CCL语料库检索系统使用说明

詹卫东（Email：ZWD@pku.edu.cn）

目 录

正 文

一 普通查询 返回目录

1.1 查 询表达式简介

查询表达式由操作符、基本项、简单项、复杂项、过滤项、子句等构成。下面依 次介绍这些单元。

1.1.1 操作符 返回目录

查询表达式中可以使用的特殊符号包括9个：SPACE | $ # + - ~ ! :

这些符号分为四组：

Operator1：SPACE |

Operator2: $ # + - ~

Operaotr3: !

Operator4: :

符号的含义如下：

（一） Operator1：Operator1是二元操作符，它的两边可以出现“基本项”（关于“基本项”的定义见1.1.2）

（1） SPACE（空格）相当于逻辑中的“并”关系。

（2） | 相当于逻辑中的“或” 关系。

在查询表达式中，Operator1可以连续多次使用，即& amp; amp; amp; amp; amp; amp; amp; amp; amp; amp; amp; nbsp; A Operator1 B Operator1 C Operator1 D ... 是合法的查询表达式（参见1.1.3）

（二） Operator2：Operator2是二元操作符，它的两边可以出现“简单项”（关于“简单项”的定义见1.1.3）

（3） $ 表示它两边的“简单项”按照左边 在前、右边在后的次序出现于同一句中。两个“简单项”之间相隔字数小 于或等于Number

（4） # 表示它两边的“简单项”出现于同一句中，不考虑前后次序。两个“简单项”之间相隔字数小于或等于Number

（5） + 表示它两边的“简单项”按 照左边在前、右边在后的次序出现于同一句中。两个“简单项”之间相隔字数刚 好等于Number

（6） - 表示它左边的“简单项”出现于句子中，并且，在右边 相隔Number个字的范围内，-号右边的“简单项”不出现。

（7） ~ 表示它左边的“简单项”出现于句子中，并且，在左边 相隔Number个字的范围内，~号右边的“简单项”不出现。

除 $ 和 + 操作符外，Operator2不能连续多次使用，即只能用Operator2连接两项：A Operator2 B，形成查询表达式。

Operator2中的 $ 和 + 可以连续多次使用，且可以混合使用。见1.2小节的说明。

（三）Operator3：Operator3是一元操作符。

（8） ！ 表示它后面的“简单项”是本次查询的主关键字符串，显 示查询结果时以该“简单项”作为中心来 进行定位。

（四）Operator4：西文冒号 : 是分隔符（delimiter）

（9） : 跟在 author，name，type，pattern 等关键字后面，用于分隔关键字和它们的取值。这样形成的查询式称之为“过滤项”（见下面 1.1.5）

注意：上述操作符不能作为 基本项在语料库中进行检索。

1.1.2 基本项　　返 回目录

指不 包含特殊符号和空格的连续字符串

1.1.3 简单项　　返回目录

简单项可以由以下三种形式的序列组成

（1） 基本项

（2） 基本项1 Operator1 基本项2 Operator1 ...

（3） (基本项1 Operator1 基本项2 Operator1 ...)

注意：在实际表达式中， Operator1 前后不能有空格

1.1.4 复杂项　　返 回目录

复杂项可以由以下三种形式的序列组成

（1） 简单项

（2） 简单项1 Operator2 Number 简单项2

（3） 简单项1 Operator2 Number Operator3 简单项2

其中第二种形式，等价于 Operator3 简单项1 Operator2 Number 简单项2，换句话说，如果以第一个简单项作为查询结果的显示中心，！可以省略。

注意：Number为0和 正整数。Operator2，Operator3前后均不能有空格

Operator2后面的 Number是必须的，不能省略。Number=0表示相邻，Number=1表示间隔1个单位，其余依此类推。

1.1.5 过滤项　　返 回目录

过滤项可以包含以下表达式：

（1）author:简单项

（2）name:简单项

（3）type:简单项

（4）pattern:简单项

（5）ch:简单项

（6）en:简单项

（7）translator:简单项

（8）enname:简单项

说明：

-- “author:简单项”的含义是指“author”关键字后面跟的表达式是上面1.1.3“简单项”所定义的字符串，其余类推 。

-- 通过指定过滤项中author（作者），name（篇名），type（文章类型），ch（中文句子），en（英文句子），用户可以缩小查询语料的范围。

-- 过滤项pattern专门用于查询汉语中的各种模式，比如“AABB”这样的重叠形式，“AB不AB”这样的反复问形式，等等。

-- 过滤项关键字（5）-（8），即ch，en，translator（译者） enname（英文篇名）等是汉英双语语料库检索系统专用的，其余关键字既可用于现代汉语、古代汉语语料库检索系统，也可以用于汉英双语语料库检索系统。

举例:

例1： 想查询“老舍”的语料，在查询表达式中输入“author:老舍”即可；

例2： 想查询“老舍”先生的文章中“A来A去”的用法，在查询表达式中输入“author:老舍 pattern:A来A去”即可。

例3： 查询 ch:以太网 en:Ethernet

意思是：查出汉语句子中包含“以太网”，英语句子中包含“Ethernet”的汉英对照句对 儿。

（**ch表示其后字符串查询范围为汉语句子；en表示其后字符串查询范围为英 语句子。**）

各过滤项的具体取值，用户可以在点击下面的超链接查看：

author　name　type　pattern

汉英双语语料库检索系统专用过滤项的具体取值，可以在汉英双语语料库检索系 统“高级查询”页面查看。

注意：下面第五节 “选择查询范围”也是一种过滤功能，用户可以指定要查询的语料文件所在的文件夹，来缩小查询范围。

1.1.6 子句　返 回目录

子句可以是以下两类表达式：

（1） 复杂项
（2） 过滤项

1.1.7 查询表达式　返回目录

查询表达式可以是以下形式的序列：

（1） 子句
（2） 子句1 子句2 ...

（子句和子句之间需要以空格隔开，表示逻辑“AND”关系）

1.2　使用 $ + 操作符的查询表达式　返 回目录

$ 符号表示间隔小于等于，如“把$10给”表示返回“把”与“给”之间少于10个字符的句子。

查询表达式支持多个“$”连用，如查询“被$10把$3给$2了”，表示 “被、把、给、了”四个关键字在一个句子中共现，并且相互之间有间隔字符的要求，“被”在“把”前出现，二者之间间隔小于10个字符。

+ 符号表示间隔等于，如“把+10给”表示返回“把”与“给”之间等于10个字符的结果。

查询表达式支持多个“$”或“+”连用，如支持查询“我$10你$3他$2 了” “你+3他+2了”。

此外系统也支持“$”与“+”的组合搜索，如“我$10你+3他$2了”， 该查询表示返回“我” 和“你”间隔小于等于10，“你”和“他”间隔等于3，“他”和“了”间隔小于等于2。

二 批量查询　　返回目录

用户可以上传查询文件，文件中可以包含多个普通查询可接受的表达式，默认允 许的最大查询数为30。

文件格式为：每一行是个合法的查询表达式。

返回的查询结果是一个网页（html文件），其中列出每一个查询表达式命中 的结果的个数，每一个查询表达式后的结果个数上有一个超链接，点击后可进入该 查询表达式对应的具体查询结果。

三 模式查询　　返回目录

在模式查询页面，用户可以检索特定的模式，比如“爱V 不 V” “有 X V X”；其中，模式“爱V 不 V”表示查询“爱”跟“不”之间间隔一个字（或词），用户可以指定V的字符个数（长度），两个V是相同的字符串。模式“有X V X”表示查询字符串中包括“有”，“有”后面紧跟的字符串“X”间隔字符串“V”后又重复出现一次，字符串“V”跟字符串“X”不相同。

3.1 模式查询表达式　返回目录

为了与文本中的字母进行区分，要求匹配的变量字符用括号括起来。比如查找模 式“爱V 不 V”，其对应的查询表达式为“爱(V)不(V)”。V的长度也可以指定，比如：

查询表达式为“爱(V,=3)不(V)”，表示要求V的长度为3；

查询表达式为“爱(V,<5)不(V)”，表示要求V的长度 不超过5；

查询表达式为“爱(V,2-5)不(V)”，表示要求V的长度介于2-5之 间。

注意：

（1）模式查询针对的语料并未分词。因此，查询“爱(V)不(V)”，也可 以匹配上“恩爱得不得了”。

（2）模式查询仅是形式意义上的匹配。不见得匹配上的实例在语义上也符合模 式的要求。比如查询“爱(V)不(V)”，也可以匹配上“他的全部的爱是不是在羞辱中消 失了”。这里的“爱是不是”不符合一般的“爱v不v”的语义模式。

（3）上 面举例中，模式查询表达式“爱(V)不(V)”的V并不表示动词（Verb），而是代号，写作X，x等等其他符号也可以，指代任意字符。因此，查询“爱(V) 不(V)”，也可以匹配上“少年，认真的恋个爱 好不好”。

对检索系统自 动返回的模式匹配结果，需要用户根据研究目的来仔细加以甄别。

附：模式查询表达式的BNF范式为

Query ::= <TERM><LPAREN><PlaceHolder><RPAREN> [(<LPAREN><PlaceHolder><RPAREN>) | <TERM>]*
PlaceHolder ::= <PLACEHOLDER><DISTANCE_START>
(
([<DISTANCE_MORE>|<DISTANCE_LESS>| <DISTANCE_EQUAL>]<NUMBER>) |
(<NUMBER><SCOPE><NUMBER>)
)
<PLACEHOLDER> ::= [a-zA-Z]
<DISTANCE_MORE> ::= “>”
<DISTANCE_LESS> ::= “<”
<DISTANCE_EQUAL> ::= “=”
<SCOPE> ::= “-”
<NUMBER>::= [0-9]*

需要注意的是：
(1) 不允许对模式中变量的长度进行多重定义，如“有(V,<8)没(V,>7)”为不合法的查询表达式。
(2) 如果变量长度未指定，则系统默认最大的长度为10。即“爱(V)不(V)”等价于“爱(V,1-10)不(V)”。

3.2 模式查询结果页面的显示 返 回目录

“模式查询”的默认结果页面跟“普通查询”的结果页面相同，参见下面第四节的说明。

在“模式查询”的结果页面上，还增加了一个“统计”按钮。点击“统计”按钮，系统对“模式查询”表达式中的“变项（如x，v等）”进行 计数，并可以按照频次降序或频次升序输出。例如：查询“爱(x)不(x)”模式，返回的默认结果页面为：

共有227条结果

[具体句例略]

点击“统计”按钮后，返回结果为：

共有43条结果

爱动不动: 3 ; (x,动)
爱去不去: 2 ; (x,去)
爱打不打: 1 ; (x,打)
爱念不念: 1 ; (x,念)

爱怕不怕: 1 ; (x,怕)
爱戒不戒: 1 ; (x,戒)
爱懂不懂: 1 ; (x,懂)
爱用不用: 1 ; (x,用)

……

根据这个统计结果可以知道，在CCL语料库中，“爱V不V”中的V有：动，去，打，念，怕，戒，懂，用，……

四 查询结果的显示与下载 返 回目录

4.1 查询结果的显示单位

查询结果以原始语料文件（纯文本格式）中的一个自然文本行为单位输出显示， 用户可以指定查询结果的显示长度（左右n个字范围）。
如果想显示查询关键字所在的整句，可以通过指定足够大的显示长度（比如1000）来实现。当用户指定的显示长度超过原文本行的字符长度时，以文本行长度为 限显示查询结果。

4.2 查询结果的“标亮”和“关键词居中” 返回目录

标亮词：在一个检索结果显示行中以红颜色标出的词，可以有多个；
中心词：是一个特殊的标亮词，显示查询结果的每行文本时，以“中心词”位于页面水平中心位置对齐。

查询表达式中的“复杂项”和“过滤项”中的pattern项目都可以作为“标亮词”。
这里“标亮词”是指跟“标亮词”匹配的字符串。

默认的中心词是第一个“标亮词”，即在用户没有用Operator2指定“中心词”的情况下，系统自动把第一个“标亮词”当作“中心词”。
如果用户用Operator2指定了“中心词”，那么该词为用户指定的“中心词”。

4.3 结果页面的显示宽度 返 回目录

显示宽度定义：
检索结果中句子长度需根据页面宽度进行裁剪（或折行）。如果是有关键词的检索，则关键词必须居中。如果检索条件没有指定关键词，则所有返回的检索结果中， 最长的结果长度不应超过页面宽度 。

根据需求，页面显示分为下面几种情况

搜索类型	指定字数	结果形式	处理方式
单语查询	最多显示字数 左XX右XX	HTML	如果指定字数不超 过页面宽度，按照指定字数； 否则，按照系统默认的最大长度显示
		下载	按照指定字数
双语查询	最多显示字数 左XX右XX	HTML	目标语言(查询输入的语言)按照指定字数；对 照语言不进行压缩
		下载	

4.4 结果页面中关键词一例一行　　返 回目录

对于有关键词的检索，如果一个文档（document）——在语料库中对应为原自然文本的一段 ——中包含了n个检索关键词，则该文档被显示n次，每次都以关键词居中显示。

为了对索引的命中数(结果数或document数)与关键词的命中数进行区分，同时兼顾执行效率，使用如下的呈现方式：
最左边使用两级编号，其中“1.1”表示第一个document的第一条查询命中；“9.1”，“9.2”，“9.3”，“9.4”，“9.5”，表示属 于同一个document，但是分别为不同的命中，即9号文档共有5个实例匹配用户指定的查询条件。

4.5 检索结果的排序　返回目录

用户可以指定按照“中心词”左边字符串排序，或按照“中心词”右边字符串排 序。排序方式为字符内码（GB码）降序。

排序依据包括：

（I）如果是返回符合检索表达式条件的句子，则可以根据关键词上下文环境中字符串的内码排序（上文、下文、上下文）。

（II）如果是返回符合检索表达式条件的句子（或段落、篇章）中的特定的词或模式，可以根据跟关键词构成搭配关系的字符串的频次排序。

除在网页上显示的检索结果支持排序外，下载的结果文件中也支持排序。

4.6 检索结果的扩展　　返 回目录

对于有关键词的检索，在返回的结果页面上有一个“上下文”链接，点击后，可 以扩展显示当前例句的上下文。

4.7 检索结果的下载 返回目录

用户可以将查询所得结果保存到自己本地计算机的磁盘上。在查询结果显示网页 上，用户可以根据需要指定下载结果的条数（缺省为500条），点击“下载”按钮，查询结果即以txt文件形式保存到本地磁盘上。每句之后在【 】内注明了该句的出处、作者、路径等信息。（如果条数较多，文件会比较大，下载速度缓慢，请耐心等待，不要重复提交下载请求）。

五 选择查询范围 返 回目录

在普通查询、批量查询、模式查询页面，系统都提供了“选择范围”按钮，点击 该按钮，系统弹出语料库目录结构的树状显示，用户可以通过鼠标点击选取框checkbox来指定查询范围。

语料库文件目录的树状结构可以在网页上“展开-收缩”显示，每个节点前有一 个选取框（checkbox），如果选中一个节点，则默认情况下，该节点的所有子孙节点都被选中，反之。如果清除一个子节点，默认情况下，该节点的所有子 孙节点都被清除。

六 在结果中检索 返 回目录

对于复杂的查询要求，可以尝试通过多次查询完成，即利用“在结果中查找”功 能，逐次逼近检索目标。

“在结果中检索”的功能是指在上一次检索基础上，用户输入新的查找条件，然后点击“在结果中检 索”按钮，系统会将此次用户输入的查找条件跟上一次的查找条件（LastQuery）合并（AND运算），执行一次查询。查询结果是上一次查询结果的一个 子集。

比如：您想查找“宁可……也”的例句，同时不希望“也”后面出现“不”这样 的否定词。

您可以先输入查询表达式“宁可$10也”，返回的结果是包含“宁可”和 “也”，且二者相隔10字以内的句子，然后您再输入查询表达式“也-4不”，这样就可以把“也”后面4字范围内有“不”的句子剔除掉了。

七 对英语词组的查询 返 回目录

词组两端用引号确定边界，比如："take care of"

八 英语词形处理 返 回目录

比如用户查 take 的时候，也可以将 took taken takes taking同时作为查询结果返回。

上面七、八两项功能仅针对汉英双语对齐语料库

九 基于IP地址的访问权限管理 返 回目录

双语语料库的访问需要用户在指定的IP地址范围内才能使用检索系统。程序对 用户IP地址进行了检查。通过扫描用户IP地址是否在预定义的IP地址列表中，来判断当前用户是否有权访问双语语料库查询系统。

十 查询举例 返 回目录

查 询式例子 1:

计算机硬件

意思是: 查出所有包含“计算机硬件”的句子。

查 询式例子 2: 返 回目录

把 被

意思是: 查出所有包含“把”，同时也包含“被”的句子，即两个关键字之间无次序限制，无距离限制，只需要在一句范围内。

查 询式例子 3: 返 回目录

把|被

意思是: 查出含有“把”或“被”的句子，两个关键字只需有一个在句中出现，就作为查询结果输出。

查询式例子 4: 返 回目录

把-4不

意思是: 查出含有“把”，但在“把”右边4个字范围内不含“不”的句子。**注 意：- 号属于opertaor2，其后必须有数字，且不能有空格。**

查 询式例子 5: 返 回目录

给~4把

意思是: 查出含有“给”，但在“给”左边4个字范围内不含“把”的句子。**注 意：~ 号属于operator2，其后必须有数字，且不能有空格。**

查询式例子 6: 返 回目录

与其$10不如

意思是: 查出同时含有“与其”和“不如”的句子，并且“与其”在先，“不如”在后出现，间隔10字以内。

查询式例子 7: 返 回目录

能力#3大

意思是: 查出同时含有“能力”和“大”的句子，且“能力”和“大”之间的间隔在3个字之内，二者的先后次序不受限制。

查询式例子 8: 返 回目录

吃+3亏

意思是: 查出同时含有“吃”和“亏”的句子，并且“吃”在先，“亏”在后出现，二者之间刚好间隔3个字。

查询式例子 9: 返 回目录

被$10!给

意思是: 查出同时含有“被”和“给”的句子，并且“被”在先，“给”在后出现，二者之间间隔10个字以内。显示查询结果时，以“给”为“中心词”，即“给”居中对 齐。

查询式例子 10: 返 回目录

(把|被)$10给

意思是: 查出同时含有“把”和“给”的句子，并且“把”在先，“给”在后出现，二者之间间隔10个字以内。
或者，查出同时含有“被”和“给”的句子，并且“被”在先，“给”在后出现，二者之间间隔10个字以内。

查 询式例子 11： 返 回目录

(把 |被)$10!给

意思是: 查出同时含有“把”和“给”的句子，并且“把”在先，“给”在后出现，二者之间间隔10个字以内。
或者，查出同时含有“被”和“给”的句子，并且“被”在先，“给”在后出现，二者之间间隔10个字以内。
显示查询结果时，以“给”为“中心词”，即“给”居中对齐。

查询式 例子 12： 返 回目录

了$0(。 |？ |， |！)

意思是：查出“了”与标点符号“。？，！”等紧邻出现的句子。这实际上就部 分地达到了查询“句尾了”（汉语学界一般所说的“了2”） 的目的。

查询式 例子 13： 返 回目录

所以 author:老舍

意思是：在现代汉语语料中查作家老舍的文章中“所以”的用例。

查询式 例子14： 返 回目录

模式查询：有一种(X,=2)叫(Y,=2)

X和Y是两个不同的字符串，且长度都为两个汉字字符。（返回结果：有一种力 量叫感动；有一种放弃叫成全；……）

查询式 例子15： 返 回目录

模式查询：爱(V,=1)不(V)

V字符串的长度为1个汉字字符； （返回结果：爱借不借；爱理不理；爱管不管……）

模式查询：爱(V,<5)不(V)

V字符串的长度不超过5个汉字字符；

模式查询：爱(V,1-5)不(V)

V字符串的长度介于1-5个汉字字符之间。

－－－　正文完　－－－