# 语料库的数据管理方式的研究

何婷婷

（华中师范大学计算机科学系，湖北 武汉 430079）

E-mail: hett@163.net

**摘要**：本文全面总结了语料库的几种数据管理方式，分析了各自的长处和不足，可以供语料库建设研究者参考。本文还提出了开发通用语料库管理系统的思想，讨论了通用语料库管理系统应该具备的功能，设计了通用语料库管理系统的体系结构，这种设计思想对于其他的文本数据库的建设也有实际意义。

**关键词**：语料库；数据管理；语料库管理系统

语料库作为语言学研究、自然语言处理研究的重要资源，得到了语言学界和自然语言学界的广泛重视，国内外建设了大量各种类型、各种规模的语料库、语料库应用研究也取得了许多丰硕成果。但是，由于语料库语言学起源于纯语言学研究，语言学工作者必然更多地关注语料的结构、语料库的应用，很少看到有文章对于语料库的数据管理方式，语料库的物理结构作全面的讨论和分析，而这个问题对于语料库的建设的质量和开发效率有重要的意义。本文讨论了语料库的逻辑结构，各种语料库物理存储方式的优缺点，提出了通用语料库管理系统的设想，并讨论了这样的系统的体系结构设计。

## 1 语料库的数据管理方式的发展阶段

现代意义上的语料库是计算机可读、可检索的语料库，所以谈语料库的数据管理方式也是谈用计算机进行的语料库管理。语料库的数据管理方式的发展是随计算机技术的发展而发展的。

(1) 文件管理方式

最早的代表性语料库 Brown 语料库 1964 年开始使用，这时，计算机操作系统已经有了文件管理模块，语料的元数据和语料样本以一个文件的形式存放，由用户规定文件的格式。

(2) 数据库管理方式

二十世纪七十年代，随着数据库技术的普及，特别是二十世纪八十年代，关系数据库系统的普及，使得用数据库进行语料管理变得非常容易，如国家语委主持的“现代汉语语料库”就是以数据库的形式存放。

(3) XML 格式的文件管理方式

二十世纪末、本世纪初，随着文本标记语言 XML 语言的提出，以 XML 语言作为语料文件的组织格式，成为国际上的主流，如国家语委主持的“现代汉语语料库管理系统”就能支持将该语料库以XML格式的文件集合备份、存储和管理。XML 格式的文件与传统的文件管理方式的区别在于：由于有了统一的文件格式规定，就不需要用户自己定义文件的格式。

(4) 多媒体语料数据的管理

笔者认为，口语语料库不仅要存储转录的文本语料，还要存储语音形式的语料、波形图形式的语料，并且要实现多种媒体的语料的同步集成，支持基于文字、声音、波形的查询、统计。现有的数据库管理系统已能很好地支持多媒体数据的存储、管理、演播，但对于支持基于声音、波形的查询、统计还得依

赖于多媒体数据库技术的进一步发展。

以上列举了语料库的几种管理方式，仅仅是按其出现的先后顺序列举，事实上，目前这些管理方式是并存的。

## 2 语料库的逻辑结构

语料库的逻辑结构是对语料数据的形式抽象，要能够比较方便、直观地表示语料的逻辑组织形式，并能很容易地转换成计算机能处理的结构。

大型平衡语料库中的数据常常包括两部分：语料的元数据和语料本身。元数据是关于语料的特性的描述，如语料来源、文体、主题、字数等等，语料库的用途不同，元数据的定义也就不同。如果把一个语料的元数据和语料本身称作一个语料记录，语料库就是语料记录的集合。

一个语料库可以是几个不同的子库的集合，如 ICE 语料库（The International Corpus of English）由 20 个结构相同的平衡的子库组成，每个子库 100 万词，分别取材于英国、美国、加拿大、澳大利亚、新西兰等把英语作为母语的国家，以及印度、新加坡、尼尔尼亚、加勒比海等把英语作为官方语言或第二语言的国家和地区。

一般语料库的逻辑结构可以定义如下：

子库名 1（语料记录号，elememt1，element2，……，elementN，语料形式 1，……语料形式 n）

……

子库名 n（语料记录号，elememt1，element2，……，elementN，语料形式 1，……语料形式 n）

其中，element i 指代元数据项。

一条语料在语料库中有多种存在的形式，如未加工的形式、已分词和加了词性标注的形式、作了句法标注的形式，口语语料库中，还有语料的录音文件、波形图、人工或机器转录的文本、作了各种标注的文本文件。每种语料形式互相参照、互相补充、各有其用途，形成有机的统一体，应通过语料的记录号保持它们的一致性。

## 3 语料库的物理结构

语料库的物理结构是指语料数据的逻辑存储方式，和具体的物理存储细节无关。

语料库可以是文件的集合，其优点是比较容易按字或按词建立索引，访问速度快，存储结构不受关系数据库管理系统的约束，开放性好，能方便地编写程序对其进行加工、检索、统计，现行的语料库加工程序的操作对象大多是纯文本文件。当语料库是纯文本文件的集合，也没有元数据时，这种结构的优点表现明显。不足之处在于：当语料记录包含元数据、语料文本以及其他格式的语料载体时，控制比较复杂，数据的一致性较难控制；当语料库的数据结构改变时，程序要作较大的修改，系统的可扩充性差。

语料库也可以利用数据库管理系统来存放，其优点是可以充分利用数据库系统已有的功能，开发效率高，语料的插入、删除、更新、备份都很容易，特别是对于语料的元数据的查询、更新、统计，非常方便。不足之处在于：早期的关系数据库管理系统（RDBMS）对于大文本数据、音频数据、图形数据的支持不够得力，现在这个问题已经解决，RDBMS 一般都支持“大对象”这种数据类型，字段的宽度可以达到几 G，能管理和支持多种媒体数据；另外，如何对于关系数据库中的大文本，按字或者按词建立索引，实现索引和关系数据库的无缝连接，还是一个有待研究的课题，一些专用的全文检索软件较好地解决了这个问题，但是这些软件缺少对语料库语言学研究的有效支持。虽然全文检索软件能快速地在大量文件中检索出含有特定的字符串的所有文件，但没有提供通用的函数库，供用户编程，完成 KWIC 所要做的基本工作。

下面分别讨论语料库的几种物理结构。

### 3.1 用文件集合管理语料库

早期的语料库一般都是以文本文件的形式存储语料库，一条语料记录作为一个文本文件，语料库是相同结构的语料记录文件的集合。在这种方式下，一般要预先定义语料文本文件的格式。由于一个语料库一种文件格式，数据和程序依赖性强，一个程序对应一个语料库，语料库共享性较差，系统的可扩展性差。在这种结构下，语料库和应用程序的关系如图 1 所示。

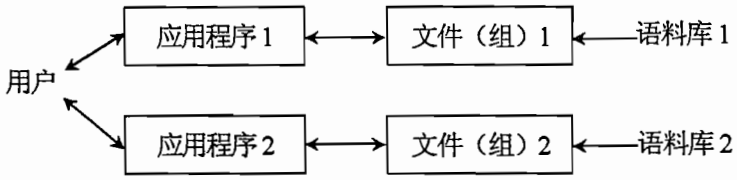


图 1 语料库物理结构示意图 1

改进的方法是用通用的标记语言 XML 语言来组织语料文件。

使用 XML 语言组织语料库，可以减少程序和数据的依赖性，提高语料库的数据独立性，从而提高语料库的共享性。这时，一个语料库的文件是一个或多个 XML 格式的文件集合，用 DTD(数据类型定义)或者 XML SCHEMA（XML 模式）来定义它们的结构，这样，通用的软件（如 IE5.0）就可以依据 DTD 来检查每个语料文件的结构是否规范，解读语料文件的程序就不用向传统的文件系统那样，过多地在程序中去解决物理存储结构的问题，从而提高语料数据和程序的独立性，提高共享性。

在这种结构下，语料库和应用程序的关系如图 2 所示。

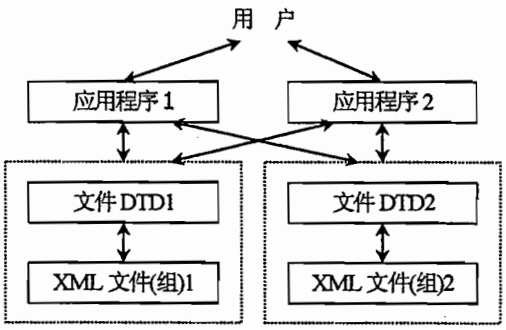


图 2 语料库物理结构示意图 2

### 3.2 用关系数据库管理系统管理语料库

用关系数据库管理系统（RDBMS Ralational Database Management System）来管理语料库时，语料库是一个关系数据库，每个子库是一个关系，一条语料记录是关系中的一个元组，多种形式的语料分别用能支持大文本、多媒体类型的数据字段来存放。这时，语料库和应用程序的关系如图 3 所示。用户通过 RDBMS 访问数据库，只需要了解语料库的逻辑结构，程序和数据具有较高的独立性。

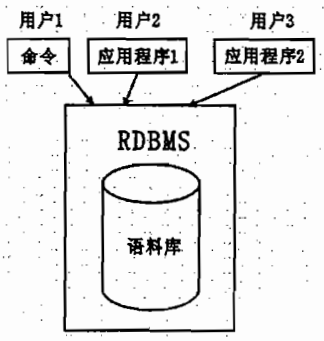


图 3 语料库物理结构示意图 3

### 3.3 元数据用关系数据库管理，语料内容用文件存放

二十世纪九十年代以前，由于关系数据库系统要求数据有规范的结构，它不能很好地处理语料库、网页等半结构化的数据，关系数据库系统也不能较好地处理大文本数据、多媒体数据。这时，如果要用关系数据库系统管理语料库，当语料文件的规模过于大时，

或者无法预知时，则可以采用折衷的策略，用关系数据库存放语料记录的元数据，而语料文本用文件形式存放，通过定义记录号和文件名的映射关系来实现语料的元数据和语料的匹配。在这种结构下，语料库和应用程序的关系如图4所示。

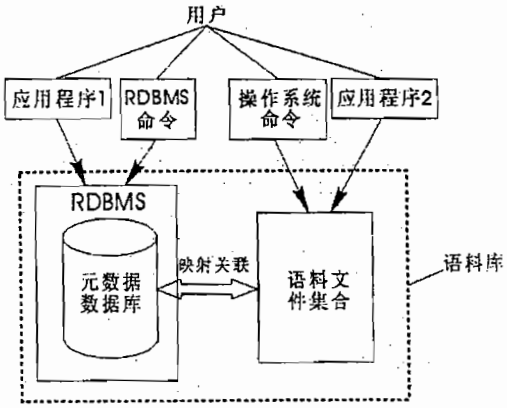


图4 语料库物理结构示意图4

这种语料和元数据分开存放的结构的优点是：语料文件结构单纯，一些和元数据无关的研究可以直接对文件进行，结构简单，程序编写方便；对元数据进行的操作可以用关系数据库进行，效率高。这种结构的不足之处也很多，第一，用户可以直接绕过关系数据库系统对语料进行操作，数据的一致性没办法保证。例如，由于用户的误操作，造成丢失语料文件、改动语料文件名等错误，语料记录的元数据和语料本身就无法匹配。第二，文件系统原有的数据和程序的依赖性强的问题仍然没有解决，哪怕是改变语料文件的存储位置，都会对系统的运行发生影响。

## 4 通用语料库管理系统设计

虽然有些语料库是以文件的形式存储，但从内容、管理方式、应用需求来看，语料库应该是一个数据库系统。目前数据库领域的研究有一个重要趋势就是研究为特定应用需求服务的数据库管理系统，解决传统的关系数据库管理系统所不能很好地解决的问题，如空间数据库、统计数据库、多媒体数据库、时态数据库等。通用语料库管理系统这一概念的提出，也是基于这一思路，它一方面要发挥现有的关系数据库管理系统的优点，而且还要增加语料库研究所需要的功能，研究这种系统的意义不仅仅在于为语料库语言学研究建立一个方便的平台，而且能为一系列文本数据库的建设提供方便，从世界各国信息系统建设的需求和发展趋势来看，建设大量的文本数据库是必然的，其中的很多技术是现在语料库建设研究正在做的工作。

### 4.1 通用语料库管理系统功能设计

通用语料库管理系统要集成语料库建设所需要的各种应用工具的功能，如语料库管理系统、语料库加工工具、语料库检索软件、语料库分析工具，使得语料库建设者与语料库研究人员能像使用文件系统一样，方便地建立语料库，使用语料库。

通用语料库管理系统的功能包括以下几个方面：语料库定义功能、词表定义功能、索引功能、数据存储功能、加工软件集成功能、语料检索和统计功能。

（1）语料库定义功能

语料库管理系统要提供语料库定义语言，让用户定义语料库的数据结构，包括命名元数据字段名称、数据类型、数据宽度等。语料库的字段的数据类型可以包括：字符型、日期型、数值型、文本型、声音数据、波形数据。文本型数据用来储存语料本身，因为长度不固定，所以与字符型数据区别开来，这样便于管理。

（2）词表定义功能

语言的基本组成单位是词，要对语料库进行加工，要分析语料库，都离不开词。语料库管理系统可以提供一些基本的词表，也应该提供词表定义功能、词表编辑功能，让用户自己设计自己的词表。同时还要提供词表查找、排序、统计等基本功能。

（3）索引功能

语料库管理系统不仅要能像关系数据库管理系统那样，按数据项建索引，还能够对语料样本文字按字或者按词建索引，以提高按关键词查找的速度，没有这种索引的支持，很难提高在大规模的语料库中检索数据的速度。

（4）数据存储功能

不管语料库管理系统以什么形式存储语料库，都应该能提供用户一个以二维表（数据库中称二维表为关系）的形式呈现的视图，用户可以直接对关系进行操作，不用了解具体的存储细节。

（4）语料库加工软件集成功能

语料库管理系统本身要具备一些语料库加工的功能，如口语语料的播放、转写、编辑同步进行的功能，根据词表进行词性标注的功能，语料库管理系统要能够集成现有的比较成功的各种语料库加工工具，并提供开放的接口，共用户调用其他的语料库加工工具。

（5）语料检索与统计功能

语料库管理系统要支持常规的关系数据库管理系统已经具备的检索和统计功能，用户界面要更加友好，还要提供常用的语料检索和统计软件所支持的功能，如按关键字串查找、按句型查找，统计字数、词数、句子数、段落数，统计词的同现现象等。

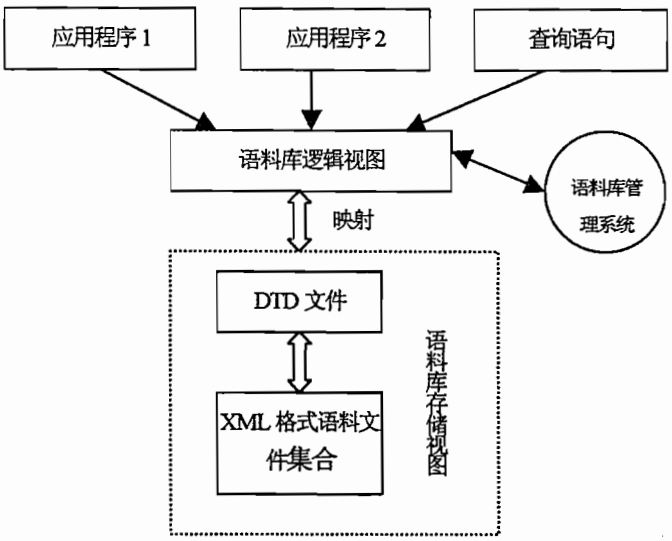

图 5 语料库系统两层视图结构

## 4.2 通用语料库管理系统体系结构设计

为了实现通用语料库管理系统，笔者设计了这种软件的体系结构，探索进一步开发和研究的可行性。

为了使语料库具有通用性、可共享性，具备数据加工方便、数据交换简单等特点，语料库管理系统的底层以 XML 的文本文件存储语料库，用文件 DTD 描述文件的组织结构，可以把这看作是语料库的存储视图；但为了让用户能以简洁、直观的界面，建设语料库、使用语料库、浏览语料，语料库管理系统要以关系的形式提供语料库的逻辑视图，这两种视图之间存在着映射关系，这种映射靠语料库管理系统来支持。

语料库系统的这种结构如图 5 所示，用户程序可以直接按语料库的存储视图对语料库进行访问，也可以通过语料库管理系统，按语料库的逻辑视图访问语料库。

必须声明的是，按这种结构建立的语料库系统，并非关系完备的系统，它虽然以关系的形式提供了逻辑视图，但它允许用户绕过语料库管理系统直接操作语料文件集合。这样，虽然牺牲了关系系统的安全性控制、完整性控制等功能（这些功能，在语料库应用中并不是最重要的），却换来了与原有的文件方式管理的语料库系统的兼容性，一方面，使原有的文件方式管理的语料库能在新系统上运行，而且，原有的一些语料加工工具也可以很方便地升级到新系统中来。

通用语料库管理系统的这种体系结构既发挥了关系数据库的特点，又充分利用了文件系统的特点，在某种程度上克服了二者的不足。

## 5 小结

本文全面总结了语料库的几种数据管理方式，分析了各自的长处和不足，可以供语料库建设研究者参考。本文还提出了开发通用语料库管理系统的思想，设计了通用语料库管理系统的体系结构，但是通用语料库管理系统的真正实现还有待进一步的研究和开发。通用语料库管理系统的应用领域不仅仅是语料库研究，对于大量的文本数据库的建设也有实际意义。

参考文献:
[1] 张小衡等. 面向语料库处理的 CDBMS 和 CSQL. 当代语言学, 1998(1).
[2]McEnery T. and Wilson, A(1996), Corpus Linguistics, Edinburgh University Press

作者简介: 何婷婷（1964－），女, 湖北武汉人, 博士生, 副教授, 主要研究领域为数据库技术, 自然语言处理

# Study On Data Management of Corpus

He Tingting

*Central China Normal University, Wuhan 430079*

E-mail: hett@163.net

**Abstract:** This article makes a summary for data management of corpus, discusses every method's merits and shortcomings. In this article, author puts forward the thinking of universal corpus management, designs its structure.

**Keyword:** Corpus; Data Management; Corpus Management System