面向语料库处理的 CDBMS 和 CSQL

张小衡　石定栩　　香港理工大学

摘要　虽然语料库与数据库在处理上有很多共同之处，但语料库技术的发展却远远落后于数据库。现有语料库的内部结构类似于早期的文件处理系统，也同样具有文件处理系统的严重缺陷。本文借鉴现代数据库的先进技术，着眼于适应语料库语言学的特殊性，提出新一代语料库处理系统的设想，以克服现有语料库的缺陷。在认真分析各种语料库的特点和用户的要求的基础上，建立一个以查询语言 CSQL 为核心的语料库管理系统 CDBMS 设计，并指出一些有待进一步研究的相关问题。

1　引言

随着现代语言学研究的深入发展，原有的研究手段已经不敷使用。语言学家不再满足于坐在摇椅上冥思苦想，而是对大量语料进行统计分析，从中找出更符合语言事实的客观规律来。正因为如此，基于计算机的大规模语料库（corpus）在当代语言学研究中正在发挥越来越大的作用。

计算机语料库处理的主要任务包括语料库的建立、语料库加工和语料库信息检索。目前，这方面的程序一般都限于解决某些局部问题，如语料分词，词性标注和关键词检索等（陈建生，1997），尚未形成完整的系统。同时，这些程序的应用范围一般都十分狭窄，往往是为某一个语料库的特定要求而专门设计的（Hafland, 1991; Sinclair, 1991）。这种单打独闹的必然后果是同一个语料库常常使用几个功能各异的独立程序，各家的语料库都使用各自的程序，而且这些程序提供不同的人机界面，互不通用，给用户带来了极大的不便。本文借鉴现代数据库管理系统 DBMS 技术，在影响广泛的数据库处理语言 SQL 的基础上讨论解决上述问题的有效方法，并试图建立面向语料库处理的 CDBMS（Corpus DBMS）和 CSQL（Coupus SQL）。

2　语料库系统与数据库系统

语料库是存放于计算机中的自然语言文本（Texts）集合。从计算机处理的角度来说，文本也是一种数据，因此完全可以把语料库看成一种（特殊）的数据库。不过，语料库处理和传统的数据库处理却在技术上存在着相当大的差别。

回顾一下数据库系统的发展史（Kroenke, 1995），就可以发现目前的语料库处理系统和作为数据库前身的文件处理系统十分相似，在基本组织结构方面几乎一模一样，如图 1 所示：

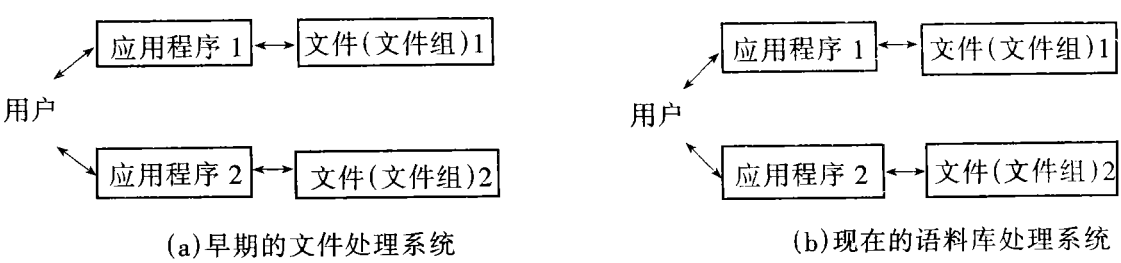

图 1：　DBMS 之前的文件处理系统和现在的语料库处理系统之间的相似结构

本文是作者在第 17 届东方语言计算机处理国际会议上的发言（Zhang, 1997）的修改扩充稿。

在这两种系统中, 供用户使用的应用程序都直接与目标数据文件发生关系, 一对一地起作用。因此, 文件处理系统的主要缺点和局限性也必然会在语料库处理系统中出现, 其中包括:

(11) 数据分离(不同语料库分属不同的语料库处理系统)

(2) 数据经常重复

(3) 应用程序依赖于(语料库)文件结构形式

(4) 不同(语料库)文件常常互不兼容

(5) 用户在使用不同(语料库的)文件时要面临不同的人机界面

在很大程度上说来, 数据库处理系统正是为克服文件系统的上述缺点而研制出来的。所以, 现存的语料库系统借用数据库技术的新系统, 也是预料之中的事情。数据库处理系统的基本结构如图 2 所示:

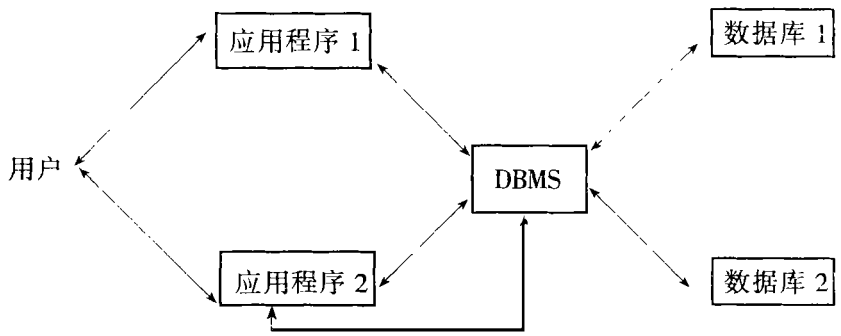

图 2　数据库系统的基本构造

在这一系统中, DBMS(Database Management System 的简称, 即数据库管理系统) 是与文件处理系统不同的地方。其作用极为重要, 既负责管理数据库, 又可以在不同的数据库上运行, 为用户和应用程序提供一个标准一致的界面。按照同样的设计思想, 新一代的语料库系统也应该采用类似的结构:

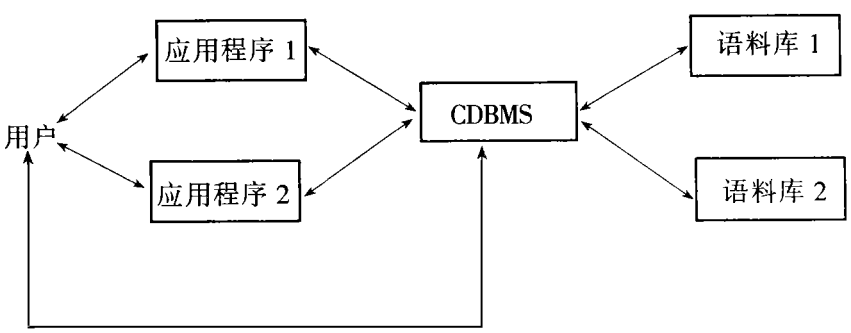

图 3　新一代语料库处理系统的基本构造

图中的 CDBMS 表示语料库 DBMS(Corpus DBMS), 作用是管理一个或多个语料库, 而语料库可根据需要灵活选用。CDBMS 还为用户编写自己的应用程序提供方便的工具和环境。CDBMS 在整个语料库处理系统中的应用与 DBMS 在数据库处理系统中的作用相似, 不仅可以克服前面提到的现有语料库处理系统的缺点, 而且能大大增强整个系统的效率。

对于数据库用户来说, DBMS 的最主要任务是提供一套方便的工作语言, 用于数据库处理, 包括数据库的建立、修改和信息检索等。用户既可用这套语言来编写自己的应用程序, 也可用人机对话的形式使用其中的个别指令。目前, 影响最大效果好的数据库处理语言是用于关系数

据库的 SQL(Kroenke, 1995), 全称为 Structuered Query Language , 即 结构化查询语言 。这是一种通用标准语言, 市面上的所有关系数据库系统都能加以支持, 而且有许多技术可直接应用于语料库处理。如果采用 SQL 的基本作法和有用技术, 不仅能大大增强语料库处理系统的语料处理功能和用户亲和度, 而且有利于语料库和传统数据库的兼容性和结合使用。这对于语言学研究来说也是极其有用的。下面两节将集中讨论如何设计面向语料库处理的 CSQL(Corpus SQL) 语言和 CDBMS(Corpus DBMS) 。

3 语料库及其用途

由于 CDBMS 是位于语料库和用户之间的接口, 要设计一个高质量的 CDBMS, 就得对语料库的用途和语料库本身有充分的了解。

3 1 各种不同的语料库

一般的数据库都具有十分严密的内部结构, 数据用纪录(records) 、域(fields) 、表(tables) 和指针(pointers) 等方式来组织和表示。然而, 大部分语料库的内部结构却是相当松散的。语料库是语言或方言语料的集合, 在计算机里一般以文本文件的形式存在。通常情况下, 语料库里的语言是为了满足某种需要, 按照一定的标准从自然语料中选取或者是特地搜集起来的。

语料库可根据形式、状态和内容等加以分类(Leech and Fligelstone, 1992) 。口语语料库的语料是从演讲、对话、独白、采访等口语材料中转写纪录得来的, 而书面语语料库的内容则直接来自各种书面文献。原始语料库仅包含生语料, 未经任何加工; 而标注语料库中的语料却是经过一定的后期加工, 如分词、词性标注、语义标注和句法结构标注等等。全文语料库收集的是整篇的文章, 而样本语料库则是些摘录的样本章节。广告语料库取材于商业或非赢利广告宣传, 新闻语料库则来自于传媒的报道。尽管语料库种类繁多, 但从计算机处理的角度来看, 都只是文本文件的集合, 一个文本文件就是一长串语言字符。

3 2 语料库的应用

语料库可以为语言研究提供大量的实际资料, 也可以提供精确的统计数字。例如, 为了研究某个词的用法或某一句型的分布, 可以从语料库中抽取所有这一现象的实例, 然后要求提供实际运用中的上下文语境索引(concordance) 。

现有的语料库检索程序一般只能为具体的词或短语建立上下文索引, 而不能为较为抽象的词组形式或句型提供索引。在当代语言学研究中, 某种句型的使用语境及其有关的词语搭配(共现) 情况则是研究的基本要求之一。简单一点的, 为了了解 不但 而且 或是 not only but also 这种连接性副词组或连词对子的使用语境, 就必须在语料库中跳过一定数量的词组, 但又必须在复句的范围内抽取语料。复杂一点的, 像要抽取所有的 AABB 类形容词词组, A 和 B 代表两个不同的字符变量, 如 大大小小 , 形形色色 和 高高兴兴 等, 就要在找出所有符合 AABB 这一表面形式的词组之后再筛选出形容词词组来。当然, 也可以反过来先找出所有的形容词词组, 再找出其中的 AABB 式来。

再抽象一点, 如果要求抽取所有的 SVO 句型, 或者所有的双重主语句型, 就不但需要能确定每个词组的范围, 而且需要确定词组的句法功能, 确定名词词组和动词之间的语义联系。要达到这一要求, 除了必须对语料库进行语义和句法结构标注之外, 还必须建立能够抽取抽象标注的索引系统。

目前, 有些语料库系统能进行简单的句型检索, 但离当代语言学的要求还有很大距离(Zhang and Cheung, 1995) 。除信息检索外, 一定的计算功能(如词频统计, 句型统计) 和数据排序与归类功能(如, 按拼音字母排序, 按语料来源归类, 按句型分类等) 也是很有用的。

4 关于设计CDBMS和CSQL的一些探讨

4.1 语料库信息检索

上面谈到的这些语料库信息检索要求,有可能借助一条数据库SQL指令可以得到简捷的解决。SQL中最重要的指令之一是 SELECT ,其基本格式如下(Kroenke, 1995; Microsoft, 1996):

SELECT 选择项目
FROM 数据库{,数据库}
[WHERE 选择的条件]
[ORDER BY 递升排列| 递降排列]
[GROUP BY 域名]
[HAVING 入选组的条件]

这里的方括号 [ ] 表示里面的内容可以出现也可以不出现,大括号 {} 表示里面的内容可出现零次或多次。竖线 | 表示前后的项目任选其一。 选择项目 通常为一个或多个域名或域函数。

虽然DBMS技术本身已经很成熟,也很值得借鉴,但在语料库处理中运用时,还是要按语言学的实际需要加以选择,并进行必要的修改,才能收到令人满意的效果。例如,面向语料库数据检索的SELECT指令可将基本格式定义为:

CSELECT 选择项目
FROM 语料库{,语料库}
WHERE 语言现象条件
[ORDER BY 排序要求]
[GROUP BY 归组要求]
[HAVING 入选组的条件]

这个语料库信息检索语句叫做CSELECT(CSELECT 是 Corpus SELECT 的简称)。由若干子句(clause)组成,各占一行。子句由关键词和参数组成,参数表示目标必须具有的条件。下面逐条加以解释。

WHERE 语言现象检索条件 子句由关键词 WHERE 和表示目标语言现象应满足的条件的参数这两部分组成。这种条件可以是(出现)某个单词、短语或句型等,也可以是这些简单条件(或它们的数学函数)的逻辑表达式。

为了提高语言现象的表达能力和实际应用上的方便性,我们直接采用UNIX, Windows和DOS等广泛使用的大家所熟悉的字符统配号 ? 和 * 。定义如下:

? 可匹配任何一个字符
* 匹配任何一个语言字符串(不含空格分词标志和标点符号)

因此,在分词连写(Zhang1997)的语料库中,单独一个 * 会匹配一个词,而 a* 或 大* 将匹配以英文字母 a 或汉字 大 开头的词。为了进一步满足语言学研究的需要,可以增加一个表示单词串和短语的统配符号 # :

匹配以分词符号(常为空格)分开的单词串(不含标点符号)。

统配符加自然数(1, 2, 3)则表示统配符变量。如 # 1 和 # 2 是两个单词串变量, * 1 和 * 2 表示两个单词变量,而 ?1 和 ?2 则表示两个单字符变量。因此,下面的中文和英文短语类型:

从 经 到
from via to

可方便地表示为:

从# 1经# 2到# 3

from# 1via# 2to# 3

前面谈到的 AABB 型短语可表示为 ?1 ?1 ?2 ?2 , 而 * 1* 1 则能匹配两个重复的单词, 包括中文的 看看 、考虑考虑 以及英文的 far far(away) 和 long long(ago) 等等。

还可以象 DBMS 一样, 通过数学函数和逻辑运算将这些表示词、短语或句型的简单条件结合成为表达力更强的语言现象表达式, 例如

(* 1者) AND (LENGTH(* 1) Q4)

表示以汉字 者 结尾, 长度等于或小于 4 个字的词。可供语料库处理直接使用的数据库函数有:

COUNT(表示某语言现象的实例的个数)

LENGTH(词或词串的长度)

AVERAGE (平均值)

等, 常用的逻辑运算有:

AND(与)

OR(或)

NOT(非)

等等。

CSELECT 选择项目 子句用来表示, 对于目标语言现象在语料库中的每次出现, 应抽取多大范围的上下文来组织索引表。常用的 选择项目有 :

N: 满足条件的语言现象及其左右各 N 个字符

line: 语言现象所在的文字行

sentence: 语言现象所在的整个句子

根据语料的结构 选择项目 也可用某些模式(pattern) 来表示, 如

[ ]: 包括目标语言现象的最小方括号及其内容

[NP]: 包括目标语言现象的最小名词短语及其内容

这对于作了语法标注的语料库检索来说是很有用的。 选择项目 中也可采用数学函数, 对查出的语言现象作统计处理。如

SELECT COUNT(大*)

表示输出目标词形 大* 的出现次数。

FROM 子句用于指定语料库查寻范围。这可能包括语料库中的特定部分(如新闻部分) , 也可能包括整个语料库或一组语料库。

ORDER BY 子句规定如何将检索出来的索引项目加以排列。一般是按某种语言特征作正序或倒叙排列。

GROUP BY 子句规定如何将检索出来的索引项目分组, 如按某个词的词形进行分组。

最后一个子句, 即 HAVING 子句, 给出合格组别的选择条件。是对 GROUP BY 子句的处理结果的进一步筛选。

下面通过实例说明 CSELECT 语句在语料库信息检索中的应用。

例 1 假设有一个语料来自中国内地、香港和台湾三地报纸的现代汉语书面语语料库 , 叫

本文作者所在的中文及双语学系就建立了这样一个语料库(Zhang, 1995)。

作 mht(Mainland, Hong Kong and Taiwan)。其中的香港语料部分用 hk 表示。那么,查询语句

```
CSELECT    10
FROM mht.hk
WHERE * 大学 exists
ORDER BY * 大学 pinyin ascending
```

将从香港的那一部分语料中找出大学名称的所有实例,并以左右邻各 10 个汉字的形式产生一个按大学名称拼音循序正序排列的 KWIT(Keyword in Context,关键字在上下文中间)的索引。如果将查寻命令中的 * 大学 改为 * ed 并假设三地语料库取材于三地的英文报纸,即

```
CSELECT    10
FROM mht.hk
WHERE  * ed esists
ORDERBY  * ed ascending
```

那么检索的结果将是香港语料中的以 ed 结尾的单词的上下文索引。

例2 本例用到统计函数和逻辑表达。CSQL 命令

```
CSELECT * 1, COUNT(* 1)
FROM mht
WHERE( * 1 exists) AND (LENGTH(* 1)¨Q4)
GROUP BY* 1
HAVING COUNT(* 1)> 10
```

将列出这样的一个单词频度表,内容是长度在四个字以下而且在语料库中出现多于十次的单词,以及它们实际出现的次数。同样,我们还可建立词语的共现频度表,如

```
CSELECT  * 1 自己 , COUNT( * 1 自己 )
FROM mht
WHERE ( * 1 自己 exists)
GROUP BY * 1
```

将会产生单词 自己 和它在语料库中的前指词的共现频度表。这个频度表可用来研究 自己 和前指词的关系。单词频度表和词语共现频度表对于基于概率统计的语料库处理是很有用的。对于基于 HMM(Hidden Markov Model,隐马尔克夫模型)的分词,词类标注等语料加工来说,则是必不可少的。

我们还可像传统的数据库那样,通过建立检索索引(indexes),来进一步提高语料库系统的效率。这样,语料信息检索就可以分两步来做:

1)通过索引表(可由计算机自动产生),找到含有目标语言现象的文本文件

2)遍历这些文件找出所要的语言信息

CSELECT 也可以人 机对话表的形式出现,为此,可借鉴 FoxPro 中的 Query Designer(Microsoft, 1996)。

4 2 其他一些语料库处理问题

除信息查询外,有时候还需要一些 CSQL 的命令来修改语料库中的语料,同一般数据库 SQL 中的 INSERT, DELETE, 和 UPDATE 等命令相似。SQL 的数据修改命令的作用对象通常是数据库中的记录,而 CSQL 命令则用于处理语料库文本文件的插入,删除和修改等。

另一件值得重视的事是为 CDBMS 和 DBMS 之间(尤其是 CSQL 和 SQL 之间)的信息交流建

立一个接口,以便将从语料库中检索出来的信息直接存放于关系数据库中。例如,一个关键词的索引表可以用一个两列的关系数据库表示,一列是该关键字的各次出现实例(含上下文),另一列是这些实例在语料库中的位置(可用文件名和行号表示)。数据库也可用来存放不同的语言现象及其在语料库中出现的次数。这些信息表可根据需要用数据库系统的 DBMS 和 SQL 作进一步加工处理。

我们在实际工作中深深地体会到,语料库系统和数据库系统是语料库语言学的两个重要工具,处理好两者之间的信息交流将使我们的努力事半功倍。此外,为了进一步提高系统的灵活性,还应考虑 CSQL 同 C, Java 等其它重要程序设计语言的相互调用或连接。

同 SQL 的情况一样,CSQL 中的指令可以是以人 机对话的形式单条使用,也可编为应用程序。单条使用时可以按命令文字行(command line)直接调用,也可使用对话表格(form)、报告(report)和菜单等形式。为此,可以借鉴微软公司的 FoxPro 和 Access 在这方面的处理技术(Microsoft, 1996, 1997)。

4 3 语料库的远程访问与分布式语料库系统

目前,有些语料库系统已经上了 intel 计算机网络,以便用户远程访问使用。设计 CDBMS 时也应考虑到这一点,以方便 WWW 和 FTP 访问。

此外,正如传统的数据库能以分布式的形式在计算机网络上运作一样,语料库系统也可以是分布式的,即分散在不同的计算机上。至少有三个方面可实行分布处理:

1) 语料文本的分布:不同的语料库或同一语料库中的不同部分存放在网络上的不同计算机上;

2) 语料库管理系统 CDBMS 的分布处理;

3) 通过自己的 CDBMS 与原来不属同一语料库系统的网上任何一个或多个语料库(根据其网址)配合,组成灵活动态的语料库系统。

5 结语

语料库与数据库在处理上有很多共同之处。然而在技术上,语料库系统的发展却远远落后于数据库系统。现在的语料库系统的内部结构类似于文件处理系统,也同样具有文件处理系统的严重缺陷。本文借鉴现代数据库系统的先进技术,着眼于适应语料库语言学的特殊性,提出新一代语料库处理系统的设想,以克服现有语料库的缺陷。在认真分析考虑各种语料库的特点和用户的要求的基础上,建立一个以查询语言 CSQL 为核心的语料库管理系统 CDBMS 设计,虽然文中的例子大多涉及中文,但整个设计并不局限于特定的语言。

当然,我们的设计还处于初步阶段,很多技术细节问题还未暇顾及,有其它许多相关问题需要讨论。

目前,语料库的结构相当松散,常以一组文本文件的形式出现。是否应该采用某种更为标准方便的模式(model)? 例如数据库就有三种规范模式(Kroenke, 1995),即层次模式(hierarchical model)、网络模式(network model)和关系模式(relational model)。如对这个问题持肯定回答,那就需要考虑如何以现有的语料库实现这一标准结构。

同数据库管理系统 DBMS 一样,语料库管理系统 CDBMS 也可以同专家系统的技术(Durkin, 1994)结合起来,构成语料库专家系统,实现智能信息检索和管理。还可考虑把 CDBMS 建成一个语料库处理系统的外壳,类似于专家系统外壳,为方便快速建造语料库系统提供良好的环境。利用传统的专家系统外壳建造专家系统时,主要工作是建立新的知识库,而语料库外壳还应注重与现有语料库的结合,形成适应不同情况而又符合要求的语料库系统。

另一个值得研究的问题是语义层次的语料检索。从语料库中检索信息不能局限于单词和语法的模式匹配,应该能借助同义词,近义词,词的分类和相关推理等知识,在更深的语义这一级上进行信息检索。例如,要给 机动车辆 这个概念产生上下文索引表,则 小轿车 , 吉普车 , 公共汽车 , 卡车 , 面包车 , 摩托车 等都应该计算在内。自动生成相关的词语匹配模式和确定查询范围对于提高语料库语义信息检索智能度很有作用,需要深入研究探讨。

徐赤裔和何克抗(1988)曾介绍过一种查询语言,用于北京师范大学的 中小学语文教材语料库 ,这是一种简单的表格式语言,自成一体。虽然与数据库的SQL语言关系不大,但其中有不少观点值得参考。

此外,除信息检索以外,语料库处理的其它方面,如自动分词,词性标注,语法和语义标注等也应该包括到CDBMS中来。

参 考 文 献

Durkin J. (1994). *Expert Systems: Design and Development*. Englewood Cliffs. NJ: Prentice Hall.

Microsoft Corporation, (1996) *Visual FoxPro* 5.0. *User s Guide*. Microsoft Corporation.

Microsoft Corporation, (1997) *Microsoft Office*97 *User s Guide*. Microsoft Corporation.

Hofland K. (1991). Concordance programs for personal computers, in Johansson, S. and Stenstrom, A. (eds.) *English Computer Corpora*. Mouton de Gruyter, pp 283– 306.

Kroenke D. M. (1995). *Database Processing: Fundamentals, Design, and Implementation*. Englewood Cliffs, NJ: Prentice_Hall.

Leech G. and Fligelstone G. (1992). Computers and corpus analysis, in Butler, C. S. (ed.) *Computers and Written Texts*. Oxford: Blackwell.

Sinclair J. (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Zhang X. and Cheung Y. S. (1995). Development of a flexible concordance program for large corpora, *Proceedings of the Joint Meeting of the Fourth International Conference on Chinese Linguistics and Seventh North American Conference on Chinese Linguistics*. Madison, USA, June 1995, pp. 496– 504.

Zhang, X. (1997). Corpus SQL, Corpus DBMS and the Internet. *Proceedings of the* 17*th International Conference on Computer Processing of Oriental Languages (ICCPOL* 97), Hong Kong. April 2– 4, 1997. pp. 600– 605.

陈建生,(1997),关于语料语言学。 国外语言学 1997年第一期 pp. 1– 11。

徐赤裔,何克抗(1988),中文语料库的表格式查询语言。 中文信息学报 第二卷第二期 pp. 10– 21。

通讯地址:香港理工大学中文及双语学系