如书籍、新闻报道、研究文章等社交媒体信息和网页，这些来源不一的文本数据构成了一个异常庞大的、具有异构性和开放性等特点的大型分布式数据库。针对这些非结构化的数据，如何有效分析并挖掘出这些数据中所隐含的内容对于信息的理解和数据之间潜在关系的寻找都具有重要的意义。因此，社会各界对于文本信息的分析和挖掘的需求非常强烈，针对文本数据的分析与挖掘理论方法的研究和应用也成为当前自然语言处理方面研究的前沿与热点问题,且应用前景广阔。文本分析与文本挖掘就是把文本型信息源作为分析的对象,利用定量计算和定性分析的方法,从自然语言文本中挖掘用户所感兴趣的模式和知识的方法技术，这种模式和知识对用户而言是新颖的，具有潜在价值。这类研究的最大挑战在于对非结构化自然语言文本内容的分析和理解。这种挑战表现在两个方面：一是文本内容几乎都是非结构化的，而不像数据库和数据仓库，都是结构化的；二是文本内容是由自然语言描述的，而不是纯用数据描述的，通常也不考虑图形图像等其他非文字形式。因此，文本分析与文本挖掘的研究与多个研究领域有密切的关系，如信息检索、信息过滤、自动摘要、文本自动聚类、文本自动分类、计算语言学、数据挖掘、人工智能、统计学等，其所涉及的技术内容也是与自然语言处理、模式分类和机器学习等相关技术密切结合的一项综合性技术。

本书正是在这样一个背景下，针对当前藏文信息处理方面的发展情况，系统性地介绍了藏文文本分析与文本挖掘的基本理论和知识框架，内容涵盖了藏文文本分析挖掘处理的多方面内容。本书一共分为 8 章，1～4 章由陈小莹编写，5～8 章由艾金勇编写。其中第 1 章对于藏文文字、藏文文法、藏文文本特征、藏文编码和藏文文本挖掘进行了总体性介绍，也为后面内容说明提供了前期的理论基础；第 2 章主要介绍藏文字符处理技术，分别针对藏文文字的结构特征、输入技术、规范化处理和结构识别方面进行了讨论；第 3 章从藏文词法分析方面出发，主要介绍了当前研究较多的藏文自动分词和词性标注的基本理论，并在后面列举了一些比较有代表性的研究内容；第 4 章主要是藏文句法分析，从藏文句法分析的主要任务出发，再结合藏文句子特点和类别特征，重点介绍了几种不同技术对于藏文句法分析的实例；第 5 章是藏文文本表示模型研究，重点关注了当前文本特征表示的方法，并在此基础上列举了一些藏文文本表示的研究实例；第 6 章是藏文文本分类算法的研究，通过对文本分类的流程、文本特征项的提取方法、文本分类算法和算法性能评价等理论知识的介绍，讨论了当前各种藏文分类算法的研究情况；第 7 章在藏文文本聚类算法研究中，主要考虑从其与藏文文本分类的对比介绍，通过聚类概念、任务和相关的几个重点问题的理论介绍，重点说明了文

本聚类的一些特殊方法，并且列举了藏文文本聚类算法的研究实例；第 8 章为藏文 web 文本挖掘方法，主要针对 web 文本这一特殊类型的文本，也是当前产生最多的文本的具体处理方法的探讨，并提出了藏文 web 文本处理的具体方法。

本书的出版是在西藏民族大学重点实验室“藏语语言资源建设研究中心”资助下完成的。本书也是西藏自治区哲学社会科学专项资金项目“基于小字符集的现代藏文音节的自动标音方法研究”（批准号：13BYY001）、西藏自治区高校青年教师创新支持计划“基于藏文 web 文本的关联知识挖掘方法研究”（批准号：QCZ2016-44）、西藏自治区科学技术厅自然科学基金项目“基于语义的藏文百科知识问答系统关键技术的研究”（批准号：2016ZR-MY-04）和“面向知识发现的藏文文献知识关联揭示方法研究”（批准号：XZ2017ZRG-56）等项目资助下所完成的项目成果之一。

本书在编写过程中，更登磋和索郎朋措为本书的藏文翻译提出了许多宝贵的意见，此外，本书的编写出版也得到了项目组、研究中心成员以及西南交大出版社编辑老师们的帮助和支持，在此一并表示衷心的感谢！

由于编著人员水平有限，加之时间仓促、可参考资源相对较少，书中难免存在不妥之处，恳请广大读者批评指正！

作　者

2020 年 7 月

目　录

第 1 章

绪　论

1.1　藏文概述

1.1.1　藏文文字的性质

藏文作为藏语的书写符号，它的文字性质也值得关注。认识藏文的文字性质对于藏文字符的分类和处理都有指导作用。藏文的 30 个基本辅音字符都能表音，从功能和结构上看，辅音字符大多能独立成字或独立成词，这意味着辅音字符蕴含了元音要素，这是辅音文字的表现。藏文中的元音符号通常不称为字母，这是因为元音符号结构上不独立书写，不能独立成字或成词。所有元音符号必须附着在辅音字符上来发声。具有元音性质的 ཨ（通俗的说法是半辅音半元音）在藏文文字体系上仍属于辅音字符。

从来源上看，藏文的创制源自中亚（波斯—伊朗）或南亚（天竺—印度）文字，这些文字自身都含有音节文字-辅音（文字）字母的痕迹，辅音文字的特点是元音不独立成符，不是严格意义上的音素字母文字。藏文的每一个音节字中有一个辅音字符，且也只有一个辅音字符具有拼读元音的作用，这个固定拼读元音的辅音字符在音节字中称为基本辅音字符，传统文法称为基字。如果一个音节字不带元音符号，那么固定拼读元音的那个辅音字符或基字一定内含了一个元音 a 的读音。音节字中其他非基字的辅音字符则是真正的音素字母。不过，由于所有辅音字符都能够出现在基字位置成为基本辅音字符，因此，藏文的辅音字符仍然可以被认定为具有音节性质。由这几方面特点可以认为，藏文的文字性质属于非典型音素文字。

从现代语言与文字关系观点来看，随着藏语的发展，当今藏语各地方言的读音与文字读音差别极大。除了 9 世纪后第三次文字厘定曾依据口语读音对文字拼写做过少量的规范外，其后文字拼写逐步固化，相当程度上失去了音素拼写口语的功能。所以有观点认为藏文的字符主要承担区别字形的作用，而不是拼读的作用。以现代拉萨话为例，音素[k]等可以由多种书写形式表示。在组合形式中不发音的字母是历史演变中脱落的音素，但为确保字形的区分在书写上仍然保留。

此外，有人认为藏文 ཀ（[ka]）与 ག（[ga]）读音不同，是不同的音素或字母（文字读音或古藏语中读音是不同音素，即书写形式是不同字母）。从现代语言学角度看，现代拉萨话的浊音已经清化，二者读音基本一致，都是[ka]。但前者的声母来源于古代清声母，后者的声母来源于古代浊声母，本族人从读音上仍感觉它们不同，这是因为按照独立字母读音已经添

加了元音，成了音节，清声母来源的读高调（[ka^{55}]），浊声母来源的读低调（[ka^{12}]或[k^{h}a^{12}]），当然感觉不一样，这种现象也说明藏文具有音节文字-辅音文字的典型特征。

从藏文的文本形式也可观察到藏文的一些有意思的性质和现象。B.A.伊斯特林曾提出文字中除基本符号外，还有 5 类特殊符号，分别是数学符号、专门的科学符号（代数、几何、化学等符号）、标点符号、字母发音符号（如变音符号）和部分大写字母。严格地说，除了大写字母，其他各类符号一定程度上均存在于藏文中，部分还获得相当程度的发展。藏文中的历算符号里就包含有一定科学价值的符号，而藏文文本中丰富的篇章起始符号应该是最为独特的文本符号，具有较强的表意性质。其他诸如敬重符号、各类标点符号也都属于表意符号。这些符号的存在使我们进一步了解到藏文文字表现的多元属性。而出现在藏文文本中的图形示意和会意符号以及宗教法器图形符号或其他象征符号更凸显出藏族文化的特质。

另外，由于藏文具有拼音文字性质，又有二维图形文字形式，书写上的不等长迫使它使用标记分隔相邻音节字，这标记就是藏文文本中的分音点，这样的文本形式在世界文字中是不多见的。例如，汉字无论简繁，相对占据的文本空间基本一致，宽高比例稳定，因此连续文本书写可不留空隙。英语字词虽然不等长，但线性书写，词与词用空格分开，不会混淆。

当然，藏文还有更多独特的文字系统特点，诸如复合元音的表示方法，转写梵文借词的方法，变体简化字形的方法，字的排序和检索方法，标点符号的表示方法，行末断行留空的方法等，藏文文本的这些特征都是在文本处理时应该予以关注的。

1.1.2 藏文文法的主要内容

藏文文法的研究伴随着藏文字的产生和发展的整个过程，具有悠久的历史和深厚的底蕴。历代文法学家都是以《三十颂》和《音势论》为蓝本，对其内容进行注释和补充，形成了传统文法研究的基本框架。在今天，随着社会的进步和藏学研究的快速发展，藏文文法研究的内涵和领域也在不断扩大，现在藏文文法研究的内容已经远远超过传统语法的内容，但是最基本的内容还是传统文法。

传统藏文文法包含有《三十颂》（全称为“授记根本三十颂”）和《音势论》（全称为“字性组织法”）两部分，两者均为偈颂体，文字言简意赅，归纳完备，沿袭了古印度语法的撰写形式。

《三十颂》共 120 句，以 4 句为一颂，全文正好 30 颂，这是《三十颂》名称的由来。《三十颂》主要论述音节结构、正字法及虚词，其内容大致有四个部分：第一部分是字母分类和文字结构；第二部分是格助词与虚词的应用；第三部分是后加字及其缀联形式的重要性；第四部分讲述语法理论的重要性。《音势论》的核心内容是字母的语音分类和动词，大致可分为三部分：第一部分是字性分类，按发音方法将藏文字母分为阴性、阳性和中性等类；第二部分是前加字和后加字的字性分类和约束规则；第三部分论述字性分类和正确缀联的重要意义。因为这两部著作内容高度概括，因而后来的学者在其细微的解释中稍有出入，普遍的观点有司徒派和扎德派之别。

随着语言教学的普及和语言信息处理的发展，加之受现代语言学理论的影响，藏文文法的内容已经根据语言运用的要求发生了根本的变化，从字母拼写、虚词的应用等基本内容延

伸到了词法、语音和句子等层面。从广义上讲，藏文文法应立足于语言应用，注重语法功能和语言结构的描述，应当包括字母、文字学、词汇、虚词、语音和句法等内容。

1. 藏文文字学

文字是交际工具，也是记录语言的工具，一种文字的发展变化，主要是出于完善记录语言的需要。作为工具性的语言，人们普遍的要求和原则是“ 高效 ”“ 便捷 ”“ 易于掌握 ”。人们所追求的语言的目标应该是“ 体系的完善化 ”“ 结构的规律化 ” 和“ 形体的简明化 ”。语言功能的强弱，取决于它的体系结构是否完善，取决于是否具备一套完备的正字法规则。作为拼音文字的藏文来说，其主要表现为字母表的完备和拼写规则的严谨，这些表现应属于藏文文字学，也是藏文正字学的基本内容。藏文字母表与古文字音系、藏文拼写规则、音节结构以及字形的规范是藏文传统文法和正字学关注的基础。但是随着藏文文字的发展，现代藏文文字研究的重要领域则是藏文字起源、形体的变迁、结构的简化以及表意音节的分布等方面的探索。

2. 藏文虚词及功能

虚词是藏文传统文法《三十颂》的核心内容。藏语属于有形态语言，具有黏着性特征，尤其是藏语格标记非常明显。老一辈学者们的分析方法和研究理论的价值取向不同，所以对虚词的分类方法也有所不同。藏语的虚词，珍贝益西扎巴大师和才理夏莺等学者认为，可分为不自由虚词和自由虚词。胡书津和华瑞 · 桑杰等学者则认为藏语虚词分为三类：格助词、不自由虚词和自由虚词。本书中根据面向文本信息处理的角度出发按照第二种的分类方式分为“ 格助词、自由虚词和不自由虚词 ”对虚词进行分类。格助词又分为位格助词（ལ་དོན་ཕྱི་རྒྱུ）、作格助词（ཉིད་ཚེག）、属格助词（འབྲེལ་ཚེག）和从格助词（འཇུང་ཁུངས།）4 类；虚词中格助词是文法说明的主要内容。

格助词是藏文中极其重要的一类虚词，是传统文法《三十颂》的主要部分。这里所说的“ 格 ”基本属于形式逻辑格，藏语的“ 格 ”在词与词之间起区别意义的作用，是确定词与词之间语法关系的一类虚词。世界上很多语言都有“ 格 ”，不同语言中格的表现形式是有区别的，有些语言是通过词的形态变化来表现的，更多的语言是通过介词来表现的。藏语格也是通过格助词这种特殊介词来表现的，藏语由于动词后置，所有格助词都以后置介词的形式附着在体词后面，表达特定的语法关系并区别意义，藏文格助词所表达的语法关系与介词基本一致。藏语格除了第一个主格和第八格呼格以词汇意义本身表达语法关系外，其他六格均以纯粹虚词的形式表达语法关系。此外，按照格的语法功能，除了上述的这 6 种以外，还加了时间格和同体格共八类，其中位格助词分为业格、于格、为格、时间格、同体格等五类。下面具体介绍各种格助词用法和意义。

第二格业格：业格表示一个动作所支配的对象、一个动作发生的地点、一种行为或思想所关注的目标和一种状态持续的地点等功能。

第三格作格：作格主要表示动作的施事者或者施事者所使用的工具、方式、手段、原因和状态等功能。作格助词附在名词后面表示一种施事结构关系，可以在句子中作状语。

第四格为格：为格主要表示动作行为的目的或者为了所需的事物而发出动作以及对动作行为所实施的对象有益等。

第五格从格：从格表示某一事物的来源、一个动作或状态发生的出处、时间或空间的起

止范围、同类事物的比较、排除事物的条件以及实施者的用法等多种功能。

第六格属格：属格表示人或事物之间的限制或领属关系。在语法上，属格是名词短语的构成形式，用前面的名词或名词短语来修饰和限制后面的中心语，相当于定语标志。修饰成分可以是名词、代词和名词短语，也可以是名词性句子。

第七格于格：于格有两种类型：第一种表示人或事物与其存在方位的关系，包含这类于格的句子末尾必须有存在动词出现，相当于英语中的存现句。常见的存在动词有 ཡོད།འདུག ཀངའ།བདོག ཐང་ཆ།མཆིས།གནས།སྡོད་།ཚགས། 等，如 དཀར་ཡོལ་དུ་ཆུ་སྡོད་།。第二种表示事物的领有关系或者某事物属于另一事物，如 བཤད་ཚུ་ཁྱོད་ལ་ཡོད་ཀྱང་བདེན་པ་ང་ལ་ཡོད།。

同体格：表示前面的词语同其后面的动词融合成另一事物或动作，使两者在动作行为变化的结果上具有不可再分的同一体性质。自性格根据结构类型的不同可以分为四种类型：第一种类型是通过某种行为和动作使一事物转变为另一种事物；第二种类型为用在形容词前面表示事物的性质；第三种类型用形容词或者形容词短语修饰动词，表示行为动作的程度、范围、时间等概念；第四种类型主要用于动词的时态表达。

时间格：用于表示一个事件发生或者事物产生的时间。

虚词中除去格助词就是非格助词，因此自由虚词和非自由虚词均可认为是非格助词。非格助词包括 དང་སྦ། ཤག་བཅས། ནི་སྦི། ཕྱི་སྦ། ཞིང་སོགས། རྫོགས་ཚོག ཆུན་སྡུད། དགག་སྦ། བདག་སྦ། འདི་དེ་སྦ། ལོ་ཐག་ཟེར་སྐད། ལོས་ཐང་ཨེ། གིན་ཀྱིན། ཀ་ལ། -ཙོག -ཁར། -ཟད། -མ་ཟད། -སྟ་ཙོག -ཅང་། -མོད། -གཤེས། -ཕྱིར། -གལ་ཏེ། -གལ་སྲིད། - མ་གཏོགས། -ཕྱིན། -དང་འཛེལ། -དང་བཅས། -དང་ཚབས་ཅིག -སྟབས། 等虚词。这些虚词按句法功能也可进一步划分为连词、助词和代词等。传统语法对这些虚词只在形式上做了简单分类，未做详细的功能分类。自由虚词包括语气词、连词、指示代词、疑问代词、否定词、概数副词、数量副词、明喻助词等。非自有虚词包括顺承助词、连词、主人助词、终结助词、离合连词、连词、引用助词、语气助词等。这些虚词的用法、数量和子类在许多藏文文法书籍中都有细微的区别。

3. 藏语词汇学

所有的语言单位（语素、词、词组、句子）都是由词语来充当、构成的，语言的功能（语言功能、交际功能）也是以词语的功能为基础而实现的。因此可以说，词汇既是语言体系存在的基础，也是语法语义规则得以体现的基础。民族语言中的词汇反映了一个民族对世界概念及其关系的认知体系。词汇学是以语言的词汇为研究对象，研究词汇的起源和发展、词的构造、构成及规范。词汇学重点研究以下几个方面的问题：

（1）词的定义。现代词汇学倾向用分解的办法给词下定义，即“ 词 ”是形态的、句法的、语义的具体特征的结合。

（2）词义分析。现代词汇学从概念意义、联想意义和社会意义 3 个方面分析词义。其中，社会意义是现代词汇学与前期词汇学最为不同也是最能体现它的现代性的地方。

（3）不同语言中词汇结构的共性成分。不同语言中存在着共性成分。凡是语言都有语音和语法的体系。语音和语法是封闭系统，有抽象的法则可循，而词汇则是开放的系统，抽象很不容易。现代词汇学重点是研究不同语言中词汇结构的共性成分。

关于藏语词的研究在多识先生的《藏文文法深义明释》和吉太加的《现代藏文文法通论》中有相关章节专门论述，特别是华瑞 · 桑杰在《藏语语法四种结构明晰》中专门论述了词格的语义区别特征。此外，除了术语规范、辞藻及词典编纂外，在传统文法中几乎未涉及词汇

学理论。20 世纪 50 年代以后，国内藏学家陆续出版了大量的历代传统语法著作对于词汇理论有所研究，《民族语文》创刊后，藏语词汇研究有了新的起色，王均、胡坦、黄布凡、瞿霭堂等著名学者在藏语词汇学研究方面取得了丰硕的成果。尽管如此，藏语词汇的研究内容比较零散，还未形成词汇学理论体系，很多的课题有待继续探讨。

4. 藏语句法学

句法学是研究句子结构的学科。句法学研究的对象是作为独立交际单位的句子。句子有三个基本特征：交际语调、述谓性和情态性。其中述谓性是指话语内容与现实的关系；情态性是指话语行为中包含的说话人的态度（确认、愿望、请求、意志等）。这三个基本特征使句子成为区别于词和词组的交际单位，句子作为形式和意义的统一体，其主要任务是描写句法结构和语义结构，以及它们之间的对应关系。根据句法学的主要任务，其研究的主要内容包括：句法成分的概念、句子成分的检测、句法关系（词类和句法成分之间映射关系）、语序与基本语序、句子结构与语言类型、句子类型、被动句与主动句、短语结构理论和句法树等。不同的语言在句法分析中存在一定的差异。

传统藏文文法把格助词虚词的应用与句子结构融合在一起，并没有专门的句法学内容，这是因为藏文的第二格（业格）、第三格（作格）、第四格（为格）、第五格（从格）、第六格（属格）和第七格（于格）决定了藏文最基本的句子结构，即藏文单句结构。第一格（主格）和第八格（呼格）与句子结构并无直接关系。因此，在传统文法《三十颂》中格助词和虚词的用法自然而然成了核心内容。随着信息技术的发展和学科间的交叉融合，当代学者在藏文文法研究中已经将现代语言学理论，尤其是句法理论应用到了藏文句法研究领域。比如吉太加的《现代藏文文法通论》和华瑞·桑杰在《藏语语法四种结构明晰》中用大量篇幅来论述藏语句法和句子结构类型的转换。除此之外，胡坦、瞿霭堂等学者也发表过藏语句法方面文章。随着计算语言学和语料库语言学的发展，对藏语句法分析提出了更高的要求，需要建立大规模的词性标注语料和句法结构标注树库，来获取更多的语法特征，构建规范的句法知识库，建立适应藏语语言特点的词法、句法和语义分析模型，搭建统一的藏语语言信息处理平台。

1.2 藏文文本特征

1.2.1 藏文文字特征

藏文是一种拼音文字，属拼音文字型，分辅音字母、元音符号和标点符号 3 个部分。其中有 30 个辅音字母、4 个元音符号以及 5 个反写字母。

藏文字形结构均以一个字母为核心，其余字母均以此字母为基础前后附加和上下叠加，组合成 1 个完整的字表结构。通常藏文字形结构最少为 1 个辅音字母，即单独由 1 个基字构成，如 ཀ། 。最多由 6 个辅音和 1 个元音构成，如 བསྒྲུབས། 。核心字母叫“ 基字 ”，其余字母的称谓均根据加在基字的部位而得名。即加在基字前的字母叫“ 前加字 ”，加在基字上的字母叫“ 上加字 ”，加在基字下面的字母叫“ 下加字 ”，加在基字后面的字母叫“ 后加字 ”，后加字之后再加字母叫“ 再后加字 ”或“ 重后加字 ”。藏文 30 个字母均可作基字，但是，可作前

加字、上加字、下加字、后加字、再后加字的字母均有限。藏文文字有以下四方面的特征：

（1）音节特征。藏文是一种拼音文字。藏文字以音节为单位，每个音节最少可由 1 个辅音字母构成（元音和上、下加字不能独立成字），最多可由 7 个字母拼合而成，各音节间用音节点分隔。

（2）拼写特征。藏字在结构上由基字、前加字、上加字、下加字、后加字、再后加字及元音以不同结构组成，它不仅具有横向拼写性，也具有纵向拼写性。组成音节时以基字为中心分为前加字、后加字和再后加字，其中前加字、基字、后加字与再后加字横向拼写，而在基字所在的竖直方向上还可能有上加字、基字、下加字和元音的纵向拼写。

（3）形态特征。

藏文的形态特征主要指文字表现出来的外部特征，主要表现的是藏文字的“ 变形特征 ”，指藏文字符在词的不同位置有着不同的显现形式，即从语义上来看，还是那个字符，但字符所显示出的形式却完全不一样，这种特征与藏文字体的结构有很大关系。如藏文字“ ར ”在作为藏文的基字、上加字和下加字时有不同的显现形式，如 ཀ་ཁ་ག་ཅ་ཉ་ར་ཏ 等。

（4）标点符号特征。藏文标点符号形体简单，其使用规则与其他文字的标点符号有别。藏文有一套独立而完整的标点符号体系，常用的标点符号主要有以下几种：一是用于书题或篇首的起始符号云头符（藏文名为 ཡིག་མགོ），用于书题或篇首，说明文章的开始；二是用于文本中每一个音节后面附着的分音点（藏文名为 ཚེག），这种符号主要起着分割音节的作用，是藏文文本中使用频率最高的符号；三是用于句子之间分割的单垂符（藏文名为 ཆོག་ཤད），这个符号大致相当于逗号、顿号和句号的作用 ;四是用于段落结束标志的双垂符（藏文名为 ཉེས་ཤད），表示一小段文字的结束；五是用于卷次末尾的四垂符（藏文名为 བཞི་ཤད），主要用于篇章结束的时候。随着社会的发展，为便于更加准确地表达语义，藏文中也已开始借鉴并使用西方文字的标点符号。

1.2.2 藏文词语特征

词是由语素构成的，词中表示基本意义的语素叫词根，词根是词的词汇意义的主要承担者。加在词根上边表示附加意义的语素叫词缀。“ 词在语言中代表一定意义的、具有固定的语音形式，是能够独立运用的最小的语言单位。” 藏语词语包含有动词时态变化、有格的变化、有助词的黏着连用等多种复杂的语言现象。

藏语的词汇可分为 3 部分：一是固有词，是词汇的中心部分，包括全部单音词，并以双音词占绝对优势，三音词极少。二是借词，比较来说，借词在词汇中占的比重较小。古代借词以汉语和印度语为主，这类借词保留在文献中的比口语中的多，如早期汉语借词“ 茶 ”“ 公主 ”等；早期印度语借词如“ 珍珠 ”“ 水牛 ”等。近代的借词主要从汉语、英语和印度语借入，如近代汉语借词“ 肉包子 ”“ 菜 ”“ 粉条 ”等；近代英语借词如“ 票 ”“ 纸烟 ”“ 球 ”等；近代印度语借词如“ 袜子 ”“ 水泥 ”等。新中国成立以后的新借词主要来源于汉语，一些原来的英语、印度语借词已逐渐为汉语借词所替代，近期汉语借词如“ 书记 ”“ 公司 ”“ 县 ”等。此外，还有一些其他语言的借词，如波斯语借词“ 白酒 ”、回鹘语借词“ 医生 ”等。三是仿制词，仿制词介于固有词和借词之间，形式上属于固有词，内容上又类似借词，早期印度语仿制词主要是译经过程中创造的源于梵语的佛经语词。新中国成立以后藏语新词术

语的发展主要是创造仿制词，在藏语中占重要地位。不过仿制词大多停留在书面上，在口语中尚未广泛使用。因此，藏语口语中的借词，与西南地区其他民族比较来说，数量要少得多。

藏文中的词就构成而言和汉语一样有单纯词和合成词。单纯词是由单个语素构成的词，单纯词又可以分为单音节和多音节。合成词是由两个或两个以上的基本语素构成的词。藏语合成词常见的构词法是合成法和派生法。合成法在现代口语里是能产的、常用的。派生法在古藏语里也是能产的，但是在现代藏语中其重要性已远不如合成法了。下面以拉萨话为例简单介绍藏语语词的构成方式。

（1）合成法。

合成法在藏语合成词里词素的结合关系主要有联合、修饰、支配和表述四种。现分述如下：

① 联合关系构成合成词的两个词素处于平等并列的地位，按照词素的意义又可分为三小类。

● 两个词素意义相同或相近的。如：“བདེ་སྐྱིད་”（幸福）的构成词素分别为“ བདེ་”（安）和“ སྐྱིད་”（乐）、“ ཡང་སྐྱར་”（再）构成词素为“ ཡང་”（又）和“ སྐྱར་”（重复）。

● 两个词素意义相反的。如：“ཚོང་ཚད་”（生意）的构成词素为“ ཚོ་”（买）和“ ཚོང་”（卖）、“ མང་ཉུང་”（多少）的构成词素为“ མང་”（多）和“ ཉུང་”（少）。

● 两个词素并没有相同或相反意义而并列的，如：“ཁ་ལག་”（饭）的词素为“ ཁ་”（口）和“ ལག་”（手）。

② 修饰关系构成合成词的两个词素之间是一个词素修饰另一个词素的关系。处于修饰关系的两个词素的次序，一般有下列两种不同的情况。

● 修饰词素在前，被修饰词素在后。由两个名词或由一个动词和一个名词合成的属于这一类。如：“མེ་ཤེང་”（柴）的构成词素为“ མེ་”（火）和“ ཤེང་”（木），“སྐྱེ་ཤེང་”（树木）的构成词素为“ སྐྱེ་”（生长）和“ ཤེང་”（木）。

● 被修饰词素在前，修饰词素在后。由一个名词和一个形容词或数词合成的词属于这一类。如：“འབུ་དམར་”（蚯蚓）的构成词素为“ འབུ་”（虫）和“ དམར་”（红），“དུས་བཞི་”（四季）的构成词素为“ དུས་”（时）和“ བཞི་”（四）。

③ 支配关系构成合成词的两个词素之间是动作和它所支配的对象的关系。在现代藏语里，由支配关系构成的动词很多。其中主要是由各种名词和表示“ 施 ”“ 放 ”“ 作 ”等义的动词合成的。如：“རྒྱལ་སྐུག་བརྒྱབ་”（反叛）的构成词素是“ རྒྱལ་སྐུག་”（反对）和“ བརྒྱབ་”（施）。由支配关系构成的动词，在两个成分之间可以插入形容词和否定成分。如：“མེ་ མདའ་བརྒྱབ་”（放枪）和“ མེ་ མདའ་མ་བརྒྱབ་”（不要放枪）之间就添加了否定成分“ མ་”。这种合成词里的动词词素，有逐渐失去其原有的词汇意义而变作构词的附加成分的趋势。

④ 表述关系构成合成词的两个词素之间是主语和谓语的关系。如：“ནམ་ལངས་”（天亮）的构成词素是“ ནམ་”（天）和“ ལངས་”（起。

此外，在合成词里，还有一部分由名词或形容词和结构助词“ ལ་”“ ཏུ་”“ ནས་”等构成的副词存在。

（2）派生法。

藏语里常用的派生法是词根加后加成分，根据后加成分体现的意义，附加法主要可以分为以下几类：

● 后加成分“ ཟ ”和动词词根结合，含有进行某项动作的“ 人 ”的附加意义，如“ ཚོང་ཟ ”（商人）就是由后加成分“ ཟ ”和“ ཚོང ”（卖）的词根结合构成。

● 后加成分“ ཟ ”和名词结合，含有与该名词有关的“ 人 ”的附加意义，如“ ཞིང་ཟ ”（农民）就是由后加成分“ ཟ ”和“ ཞིང ”（田地）结合构成。在地名后加“ ཟ ”成分，表示某地方的人，如：“ གཙང་ཟ ”（后藏人）就是“ གཙང ”（后藏）后加“ ཟ ”构成。

● 后加成分“ མ ”“ ཁ ”等与动词词根结合，体现着由某项动作产生的结果或者与某项动作有关的事物的附加意义。如：“ ཟིམ་མ ”（抄本）就是由“ ཟིམ ”（写）的词根后加“ མ ”构成。

● 后加成分“ མ ”“ ཆ ”等与名词词根结合，构成与原来名词意义有关的另一个名词。如：“ རྫ་མ ”（瓦罐）就是由“ རྫ ”（陶土）的词根后加“ མ ”构成。

1.2.3 藏文句子特征

一般而言，藏语的句子又是以动词为中心来组织的，动词居于句子末尾，制约着全句的格局，决定着格助词的添接规则。藏语句子的组织过程就是在词与词、短语与短语之间添加格助词并与句末动词有效结合的过程。藏语句子结构的主要特征概括起来有以下三点：

（1）语序特征。藏语的语序相对稳定。藏语属于 S（主）O（宾）V（谓）型语言，即谓语动词后置型语言。动词谓语是藏文句子的核心，动词谓语决定着主语带不带格标记以及带什么样的格标记。格标记像纽带一样把主语和谓语联结成一个紧密的整体。主语所带的标记，实际上是主谓结构的标记。藏文句子主谓语存在相互照应的一致性关系，具体表现在主语与静态动词的照应及主语与句尾助词的照应两个方面。

① 主语与静态动词。

静态动词包括判断动词、领有动词和存在动词。静态动词与主语在人称上有照应关系，主语的不同人称对应于静态动词的不同变体。如：ང་བོད་རིགས་ཡིན། （我是藏族）中第一人称“ ང ”对应的动词谓语是“ ཡིན ”，而 ཁོང་བོད་རིགས་རེད། （他是藏族）中第三人称“ ཁོང ”对应的动词谓语是“ རེད ”。主谓语照应关系的特点是，即使句中照应关系的一方未出现（如主语未出现），也可根据照应关系中所出现的一方（如谓语）而推知。

② 主语与句尾助词。

藏语位于动态动词之后的句尾助词，是谓语动词形态的延伸和补充，其与主语在人称上保持照应关系。其照应规则类似于静态动词谓语句。如：“ ཁ་ལག་བཟས་ཡ་ཡིན། ”（吃过饭了）“ ནང་ལ་སོག་འགྲོ་གི་རེད། ”（回家去）两句在助词上分别用“ ཡིན ”和“ རེད ”来描述不同的主语对象。

（2）藏语是动词居尾类的一种语言，所有名词都在动词之前，排列成一个名词群并形成一个以动词为中心的语义格系统。在藏语里，每类语义格都带有格标记，形成一个与语义系统相应的格标记系统。因此，句子语义主要借助格助词来表达，藏语句子的主要成分一般都要与格助词相关联，只有这样才能把句子各成分之间的语义关系表达清楚。比如：“ ཐག་གིས་ཤིང་གཅགས། ”（电把树劈开了），如果没有格助词 གིས 就没有办法表达句子的含义。

（3）藏语具有后置性修饰语。在藏语语句中，中心词为名词时，藏语修饰词位置可以前置也可以后置，充当修饰词的形容词、数词、数量词、名词、代词和动词等均可后置，且充当后置性修饰词的词类要远远多于前置性修饰语词类。

1.3 藏文编码标准情况

1.3.1 ASCII 码

计算机领域把诸如文字、标点符号、图形符号、数字等统称为字符。由于所有字符在计算机中都是以二进制来存储的，那么一个字符，比如 a 用哪一个二进制数值表示呢？从理论上说，用任何一个二进制表示都可以，但为保证人类和设备、设备和计算机之间能进行正确的信息交换，人们编制了统一的信息交换代码。最早的计算机内部字符编码表示是 ASCII 码，它的全称是“美国信息交换标准代码”。美国信息交换标准代码（American Standard Code for Information Interchange, ASCII）是由美国国家标准学会（American National Standard Institute, ANSI）制定的标准的单字节（一个字节）字符编码方案。它起始于 20 世纪 50 年代后期，在 1967 年定案，最初是美国国家标准，供不同计算机在相互通信时用作共同遵守的西文字符编码标准，后来经国际标准化组织 ISO 批准成为相应的国际标准，即国际标准 ISO/IEC 646 标准，是目前世界范围内应用最为广泛的编码字符集。

ASCII 码规定字符集中的每个字符均由一个字节表示，并指定了字符表编码表，称为 ASCII 码表。一个字节虽然有 8 位二进制，8 位二进制有 256 个数值，但在制定 ASCII 时，可能是认为英文字符没有 256 个，用 7 位 128 个数值就能表示英文字符，也就用了 7 位二进制来对英文进行编码。7 位的 ASCII 编码称为标准 ASCII 编码，即每个字符的 ASCII 码由七位二进制数组成，标准 ASCII 最高位（b7）可用作奇偶校验位。所谓奇偶校验，是指在代码传送过程中用来检验是否出现错误的一种方法，一般分奇校验和偶校验两种。奇校验规定：正确代码的一个字节中“1”的个数必须是奇数，若非奇数，则在最高位 b7 添“1”；偶校验规定：正确代码的一个字节中“1”的个数必须是偶数，若非偶数，则在最高位 b7 添“1”。

这种 ASCII 码版本包括 10 个阿拉伯数字、52 个英文大小写字母、32 个标点符号和运算符以及 34 个控制码，总共 128 个字符，比如空格“SPACE”十进制编码是 32（二进制 00100000），大写的字母 A 十进制编码是 65（二进制 01000001）等。这 128 个符号只占用了一个字节的后面 7 位，最前面的 1 位统一规定为 0。标准 ASCII 码所包含的字符包括两个部分：一部分是可显示的字符，对应码表中的 32～126，共 95 个，其中 32 是空格，48～57 为 0～9 的 10 个阿拉伯数字，65～90 为 26 个大写英文字母，97～123 号为 26 个小写英文字母，其余为一些标点符号、运算符号等；另一部分为控制字符或通用专用字符，对应码表中的 0～31 及 127，共 33 个，包含的控制符如 LF（换行）、CR（回车）、FF（换页）、DEL（删除）、BS（退格）、BEL（振铃）等；通信专用字符如 SOH（文头）、EOT（文尾）、ACK（确认）等。其中 ASCII 值为 8、9、10 和 13 的分别代表退格、制表、换行和回车字符。它们并没有特定的图形显示，但会依不同的应用程序，而对文本显示有不同的影响。常用 ASCII 码如表 1.1 所示。

表 1.1 常用 ASCII 码

低位	高位							
	000	001	010	011	100	101	110	111

0000	NULL	DLE	SP	0	@	P	`	p
0001	SOM	DC1	!	1	A	Q	a	q
0010	STX	DC2	“	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	,	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	：	J	Z	j	z
1011	VT	ESC	+	；	K	[	k	{
1100	FF	ES	‘	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	～
1111	SI	US	/	?	O	_	o	DEL

表中，上横栏为 ASCII 码的前三位（即高位），左竖栏为 ASCII 码的后四位（即低位）。要确定一个字符的 ASCII 码，可先在表中查出它的位置，然后确定它所在位置对应的行和列。根据行数可确定被查字符低位的四位编码。根据列数可确定被查字符高位的三位编码，由此组合起来可确定被查字符的 ASCII 码。例如，字符 A 的 ASCII 码是 01000001，十进制码值是 65。

由于标准 ASCII 字符集字符数目有限，无法满足实际应用的要求。为此，国际标准化组织（ISO）及国际电工委员会（IEC）又联合制定了 ISO/IEC 2022 标准，即扩展 ASCII 码。它在规定了保持与 ISO 646 兼容的前提下将 ASCII 字符集扩充为 8 位代码的统一方法。扩展 ASCII 码允许将每个字符的第 8 位用于确定附加的 128 个特殊符号字符、外来语字母和图形符号。

1.3.2 中文字符的编码

中文的基本组成单位是汉字，加上需兼容 ASCII 码的几百个英文字符，使用 7 位或 8 位二进制无法表示。加上目前汉字的总数超过 6 万字。数量大，字形复杂，同音字多，异性字多，这就给汉字在计算机内部的表示和处理、汉字的传输与交换、汉字的输入输出等带来了一系列的问题。为此，我国于 1981 年公布了“国家标准信息交换用汉字编码基本字符集（GB 2312—1980）”。该标准规定：一个汉字用两个字节（256×256＝65 536 种状态）编码，同时用每个字节的最高位来区分是汉字编码还是 ASCII 码，这样每个字节只使用低 7 位，这就是所谓的双 7 位汉字编码（128×128＝16 384 种状态），称作汉字的交换码，又称国标码（GB 码）。格式如表 1.2 所示。国标码中每个字节的定义域在 21H～7EH 之间。

表 1.2 国标码格式

b7	b6	b5	b4	b3	b2	b1	b0	b7	b6	b5	b4	b3	b2	b1	b0
0	x	x	x	x	x	x	x	0	x	x	x	x	x	x	x

GB 2312—1980 一共收录了 7 445 个字符，包括 6 763 个汉字和 682 个其他符号。汉字区的内码范围高字节从 B0～F7，低字节从 A1～EE，占用的码位是 72×94＝6 768。其中有 5 个空位是 D7FA～D7EE。因此，GB 2312 支持的汉字还远不能满足要求。1995 年的汉字扩展规范 GBK 1.0 收录了 21 886 个符号，它分为汉字区和图形符号区。汉字区包括 21 003 个字符。2000 年的 GB 18030 是取代 GBK1.0 的正式国家标准，该标准收录了 27 484 个汉字，同时还收录了藏文、蒙文、维吾尔文等主要的少数民族文字。GB 2312、GBK 到 GB 18030 都属于双字节字符集。其中 GB 18030 是中国所有非手持/嵌入式计算机系统的强制实施标准。因此，现在的 PC 平台必须支持 GB 18030，对嵌入式产品暂不做要求，所以手机一般只支持 GB 2312。从 ASCII、GB 21312、GBK 到 GB 18030，这些编码方法是向下兼容的，即同一个字符在这些方案中总是有相同的编码，后面的标准支持更多的字符。在这些编码中，英文和中文可以统一处理。

由于汉字目前既有中国内地地区使用的简体字，也有中国港澳台地区使用的繁体字，因此，汉字编码并不统一，中国大陆地区使用的是 GB 码，而中国台湾地区使用的是 BIG5 码，主要针对繁体字。BIG5 码编码规则是这样的：每个汉字由两个字节构成，第一个字节的范围为 0x81～0xFE，共 126 种。第二个字节的范围分别为 0x40～0x7E，0xA1～0xFE，共 157 种。也就是说，利用这两个字节可定义出 126×157＝19 782 种汉字。这些汉字的一部分是我们常用到的，如一、丁，这些字称为常用字，常用字 BIG5 码的范围为 0xA440～0xC671，共 5 401 个。较不常用的字，如滥、调，称为次常用字，范围为 0xC940～0xF9FE，共 7 652 个，剩下的便是一些特殊字符。

1.3.3 藏文字符编码国家标准

我国藏文信息处理在 20 世纪 90 年代与国际上基本同步，但标准化建设和实际应用却相对滞后。20 世纪 90 年代，藏文信息技术研发单位缺乏沟通和合作，藏文编码是根据机构和企业各自的需要设计的。21 世纪以来，随着各类国际标准的成功应用，为了推动国内藏文信息化建设，国内一些专家提出了两个路线的方针：一是继续对国际标准小字符集技术的深入研究，二是根据国内信息化研究情况研制藏文大字符集国家标准，以此统一国内藏文编码，达到资源共享、避免重复开发的目的。其中藏文大字符集为国家标准，本节将进行具体讲解。

国内大字符集统称或俗称大丁字符集，是音译藏名称 བདར་ཏེན་（brdarten）而来，准确应称为预组合字符集。大丁字符集研究的目标是“ 根据我国现有技术水平、用户需求和产业发展现状，制定适合于我国现有藏文信息处理技术、在国产藏文信息处理软件的实现中具有较高可行性的藏文编码字符集标准。”

大丁字符集研究的指导纲要为“ 在国际标准框架下制定藏文大字符集编码国家标准，定义垂直预组合的藏文字符，应作为我国藏文信息处理发展的策略；同时，不排斥小字符集的技术方案，并积极跟踪研究动态组合技术 ”。

大丁字符集包含基本藏文字符集、扩充集 A 和扩充集 B 三大块。

基本藏文字符集（Basic Set）已经在 0F00～0FFF 编码的全部藏文字符（共有 201 个编码字符和 9 个未用的编码位置）。所收集的字符及各种符号分别由“ 非组合字符 ”和“ 组合字符 ”组成。

扩充集 A（Extension set A）：由基本字符纵向叠加而成的结构稳定的藏文字符和最常用梵音转写字符的集合。扩充字符集 A 共有 1 536 个垂直预组合字符，包括现代藏文（三次规范后的藏文书写形式），如：ཀྱི（kyi）、ཀྲི（kri）、ཀྱུ（rku）等。古藏文（规范之前藏文书写形式），如：རྫི（rdz.i）、སླི（sl.i）和已成为藏文部分的梵音转写藏文字符，如：ཌྷུ（d.dhu）共 962 个字符。还有 574 个最常用梵音转写藏文字符。扩充字符集 A 在 GB 13000 的基本多文种平面专用用户区编码，其编码位置是 F300～F8FF，共占用 1 536 个编码位量。

扩充集 A 中所收藏文预组合垂直结构的结构方式有：

- 辅音+元音：ཏུ（tu）、ཁི（kh.i）；
- 基字+下加字：ཀྱ（kya）、ཁྲ（khra）、གླ（gla）；
- 基字+下加字+元音：ཀྱོ（kyo）；
- 基字+上加字：རྐ（rka）、རྒ（rga）、རྐྷ（rkha）；
- 基字+上加字+元音：སྐི（ski）；
- 基字+上加字+下加字：སྐྲ（skra）；
- 基字+上加字+下加字+元音：རྒྱོ（rgyo）。

上述七种结构方式符合现代藏文的基本结构方式。其中把 ཀྵ、ཧྲ、ཧྭ、ཀྵྲ、ཧླ 作为一个整体，在组合中充当基字，下加字，如：ཧླྀ（ghla）。此外还有两种不符合现代藏文结构的组合方式，具体如下：

- 在上述七种组合形式上添加附加符号、变音符、长音符等。如：ཀྱཱ、ཀྲཱུ。
- 不符合藏文组合规则的梵音转写垂直组合结构。这种结合一般有层叠加和三层叠加结构，如：རྦྷ，以及在这些结构上添加附加符号所构成的叠加结构形式。

扩充集与藏文基本集最大的不同在于，它是在 ISO/IEC 10 646 编码体系结构的框架内对藏文中由基本字符纵向叠加而成，具有稳定结构且使用频繁的藏文和梵源藏字字丁进行预组合编码，即把藏文垂直方向的叠置字符形式看作一个不可分割的整体，只用一个编码来表示，这样就将复杂的二维动态技术转化为一维线性排列技术。比如 བློ་གྲོས“智慧”这个词预组合之后，只需要在横向组合 ཨྲོ、ཤྲ、ས 3 个字符。下图清晰地表示出了藏文扩充集中的藏文字丁。

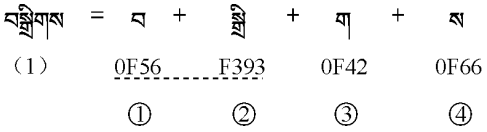

图 1.1　包含扩充集字丁的编码顺序

如果按照小字符集方案编码，上述将会用 7 个编码，而在大丁字符集中只用 4 个编码。基本集与扩充集相比，各有优缺点：使用藏文基本集表示字丁，其优点在于只需要对构成藏文字丁的图形元素进行编码，码点空间的占用量很少；使用藏文扩充集则需要为每一个不同字丁分配一个独立的码点，需要较大的编码空间。但是从藏文存储的角度来说，采用藏文基本集对藏文字丁编码，每个字丁的编码长度取决于构成字丁的元音、辅音的数目，一般需多个编码字符组合而成；采用藏文扩充集对藏文字丁编码，每个字丁对应一个码点。因此，采用藏文基本集编码藏文字丁较藏文扩充集来说，需要使用更多存储空间。

扩充集 B 共有 5 669 个垂直组合字符。它以西藏收集的大字符集、藏学中心提供的梵音转写藏文字符和其他佛教经典中出现的梵音转写藏文字符为主要依据，确定了 5 669 个常用

梵文字符。除扩充集 A 收录的部分字符外,其余都收录于扩充集 B 之中。扩充集 B 在 GB 13000 专用平面 0F 平面上的编码，共占用从 000F0000 到 000F1624 位置的 5 702 个编码位置。

1.3.4 国际字符编码 UNICODE 及藏文字符编码国际标准

由于不同编码在各国家或地区编码时并未考虑其他国家或地区的字符编码，导致了编码空间和编码内容有重叠，同一个二进制数字可能被解释成不同的符号。因此，打开一个文本文件或者页码文件时必须知道原来的编码方式，否则就可能出现乱码。这种问题在信息内容快速和随机传播的互联网时代变得更为突出。

为解决这个问题，历史上有两个独立的创立单一字符集的尝试项目：一个是国际标准化组织 ISO 的 ISO 10646 项目，另一个是由多语言软件制造商组成的协会组织的 UNICODE 项目。1991 年前后，两个项目的参与者都认识到，世界不需要两个不同的单一字符集，故它们合并双方的工作成果，并为创立一个单一编码表而协同工作。目前，两个项目仍都存在并独立地公布各自的标准，但 UNICODE 协会和 ISO 都同意保持 UNICODE 和 ISO 10646 标准的码表兼容，并密切地共同调整任何未来的扩展。

国际字符编码 UNICODE 是一种国际标准的字符集，它包括目前地球上几乎所有正在使用的文字，英文、简体中文、繁体中文等各种复杂的语言都可以正常地显示。这样，只要操作系统和浏览器支持 UNICODE，就可以毫无困难地显示各种字符，不会出现繁体系统无法显示简体中文或相反的情况。UNICODE 标准额外定义了许多与字符有关的语义符号，用于为实现高质量的印刷出版系统提供更好的支持。UNICODE 详细说明了绘制某些语言（比如阿拉伯语）表达形式的算法，处理双向文字（比如拉丁与希伯来文混合文字）的算法和排序与字符串比较所需的算法，以及其他许多相关内容。

藏文字符国际标准编码也是依托于 UNICODE 字符集的，其制定主要包括三个方面的工作：第一，制定藏文编码字符集标准；第二，制定藏文字符键盘布局标准；第三，制定藏文字形标准。

英国标准局 1988 年 7 月 12 日首先提交的 ISO/DP 10 646 藏文提案,给出了 63 个编码点。该提案在很大程度上决定了藏文将以拼音文字的方式，进入 ISO 的 BMP 平面的 A 一区，称为基本多文种平面 BMP（Basic Multilingual Place）。1993 年国家技术监督局、电子工业部和国家民委下达厂研制藏文编码标准的任务。1994 年 5 月藏文编码的第一份中国提案，是一个含 500 多个藏文整字（藏文纵向组合体）的中字符集，提交给了在土耳其召开的国际标准化组织的 WG2 第 25 届会议。1994 年 10 月，按拼音文字结构特征起草的藏文编码中国提案提交到 WG2 第 26 届旧金山会议，这份提案确定了藏文编码国际标准的框架和模式。1995 年 3 月，在瑞士日内瓦召开的 WG2 第 27 届会议上，统一编码联盟组织（Unicode）提交了与我国提案既相近又有差异的另一个藏文编码提案。同年 4 月，我国专家赴美参加在硅谷召开的藏文编码会，会上我国专家以翔实的资料和充分的证据论述了中国的提案，最终形成了以我国提案为主的双方一致确认的藏文编码国际标准提案。该提案提交给 1995 年 6 月在芬兰赫尔辛基举行的 WG2 第 28 届会议，通过 WG2 一级审查，进入 SC2（第二分委员会）一级投票处理阶段。1996 年 4 月，在丹麦哥本哈根举行的 WG2 第 30 届会议上，英国、爱尔兰又提交了藏文编码扩充方案，经过辩论，原藏文编码方案进入第二轮投票阶段。1997 年 6 月 30 日至 7

月 4 日，WG2 第 33 届会议及 SC2 全会在希腊举行，两会在决议中分别宣布：藏文编码已经通过最后一级投票表决，正式形成 ISO/IEC10 646《通用多八位编码字符集》藏文编码国际标准字符集方案。我国也在 1998 年 1 月正式发布了藏文小字符集国家标准《信息技术——信息交换用藏文编码字符集基本集》（GB 16959—1997）。同时，国家标准《信息技术——藏文编码字符集（基本集）24×48 点阵字形第一部分：白体》（GB/T 16960.1—1997）也正式发布。随后在 1999 年发布国家标准《信息技术一藏文编码字符集(基本集)键盘字母数字区的布局》。

建立在 ISO 10646-1 的基本平面 00 组 00 平面的藏文《基本集》占用 192 个码位，机内码为 0F00～0FBF，提供了 168 个编码字符，空缺 24 个码位。藏文的叠置书写使一些辅音在充当上加字和下加字时，形式会发生变化，如 ཝ->ྭ（w）、ཡ->ྱ（y）、ར->ྲ（r）。针对这种情况，必须给这些变形符号赋予新的编码，同时在藏文文本中还常常会遇到一些未编码的图形符号。

为了进一步完善编码，在 1999 年 9 月发布的 Unicode 3.0 版中，增补了部分藏文字符，共涉及 25 个字符。该版藏文字符集增加了 9 个字符或组合用字符，其中包括“ya，ra，wa”的变形形式，分别是 ྱ（0FAD）、ྱ（0FB1）、ྲ（0FB2）；为了满足基字是 nya 时，上加字 ra 不变形的情况，又增加 ཻ（0F6A），其他增加的是 ྷ（0F96）、ྺ（0FAE）、ྻ（0FAF）、ྼ（0FB0）、྽（0FB8）等字符或组合用字符以及图形符号 ྾（0FBE）、྿（0FBF），添补过去旧版中的空缺码位。在此基础上追加了部分空间，机内码从原来的 0F00-0FBF 扩充到 0F00～0FCF，并追加了 14 个图形符号：྅（0FC0）、྆（0FC1）、྇（0FC2）、ྈ（0FC3）、ྉ（0FC4）、ྊ（0FC5）、ྋ（0FC6）、ྌ（0FC7）、ྍ（0FC8）、ྎ（0FC9）、ྏ（0FCA）、ྐ（0FCB）、ྑ（0FCC）、ྒྷ（0FCF）。

Unicode 4.0 版中没有增加藏文字符，但是藏文编码空间进一步扩充到 0FFF 范围，共计 256 个码位。该标准还规定了藏文字丁的编码顺序与藏文字丁的书写顺序一致。图 1.2 清晰地表明了 ISO/IEC 10 646 标准对藏文字丁的编码顺序。

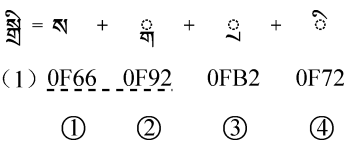

图 1.2 编码顺序与书写顺序

其中字丁（1）的编码为 0x0F66+0x0F92+0x0FB2+0x0F72，而这些附加辅音在基本集中使用“组合用字符”来表示。

藏文小字符集字符包括辅音字符、元音字符、数字、标点、其他符号、减文转写梵文来源的辅音字符、藏文转写梵文来源的元音字符等。藏文小字符标准颁布后，字符叠置技术得到发展。微软公司等国外的专家针对藏文编码的复杂性以及基本字符集技术在应用中动态组合所存在的问题，以叠置引擎技术和 OpenType 字库技术来解决。所谓 OpenType 字库，即在编码字符之外建立 OpenType 字体文件，用来存放所有可能的藏文字符字形（包括叠加字形），然后通过字符编码与字形特征之间的映射关系来显现藏文 Opentype 字体。例如，直接将 ཤེས་རབ “智慧”按照 ཤ、ེ、ས、་、ར、ཕ、ཝ 7 个字符从横向和垂直两个维度上动态组合起来。

采用小字符集方案和前面的大字符集方案都能适用于藏文的书面形式的表示、传输、交换、处理、存储、输入及显现，只是采用的技术路线有所不同。特别是以微软公司为代表的国外集团投入大量人力和财力在 UNICODE 编码体系内采用 OpenType 技术初步解决了藏文小字符集排版、打印、外观、质量问题，他们的设计体系已经形成，产品也已推出，因此对我国采用大字符集方案肯定有不同的意见。如果重新采用大字符集编码方案，将对他们已经成型的技术产品造成经济损失。更何况，国外其他一部分机构、公司不断开发以小字符集标准为基础的各种藏文处理软件和平台，并在此基础上建立了为数不少的藏文资源数据。在这

种情况下，如果采用预组合大字符集方案，现有资源的利用也将成为严重的问题，因此，他们坚持基本字符集符合他们的经济利益需要。而国内企业采用预组合编码方案多年，重新开发小字符集编码体系也不是容易的事情，首先是技术上有一些困难；其次，国内也形成了大量的以大字符集编码的基础资源，放弃大字符集改用小字符集编码开发也将浪费许多资源，这也是当前两种编码技术方案长期共存的重要原因。但是从长远来讲，我们应该看到，采用小字符集方案遵循了藏文国际、国家标准，是无可非议的，这种方案也是将来字符标准化发展的重要方向。

1.4 藏文文本挖掘

1.4.1 文本挖掘基本概念

文本数据挖掘（Text Data Mining）是数据挖掘的一个分支，它是把文本型信息源作为分析的对象，利用定量计算和定性分析的方法，从自然语言文本中挖掘用户所感兴趣的模式和知识的方法和技术，这种模式和知识对用户而言是新颖的，具有潜在价值。文本数据挖掘有时候也简称为文本挖掘（Text Mining）。这里所说的文本包括普通 txt 文件、doc/docx 文件、pdf 文件和 html 文件等各类以语言文字为主要内容的数据文件。

与广义的数据挖掘技术相比较，除了解析各类文件（如 doc/docx 文件、pdf 文件和 html 文件等）的结构所用到的专门技术以外，文本数据挖掘的最大挑战在于对非结构化自然语言文本内容的分析和理解。这种挑战表现在两个方面：一是文本内容几乎都是非结构化的，而不像数据库和数据仓库，都是结构化的；二是文本内容是由自然语言描述的，而不是纯用数据描述的，通常也不考虑图形和图像等其他非文字形式。当然，文档中含有图表和数据也是正常的，但文档的主体内容是文本。因此，文本数据挖掘是自然语言处理（Natural Language Processing，NLP）、模式分类（Pattern Classification）和机器学习（Machine Learning，ML）等相关技术密切结合的一项综合性技术。

所谓的挖掘通常带有“发现、寻找、归纳、提炼”的含义。既然需要去发现和提炼，那么，所要寻找的内容往往都不是显而易见的，而是隐蔽和藏匿在文本之中的，或者是人无法在大范围内发现和归纳出来的。这里所说的“隐蔽”和“藏匿”既是对计算机系统而言，也是对用户而言。但无论哪一种情况，从用户的角度，肯定都希望系统能够直接给出所关注问题的答案和结论，而不是像传统的检索系统一样，针对用户输入的关键词送出无数多可能的搜索结果，让用户自己从中分析和寻找所要的答案。

粗略地讲，文本挖掘类型可以归纳成两种：一种是用户的问题非常明确、具体，只是不知道问题的答案是什么，如用户希望从大量的文本中发现某人与哪些组织机构存在什么样的关系；另一种情况是用户只是知道大概的目的，但并没有非常具体、明确的问题，如医务人员希望从大量的病例记录中发现某些疾病发病的规律和与之相关的因素。在这种情况下，可能并非指某一种疾病，也不知道哪些因素，完全需要系统自动地从病例记录中发现、归纳和提炼出相关的信息。当然，这两种类型有时并没有明显的界限。

文本挖掘技术在国民经济、社会管理、信息服务和国家安全等各个领域中都有非常重要的应用，市场需求巨大，如对于政府管理部门来说，可以通过分析和挖掘普通民众的微博、

微信、短信等网络信息，及时准确地了解民意、把握舆情；在金融或商贸领域通过对大量的新闻报道、财务报告和网络评论等文字材料的深入挖掘和分析，预测某一时间段的经济形势和股市走向；电子产品企业可随时了解和分析用户对其产品的评价及市场反应，为进一步改进产品质量、提供个性化服务等提供数据支持；而对于国家安全和公共安全部门来说，文本数据挖掘技术则是及时发现社会不稳定因素、有效掌控时局的有力工具；在医疗卫生和公共健康领域可以通过分析大量的化验报告、病例、记录和相关文献、资料等，发现某种现象、规律和结论，等等。

文本挖掘与多个研究领域有密切的关系，如信息检索、信息过滤、自动摘要、文本自动聚类、文本自动分类、计算语言学、数据挖掘、人工智能、统计学等。文本挖掘作为多项技术的交叉研究领域起源于文本分类（Text Classification）、文本聚类（Text Clustering）和文本自动摘要（Automatic Text Summarization）等单项技术。大约在 20 世纪 50 年代，文本分类和聚类作为模式识别的应用技术崭露头角，当时主要是面向图书情报分类等需求开展研究。当然，分类和聚类都是基于文本主题和内容进行的。1958 年 H.P. Luhn 提出了自动文摘的思想[Luhn，1958]，为文本挖掘领域增添了新的内容。20 世纪 80 年代末期和 90 年代初期，随着互联网技术的快速发展和普及，新的应用需求推动这一领域不断发展和壮大。美国政府资助了一系列有关信息抽取（Information Extraction，IE）技术的研究项目，1987 年美国国防高级研究计划局（DARPA）为了评估这项技术的性能，发起组织了第一届消息理解会议（Message Understanding Conference，MUC1）。在随后的 10 年间连续组织的 7 次评测使信息抽取技术迅速成为这一领域的研究热点。之后，文本情感分析（Text Sentiment Analysis）与观点挖掘（Opinion Mining）、话题检测与跟踪（Topic Detection and Tracking，TDT）等一系列面向社交媒体的文本处理技术相继产生，并得到快速发展。今天，这一技术领域不仅在理论方法上快速成长，而且在系统集成和应用形式上也不断推陈出新。

1.4.2 文本挖掘主要技术

正如前面所述，文本挖掘是一个多项技术交叉的研究领域，涉及内容比较宽泛。在实际应用中通常需要几种相关技术结合起来完成某个应用任务，而挖掘技术的执行过程通常隐藏在应用系统的背后。例如，一个问答系统（Question and Answering，Q&A）通常需要问句解析、知识库搜索、候选答案推断和过滤、答案生成等几个环节，而在知识库构建的过程中离不开文本聚类、分类、命名实体识别（Named Entity Recognition，NER）、关系抽取（Relation Extraction）和消歧等关键技术。因此，文本挖掘通常不是一个单项技术构成的系统，而是若干技术的集成应用。以下对几种典型的文本挖掘技术做简要的介绍。文本挖掘的主要目标是获得文本的主要内容特征，如文本涉及的主题、文本主题的类属、文本内容的浓缩等。目前，这些技术在处理网络信息资源时非常有效。文本挖掘的具体实现技术主要有如下几种：

1. 文本分类

文本分类是模式分类技术的一个具体应用，其任务是基于内容将自然语言文本自动分配

给预定义的类别。文本分类技术类似于数据库挖掘中的分类技术，不同之处在于它需要预先对文本进行特征抽取，利用文本特征向量对文本进行分类。例如，“中国西藏网”首页划分的内容类别有新闻、时政、文化、援藏、藏医药、文史、宗教、视频、教育等多项，针对一篇新产生的新闻，对于给定的类别，如何根据新闻的内容自动将其划归为某一种类别，是一项具有挑战性的任务，这也是文本分类要解决的现实问题。

2. 文本聚类

聚类就是将一个数据对象的集合分组成为多类或簇。它的分析并不依赖于已知类标记的数据对象。在通常情况下，聚类的训练数据样本没有类标记，它要划分的类是未知的，通过聚类可以产生这种类标记。文本聚类是对给定的文本集根据文本相似度进行聚类的方法。

文本聚类和文本分类的根本区别在于：分类事先知道有多少个类别，分类的过程就是将每一个给定的文本自动划归为某个确定的类别，打上类别标签。而聚类则事先不知道有多少个类别，需要根据某种标准和评价指标将给定的文档集合划分成相互之间能够区分的类别。但两者又有很多相似之处，所采用的算法和模型有较大的交集，如文本表示模型、距离函数、K-means（K-均值）算法等。对于文本聚类而言，通常情况下从不同的角度可以实现不同的聚类结果，如针对统一文本数据集，根据文本内容可以将其聚类成新闻类、文化娱乐类、体育类或财经类等；而根据作者的倾向性可以将其聚成褒义类（持积极、支持态度的正面观点）和贬义类（持消极、否定态度的负面观点）等。

3. 文本表示

文本表示是指用文本的特征信息集合来代表原来的文本。文本的特征信息是关于文本的元数据，可以分为外部特征和内容特征两种类型。文本的外部特征包括文本的名称、日期、大小、类型、文本的作者、标题、机构等信息，文本的内部特征包括主题、分类、摘要等信息。文本的内容特征需要通过分析处理才能得到。文本表示模型也就转换成文本特征的表示模型。文本特征分为一般特征和数字特征，其中一般特征主要包括名词和名词短语；数字特征主要包括日期、时间、货币以及单纯数字信息。特征是概念的外在表现形式，特征抽取是识别潜在概念结构的重要基础。

通常情况下文本特征指的是文本的主题特征。每一篇文章都有一个主题和几个子主题，而主题可以用一组词汇表示，这些词汇之间有较强的相关性，且其概念和语义基本一致。我们可以认为每一个词汇都通过一定的概率与某个主题相关联。反过来，也可以认为某个主题以一定的概率选择某个词汇，由此可以计算出文档中每个词汇出现的概率。为了从文本中挖掘隐藏在词汇背后的主题和概念，人们提出了一系列统计模型，称为主题模型（topic model）。

4. 情感分析与观点挖掘

所谓的文本情感是指文本作者所表达的主观信息，即作者的观点和态度。因此，文本情感分析（Text Sentiment Analysis）又称文本倾向性分析或文本观点挖掘（Opinion Mining），其主要任务包括情感分类（Sentiment Classification）和属性抽取等。情感分类可以看作文本分类的一种特殊类型，它是指根据文本所表达的观点和态度等主观信息对文本进行分类，或者判断某些（篇）文本的褒贬极性。例如，某一特殊事件发生之后（如马航 MH370 飞机失

联等），互联网上有大量的新闻报道和用户评论，如何从这些新闻和评论中自动了解各种不同的观点（倾向性）呢？某公司发布一款新的产品之后，商家希望从众多用户的网络评论中及时地了解用户的评论意见（倾向性）、用户年龄区间、性别比例和地域分布等，以帮助公司对下一步决策做出判断。这些都属于文本情感分析所要完成的任务。

5. 话题检测与跟踪

话题检测（Topic Detection，TD）通常指从众多新闻事件报道和评论中挖掘、筛选出文本的话题，而多数人关心、关注和追踪的话题称为“热点话题”。热点话题发现（Hot Topic Discovery）、检测和跟踪是舆情分析、社会媒体计算和个性化信息服务中一项重要的技术，其应用形式多种多样。例如，“今日热点话题”是从当日所有的新闻事件中筛选出最吸引读者眼球的报道。“2018 热门话题”则是从 2018 年全年（也可能是自 2018 年 1 月 1 日起到当时某一时刻）的所有新闻事件中挑选出最受关注的前几条新闻。

6. 信息抽取

信息抽取是指从非结构化、半结构化的自然语言文本（如网页新闻、学术文献、社交媒体等）中抽取实体、实体属性、实体间的关系以及事件等事实信息，并形成结构化数据输出的一种文本数据挖掘技术[Sarawagi，2008]。

信息抽取中的关系通常是指两个或多个概念之间存在的某种语义联系，关系抽取就是自动发现和挖掘概念之间的语义关系。事件抽取通常是针对特定领域的“事件”对构成事件的元素进行抽取。这里所说的“事件”与日常人们所说的事件有所不同。日常人们所说的事件与一般人的理解是一致的，是指在什么时间、地点、发生了什么事情，所发生的事情往往是一个完整的故事，包括起因、过程和结果等很多详细的描述，而事件抽取中的“事件”往往是指由某个谓词框架表达的一个具体行为或状态。如“市长约见相关负责人”是一个由谓词“约见”触发的事件。如果说一般人所理解的事件是一个故事的话，那么，事件抽取中的“事件”只是一个动作或状态。

7. 文本自动摘要

文本自动摘要简称自动文摘（Automatic Summarization），是指利用计算机分析文章的结构，找出文章的主题语句，然后经过整理、组合、修饰，构成文摘的过程。在信息过度饱和的今天，自动文摘技术具有非常重要的用途。例如，信息服务部门需要对大量的新闻报道进行自动分类，然后形成某些事件报道的摘要，推送给可能感兴趣的用户，或者某些公司想大致了解某些用户群体所发布言论（短信、微博、微信等）的主要内容，自动摘要技术就派上了用场。

1.4.3 文本挖掘的一般过程

文本挖掘过程一般包括文本准备、特征标引、特征集缩减、知识模式的提取、知识模式的评价、知识模式的输出等过程，如图 1.3 所示。

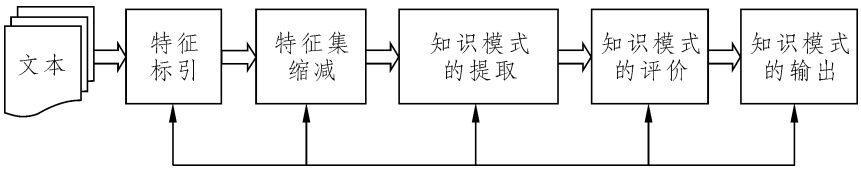

图 1.3 文本挖掘的一般过程

● 文本准备阶段是对文本进行选择、净化和预处理的过程，用来确定文本信息源以及信息源中用于进一步分析的文本。具体任务包括词性的标注、甸子和段落的划分、信息过滤等。

● 特征标引是指给出文本内容特征的过程，通常由计算机系统自动选择一组主题词或关键词可以作为文本的特征表示。

● 特征集缩减就是自动从原始特征集中提取出部分特征的过程，一般通过两种途径：一是根据对样本集的统计分析删除不包含任何信息或只包含少量信息的特征；二是将若干低级特征合成一个新特征。特征集包括过多的特征会增加挖掘的难度，因此，需要在不影响挖掘精度的前提下减少特征项的个数。

● 知识模式的提取是发现文本中的不同实体、实体之间的概念关系以及文本中其他类型的隐含知识。

● 知识模式评价阶段的任务是从提取出的知识模式集 A 中筛选出用户感兴趣的、有意义的知识模式。

● 知识模式输出的任务是将挖掘出来的知识模式以多种方式提交给用户。

1.4.4 文本挖掘面临的困难

文本挖掘工作是一项极具挑战性的任务，一方面，文本挖掘需要处理的对象是自然语言文本，而自然语言处理的理论体系尚未完全建立，因此，文本的分析能力在很大程度上仅仅是基础信息的“处理”阶段，远远未能达到类似于人类一样的文本语义理解基础上的分析处理。另一方面，由于自然语言是人类表达情感、抒发情怀和阐述思想最重要的工具，当人们针对某些特殊的事件或现象表述自己观点的时候，往往采用委婉、掩饰甚至隐喻、反讽等修辞手段，从而使得文本挖掘面临很多特殊的困难，很多在图像分割、语音识别等其他领域能够取得较好效果的机器学习方法在自然语言处理中往往难以大显身手。归纳起来，文本挖掘的主要困难大致包括如下几点。

（1）文本噪声或非规范性表达使自然语言处理面临巨大的挑战。自然语言处理通常是文本挖掘的第一步。由于文本挖掘处理的数据来源于所有可能产生文本的环境，因此，数据结构与规范的书面语相比可能会存在大量的非规范表述，这其中就包括大量的网络文本。根据[宗成庆，2013]对互联网新闻文本进行的随机采样调查，中文网络新闻中词的平均长度约为1.68 个汉字，句子平均长度为 47.3 个汉字，均短于规范的书面文本中的词长和句长。相对而言，网络文本中大量使用了口语化的甚至非规范的表述方式，尤其在网络聊天文本中非规范的表述比比皆是。

噪声数据和非规范语言现象的出现使常规的自然语言处理工具的性能大幅下降，如在《人民日报》《新华日报》等规范文本上训练出来的汉语分词工具通常可以达到 95%以上的准

确率，甚至高达 98%以上，但在网络文本上的性能立刻下降到 90%以下。根据[张志琳，2014]实验的结果，采用基于最大熵（Maximum Entropy，ME）分类器的由字构词的汉语分词方法（Character-Based Chinese Word Segmentation），当词典规模加大到 175 万多条（包括普通词汇和网络用语）时，微博分词的性能 F1 值只能达到 90%左右。而根据众多汉语句法分析方法研究的结果，在规范文本上汉语句法分析器（Syntactic Parser）的准确率可以到达 86%左右，而在网络文本上句法分析器的准确率平均下降 13 个百分点[Petrov and McDonald，2012]。这里所说的网络文本还不包括微博、微信中的对话聊天文本。

（2）歧义表达与文本语义的隐蔽性。

歧义是自然语言文本中常见的现象，如汉语词汇“苹果”既可以指代一种数码产品，也可以表示一种水果，这些都要根据其所处环境进行判断。此外，句法结构歧义同样大量存在，如句子“关于鲁迅的文章”既可以理解为“关于[鲁迅的文章]”，也可以理解为“[关于鲁迅]的文章”。如何解析这种固有的自然语言歧义表达早已成为自然语言处理领域研究的基础问题，但令人遗憾的是这些问题至今没有十分奏效的处理方法，在实际网络对话文本中却又出现了大量人为的千奇百怪的“特殊表达”，例如，“木有”“坑爹”“奥特”等。

有时候为了回避某些事件或人物，故意使用一些特殊用词或者使用英文单词代替某个词汇，如“CBA”等，或者故意绕弯儿，如“请问×××的爸爸的儿子的前妻的年龄是多大？”。请看下面的一则新闻报道：

张小五从警 20 多年来，历尽千辛万苦，立下无数战功，曾被誉为孤胆英雄。然而，谁也未曾想到，就是这样一位曾让毒贩闻风丧胆的铁骨英雄竟然为了区区小利铤而走险，痛恨之下昨晚在家开枪自毙。

对于任何一位正常的读者，无须多想就可以完全理解这则新闻所报道的事件，但如果基于该新闻向一个文本挖掘系统提出如下问题：张小五是什么警察？他死了没有？恐怕目前很难有系统能够给出正确的回答，因为文本中并没有直接说张小五其人是警察，而是用“从警”和“毒贩”委婉地告诉读者他是一名缉毒警察，用“自毙”说明他已经自杀身亡。这种隐藏在文本中的信息需要通过深入的文本分析和推理技术才有可能将其挖掘出来，而这往往是困难的。

（3）样本收集和标注困难。

目前主流的文本挖掘方法是基于大规模数据的机器学习方法，包括传统的机器学习方法和深度学习（Deep Learning，DL）方法，需要大量标注的训练样本，收集和标注足够多的训练样本是一件非常困难的事情。一方面，因为很多文本内容涉及版权或隐私权的问题而难以任意获取，更不能公开或共享；另一方面，即使能够获取一些数据，处理起来也是非常耗时费力的事情，因为这些数据往往含有大量的噪声和乱码，格式也不统一，而且没有数据标注的标准。另外，能够收集到的数据一般属于某个特定的领域，一旦领域改变，数据收集、整理和标注工作又得重新开始，而且很多非规范语言现象（包括新的网络用语、术语等）随领域而异，且随时间而变，这就极大地限制了数据规模的扩大，从而影响了文本挖掘技术的发展。

（4）挖掘目标和结果的要求难以准确表达和理解。

文本挖掘不像其他理论问题，可以清楚地建立目标函数，然后通过优化函数和求解极值

最终获得理想答案。在很多情况下，我们并不清楚文本挖掘的结果将会是什么，应该如何用数学模型清晰地描述预期想要的结果和条件。例如，我们可以从某文本中抽取出频率较高的、可以代表这些文本主题和故事的热点词汇，但如何将其组织成以流畅的自然语言表达的故事摘要，却不是一件容易的事情。

（5）语义表示和计算模型不甚奏效。

如何有效地构建语义计算模型是长期困扰自然语言处理和计算语言学（Computational Linguistics）领域的一个基础问题。自深度学习方法兴起以来，词向量（Word Vector）表示和基于词向量的各类计算方法在自然语言处理中发挥了重要作用。但是，自然语言中的语义毕竟与图像中的像素不一样，像素可以精确地用坐标和灰度描述，而如何定义和表征词汇的语义，如何实现从词汇语义到短语语义和句子语义，最终构成段落语义和篇章语义的组合计算，始终是语言学家、计算语言学家和从事人工智能研究的学者们共同关注的核心问题之一。迄今为止，还没有一种令人信服的、被广泛接受且有效的语义计算模型和方法。目前大多数语义计算方法，包括众多词义消歧方法、基于主题模型的词义归纳方法和词向量组合方法等，都是基于统计的概率计算方法，从某种意义上讲统计方法就是选择大概率事件的“赌博方法”，无论在什么情况下，只要概率大，就会成为最终被选择的答案。这实际上是一种武断的甚至是错误的权宜之计，由于计算概率的模型是基于训练样本建立起来的，而实际情况（测试集）未必都与训练样本的情况完全一致，这就必然使部分小概率事件成为“漏网之鱼”，因此，一律用概率来衡量的“赌博方法”只能解决大部分容易被统计出来的问题，却无法解决那些不易被发现、出现频率低的小概率事件，而那些小概率事件往往都是难以解决的困难问题，也就是文本挖掘面临的最大“敌人”。

通过上面叙述可以发现，由于文本挖掘处理对象的特殊性和处理方式的多样性，使得文本挖掘过程需要与许多语言、语义、图形、图像等抽象对象处理技术相结合，而且这一领域的理论体系尚未建立。所以必将需要一段长期而艰辛的探索过程，但是数据挖掘技术的应用前景极其广阔，且对于智能化发展有巨大的促进作用，因此将来文本挖掘必将成为一个备受瞩目的研发热地，必定会伴随相关技术的发展而迅速成长壮大。