

# 大规模中文语料库检索技术研究

余一骄<sup>1</sup> 刘 芹<sup>2</sup>

(华中师范大学语言学系 武汉 430079)<sup>1</sup> (武汉大学计算机学院 武汉 430072)<sup>2</sup>

**摘 要** 大型中文语料库的检索需求与通用文本检索系统差异很大,需要研究专门的中文语料库检索技术。Cici 是一个面向 GB 规模的中文语料库检索系统,它高效地实现了 4 种针对汉语研究的检索功能,涉及词性的检索、词或短语的重叠式检索、带通配符的汉字串检索、汉字串频次检索。实现以上检索功能的关键是:先统计语料库的 N-gram 汉字串频次,并将统计结果分别按频次大小及汉字串 Unicode 编码进行倒排序索引。对用户输入的检索请求,先检索汉字串频次统计结果,向用户反馈一个备选汉字串集合;然后让用户参与检索优化过程,选择正确性较高的汉字串;最后在语料库中检索用户选定的检索词。

**关键词** 汉字,语料库,检索,词性,N-gram

**中图分类号** TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.045

## Key Retrieval Technologies in Large-scale Chinese Corpus

YU Yi-jiao<sup>1</sup> LIU Qin<sup>2</sup>

(Department of Linguistics, Central China Normal University, Wuhan 430079, China)<sup>1</sup>

(Computer School, Wuhan University, Wuhan 430072, China)<sup>2</sup>

**Abstract** The query requirements of large-scale Chinese corpora are different from those of general text retrieval system. Cici v2.0 is a Chinese corpus search system and provides linguistic query services: part-of-speech search, reduplicated words search, wildcard search, and Chinese N-gram string occurrence search. The N-gram string occurrences are accounted and indexed by Unicode and frequency respectively. The search procedure is divided into three steps. First, the Chinese N-gram occurrence statistic records are searched and the candidate n-gram strings are produced. Then, keywords are searched according to user's linguistic need; At last, these Chinese strings are searched selected by users in the corpora and the final results are returned.

**Keywords** Chinese character, Corpus, Information retrieval, Part-of-speech, N-gram

## 1 前言

充分利用大规模语料库的语言统计和检索结果来开展语言学是当前应用语言学的经典模式<sup>[1]</sup>。汉语研究者不仅期望语料库规模大、语料来源广,还希望语料检索软件功能强大且简单易用。随着从互联网中获取中文语料难度的降低,一些中文文本语料库已达到 GB 规模。获得更多的文本生语料已不存在太多技术挑战,然而语料库检索软件尚存在以下不足,还有待进一步改进。

第一,汉语研究者的检索需求没被软件开发人员充分理解,有些涉及汉字串查询的检索功能没有提供。大多数情况下汉语研究者不是到语料库中找确定的汉字串,而是查找符合指定语言学规律的汉字串。例如输入“AABB”,期望立即得到语料库中全体 AABB 重叠格式的汉字串(如“高高兴兴”、“弯弯曲曲”等),并按频次从高到低排序;然后用户从中选择感兴趣的汉字串;最终获得感兴趣重叠串所在的语料文本。一些现有的语料库检索软件只是照搬通用文本检索技术,不具备上述重叠式检索功能。第二,语料库检索系统不支

持词语搭配、词性搭配检索功能,而这些功能却是汉语研究者亟需的。例如语法研究者期望输入“V 一把”(V 表示一个动词),立即获得“玩一把”、“赌一把”、“抓一把”等检索结果。通用文本检索技术研究不考虑涉及词性的检索,但它对提高汉语研究者的检索效率却至关重要<sup>[2]</sup>。由此可知,面向汉语研究的中文语料库检索技术与通用的中文文本检索技术有差异。语料库检索系统开发者须全面了解语料库用户的特定检索需求,然后在基于汉字串匹配的基础上,研制新的语料库检索系统。

2011 年至今,我们开发了针对 GB 规模语料库的中文文本 N-gram 串统计与检索软件 Cici<sup>[3]</sup>,以及规模为 1.08GB 的现代汉语生语料。Cici 已在华中师范大学语言学系测试、使用两年多,为汉语研究提供了有力的支持<sup>[4]</sup>。Cici 主要涉及大规模中文文本 N-gram 汉字串频次统计以及中文语料检索技术。关于 N-gram 汉字串频次统计技术已在文献<sup>[3]</sup>中深入讨论,本文只讨论大型中文语料库检索技术。

本文第 2 节分析大型中文语料库特殊的检索需求;第 3 节讨论如何实现确定性 N-gram 汉字串频次检索;第 4 节讨

到稿日期:2014-03-17 返修日期:2014-06-13 本文受教育部人文社会科学研究项目:逻辑推理与词义匹配相融合的中文网页语义检索技术研究(10YJA740120),湖北省教育厅人文社会科学研究项目:基于语义理解的中文网页检索方法研究(2010b032)资助。

余一骄(1978-),男,博士,副教授,主要研究方向为计算语言学、计算机网络,E-mail: yjyu@mail.ccnu.edu.cn;刘 芹(1978-),女,博士,副教授,主要研究方向为计算机网络、信息检索。

论汉字串重叠式检索;第5节论述带通配符的汉字串检索;第6节讨论涉及汉语词性的检索,最后总结全文。

## 2 针对中文语料库的特殊检索需求

除了必须支持传统的汉字串匹配检索功能外,汉语研究者还期望中文语料库检索系统具有以下功能。

### 2.1 准确的汉字串频次检索

通用文本检索系统大多只提供所包含汉字串的文本数,却不提供汉字串的总频次。一个汉字串在一篇语料中可能多次出现,汉语研究者不仅希望获得包含该汉字串的语料篇数,还想获得汉字串在语料库中出现的总次数(本文称为汉字串的频次)。频次大小有助于判断汉字串是否高频出现,所出现的语料篇数能反映汉字串的通用程度。如果某个高频汉字串只在某篇作品或极少数作品中出现,语言学家也不会太重视这样的高频串。

在语料不变的前提下,无论何时输入同一检索串,语料库检索系统反馈的结果须保持一致,否则该语料库检索系统的正确性会受语言学家质疑。通常 Google、百度提供的检索结果数量与反馈网页数量相差甚远。以 2014 年 1 月 8 日在 Google 中检索“余一骄”为例,Google 最初显示有 32200 个结果,但实际仅提供 43 个网页链接。在中文语料库检索中,不能出现以上不一致的现象。

### 2.2 汉字串重叠式检索

汉语研究者通常不是对某一个词或短语进行研究,而是对一类词、短语进行研究。一篇语言学论文中往往会列出数十个甚至上百个属于同一类型的词或短语。汉语研究者输入的检索式大多不是确定的汉字串,而是一个具有特定语言学含义的英文串或中英文混合串。在语料库语言学中,把这些检索式称为“人工语言检索”或“自然语言和人工语言相结合的检索”<sup>[5]</sup>。

重叠式是汉语语法研究中的经典问题<sup>[6-8]</sup>。例如汉语研究里常用 ABB 表示一类三字串重叠式,其中 A 和 B 代表不同的汉字,“喜洋洋”就是一个 ABB 类型叠词。汉语研究者期望输入特定的重叠格式后立即获得满足该格式的所有汉字串。例如,输入“ABAC”,获得“流里流气”、“不三不四”等汉字串,然后从中选择部分高频的汉字串做进一步检索。除了 ABB、ABAC 等纯英文串表示的重叠式外,用户还会输入中英文混合的重叠格式,例如“AA 地”(其含义是:第三个字必须是“地”,前面两个汉字须重叠)。中英文混合的检索式,有利于用户快速地找到特定的语料。由于汉字数量众多,中英文混合的检索式难以计数。中文语料库检索系统不但要支持纯英文串的重叠式检索,还必须支持用户自定义中英文混合的检索式。

### 2.3 带通配符的汉字串检索

在词语搭配、短语结构分析等研究中,常需要进行带通配符的语料检索。例如文献[9]研究“X 以上”的语法规律,“两瓶以上”、“六十分以上”、“两瓶或两瓶以上”等都是该文中的有效语料。“两瓶”和“六十分”是短语,“两瓶或两瓶”是非词非短语(此处“两瓶以上”才是短语)。文献[10]研究“非 X 不可”,“非去不可”、“非你不可”、“非成功不可”都是对该文研究有帮助的语料。“去”和“成功”是动词,“你”是代词。

由上述例子可知,X 仅代表一个汉字串,它不对该汉字串是否为词语、短语等进行限制。X 到底代表几个汉字,在检索开始时检索者往往没有明确指定。在涉及到带通配符的检索

中,用户输入的检索式都是中英文混合串,不会是纯英文字母串。如果是纯英文字母串,则意味着检索者不知道要检索什么。

### 2.4 关于词性的检索

汉语研究常涉及对词性搭配的检索。文献[11]研究“V 一把”,文献[12]研究“别 V 着”,其中 V 代表动词。“V 一把”表示是动词与“一把”的搭配,因此“他一把抓住你”中的“他一把”就不是用户期望的检索结果。文献[13]研究“A 不到哪里去”,其中 A 代表形容词。文献[14]研究“N 们”,其中 N 代表名词。既然检索式涉及词性,对语料进行分词和词性标注就必不可少。这就让检索从关于汉字串的检索,升级到对语言学单位的检索。

目前应用最广的北京大学 CCL 中文语料库是以汉字为基本单位进行检索,不对语料进行分词处理<sup>[15]</sup>,不提供关于词性的检索服务。“汉语检索通”采用分词软件对生语料自动分词,具有词串检索等功能,但它未提供支持汉语词性搭配检索功能<sup>[16]</sup>。汉语研究者不得不先检索出包含“一把”的语料,然后人工判断“一把”前是否为动词,再从中挑选出合适的语料,这种语料获取方法很慢。近年来汉语研究者对支持词性搭配的检索需求越来越强烈。

Cici 具有以上 4 种面向语言学研究的语料库检索功能,以下将逐一讨论实现这 4 种检索功能的技术细节。

## 3 索引机制与汉字串频次检索

N-gram 汉字串和 N-gram 汉字串频次是计算语言学中的重要概念,在基于概率的分词、词性标注等领域都被广为使用,以下先举例说明这两个概念。对于中文句子“在钓鱼岛上钓鱼”,它的 N-gram 频次统计结果如下:七字串“在钓鱼岛上钓鱼”1 次;六字串“在钓鱼岛上钓”、“钓鱼岛上钓鱼”各 1 次;五字串“在钓鱼岛上”、“钓鱼岛上钓”、“鱼岛上钓鱼”各 1 次;四字串“在钓鱼岛”、“钓鱼岛上”、“鱼岛上钓”、“岛上钓鱼”各 1 次;三字串“在钓鱼”、“钓鱼岛”、“鱼岛上”、“岛上钓”、“上钓鱼”各 1 次;二字串“在钓”、“鱼岛”、“岛上”、“上钓”各 1 次,“钓鱼”2 次;单字串“在”、“岛”、“上”各 1 次,“钓”、“鱼”各 2 次。N-gram 汉字串频次统计既不是汉语词统计,也不是短语统计。

在 Cici 开发初期,汉语研究者提出的软件需求是发现语料库中的高频汉字串分布规律。因此,Cici 需要一次性把语料库中不同长度的 N-gram( $N=1,2,3,\dots,10$ )汉字串频次信息统计出来。对 1.08GB 的中文语料文本,Cici 在一台笔记本电脑(配置如下:4 核 AMD A6 处理器、8GB 内存、Win7 操作系统)上花了近 11 个小时才完成统计,频次统计结果占 16GB。表 1 列出了 1.08GB 中文语料库,长度为 2-10 的汉字串数量和存储该统计结果的 TXT 文本文件大小。表 1 中的“字串数量”是指不同的汉字串数量。例如“在钓鱼岛上钓鱼”中包含 5 个不同的二字串,虽然“钓鱼”在其中出现 2 次,但只能算是 1 个二字串。从表 1 可知:检索一个六字串,最坏情况下需要遍历 2.82GB 的 TXT 文件,并执行 180725570 次六字串匹配计算。

如何存储和索引汉字串频次统计结果对提高汉字串频次检索速度十分关键。根据汉字串频次分布极不均衡的状况,文献[3]按频次大小将汉字串频次信息分块存储,该方案有利于快速检索出高频汉字串。随着 Cici 的广泛应用,汉语研究者又提出要快速查询任意汉字串的频次。按频次排序存储汉

字串统计结果,对低频汉字串特别是没出现过的汉字串的检索速度慢。因此,有必要按汉字串的 Unicode 编码来存储、索引汉字串频次信息。

表 1 N-gram 串统计结果信息

字串长度	2	3	4
字串数量	4029779	43697074	116935223
文件大小	33MB	431MB	1.37GB
字串长度	5	6	7
字串数量	169111110	180725570	166956581
文件大小	2.31GB	2.82GB	2.94GB
字串长度	8	9	10
字串数量	144484470	120936642	99518860
文件大小	2.82GB	2.59GB	2.33GB

汉字串频次统计结果的存储与索引有以下可能的方案:第一,仅保存按 Unicode 编码排序的汉字串频次统计结果,并根据汉字串 Unicode 编码来建立索引;第二,仅保存按频次大小排序的汉字串频次统计结果,根据频次大小建立索引;第三,同时保持按频次排序、按 Unicode 编码排序的汉字串频次信息,并分别建立索引。只保持一份汉字串频次统计信息显然更节省存储空间。但在 Cici 应用中发现,汉语研究者做非确定性汉字串检索和确定性汉字串检索都很频繁。为了使两类检索的响应速度都很快,Cici 采用牺牲存储空间来提高检索速度的策略,使用图 1 所示的存储、索引机制。

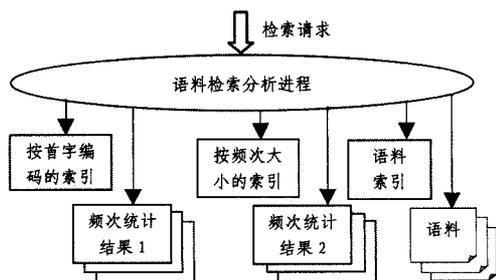


图 1 Cici 中的索引与数据存储机制

“语料检索分析进程”先分析用户提交的每一个检索请求,根据检索请求类型查询合适的索引表。如果是查询确定性汉字串的生语料文本,则读取“语料索引”,最终访问生语料;如果是查询首字确定的汉字串,则先访问“按首字编码的索引”,再访问 N-gram 汉字串“频次统计结果 1”;如果是查询首字不确定的汉字串,则读取“按频次大小的索引”,再访问 N-gram 汉字串“频次统计结果 2”。

图 1 中的“语料索引”是指对生语料的索引。基于布尔模型的中文文本倒排序索引已被研究得很透彻,“语料索引”的实现方法不在此讨论。“频次统计结果 2”是根据汉字串频次从大到小存储,其存储、索引方法在文献[3]中已有详细介绍。“频次统计结果 1”是根据汉字串 Unicode 编码从小到大为序存储,例如以“一”为首字符的汉字串及其频次信息存储在一起。以下仅讨论按 Unicode 编码次序的 N-gram 汉字串频次存储与索引。

按汉字串的 Unicode 编码次序存储汉字串频次统计信息,有必要先考察不同首字所对应的汉字串数量。语料库中不同汉字的频次分布极不均匀,图 2 显示了 1.08GB 语料库中 9193 个不同的汉字的频次分布特征,其中横坐标是汉字频次的对数值( $\lg frequency$ ),纵坐标是同一频次汉字数量的对数值( $\log num$ )。在 9193 个汉字中,有 1100 个汉字只出现了一次,而“的”、“一”、“了”、“是”、“不”等 10 个最高频汉字的频次之和却占总字数的 14.985%。汉字频次分布不均必定导

致以同一个汉字为首字的汉字串数量也极不均衡。

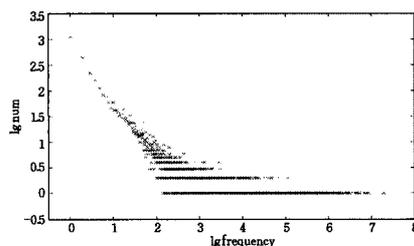


图 2 语料中的汉字频次分布

汉字在句子中所处位置决定了它能否成为 N-gram 串汉字首字。例如某汉字仅在语料库中出现一次,且位于句子的最后位置,它就不可能成为二字串、三字串的首字符。1.08GB 的语料库中出现了 9193 个不同的汉字,但作为三字串首字符的却只有 8576 个汉字。随着汉字串串长增大,首字数量不断变少。图 3 是语料库中以不同汉字为首字的三字串数量。语料库中的汉字分布在 Unicode 编码的 CJK 区(19968~40869 之间),因此图 3 的横坐标是 19968~40869 之间的整数;纵坐标则是以该汉字为首字的三字串数量的对数值( $\log num$ )。从图 3 可知,不同汉字所对应的三字串数量差异极为显著。如果简单地根据汉字串首字来分别存储三字串频次信息,需要建立 8576 个文件,显然存在文件太多、文件大小相差悬殊等缺陷。

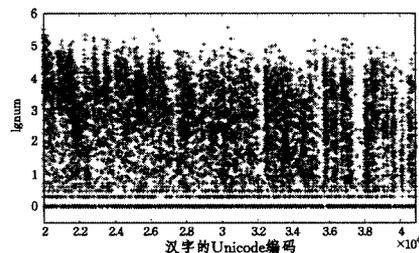


图 3 不同首字所对应的三字串数量

表 2 列出了 1.08GB 语料库中最高频的 10 个汉字,以及拥有最多三字串的 10 个首字。以“的”、“一”、“是”等为首字的不同三字串超过 17 万个;而有 1021 个汉字作为首字的三字串只有 1 个。比较三字串的首字序列与最高频汉字序列,可发现二者也不一致。不同长度汉字串的首字所对应汉字串数不一致,首字排序也不一致。因此,对不同长度汉字串频次信息,需独立地按首字建立倒排序索引。

表 2 最高频的汉字和三字串首字

排序	单个汉字		三字串	
	汉字	频次	首字	三字串数
1	的	18869684	的	367859
2	一	8622526	一	333204
3	了	8043890	是	306423
4	是	7299889	不	222379
5	不	6783174	了	218472
6	我	5185127	在	212837
7	在	4758181	人	203185
8	人	4543240	有	200660
9	有	4449010	大	180660
10	这	4329332	上	179682

Cici 先把全体同一长度的汉字串频次信息按汉字串的 Unicode 编码排序。然后,把首字所对应汉字串数量超过 10 万个(如表 2 中“的”、“一”、“是”等)的汉字串频次统计信息单独存在一个文本文件中,对首字所对应汉字串没超过 10 万个的,则将多个相邻的汉字串频次信息合并存储在一个频次统

计结果文件中,且按每个文件约存储 10 万个汉字串频次信息规模来划分。为了避免以同一个汉字为首字符的汉字串分布在两个频次统计结果文件中,有些频次信息文件会略多于 10 万个汉字串,有些会略少于 10 万个。使用该方法,N-gram 汉字串频次统计信息文件数量一定少于汉字串首字数量,且除了少数文件较大以外,大多数文件的大小比较均衡(都是存储约 10 万个汉字串频次信息)。当然,当语料库规模继续增大时,可以对这一标准(10 万个)做调整。

“按 Unicode 编码排序的三字串频次信息索引表”结构如表 3 所列。其中第一列的“首字”是指三字串中的第一个汉字,表 3 按首字的 Unicode 编码从小到大排序。汉字串频次统计结果文件名的编排规则是“首字-n”,n 指汉字串长度,例如对全体以“一”为首个汉字的三字串为“一-3.txt”,以“一”开始的四字串则为“一-4.txt”。“开始位置”和“结束位置”是该首字所对应的第一个汉字串和最后一个汉字串在频次统计结果文件内部所对应的位置。检索进程根据“开始位置”和“结束位置”信息,可以快速地利用折半查找算法来查询汉字串。折半查找时间复杂度为  $\log_2 m$ (m 为以该汉字为首字符的汉字串的总数),即使是对以“的”为首字符的三字串(367859 个不同的三字串)进行查找,所需执行的汉字串匹配操作也不超出 20 次。实践表明,表 3 所描述的频次信息存储和索引方式简单、高效。

表 3 按 Unicode 编码排序的三字串频次信息索引表

首字	文件名	开始位置	终止位置
一	一-3	0	333203
丁	丁-3	0	13536
七	七-3	13537	37610
上	上-3	37611	37641
...	...	...	...

## 4 汉字串重叠式检索

鉴于汉语重叠构式研究近 30 年来经久不衰,Cici 提供两种类型的重叠式检索:第一,纯粹用英文字符来表示的重叠式,例如“AABB”;第二,中英文混合重叠式,例如“白 AA”。以下分别讨论其实现方法。

### 4.1 纯英文字符表示的重叠式检索

汉语词主要是单字词(不存在重叠式)、双字词、三字词、四字词,因此对汉字串重叠式的检索主要是对二字串、三字串和四字串的检索。根据排列组合理论可知,二字串、三字串、四字串一共存在表 4 所列的 19 种重叠类型,其中 A、B、C 代表不同的汉字。表 4 列出在 1.08GB 语料库中,二字串、三字串、四字串所包含特定重叠式的汉字串数量,以及在整个 N-gram 串中所占比率(即满足该重叠式的汉字串数量除以该长度的全体汉字串数量)。表 4 中的“比率”是用其中的“数量”和表 1 所列汉字串数量做除法所得。例如,语料库包含 116935223 个不同的四字串,有 15486 个四字串满足 AABB 类型,故 AABB 类型的四字串比率为  $15486/116935223 = 0.01\%$ 。

从表 4 所列的比率可知,重叠式汉字串数量远少于该长度的全体汉字串数量。鉴于汉字串重叠式检索是高频操作,因此可以把满足特定重叠式的汉字串预先从 N-gram 频次统计信息表中抽取出来,存放在特定的重叠式汉字串频次表中。重叠式汉字串表生成过程如下:第一,访问图 1 所示按汉字串

Unicode 编码排序存储的“统计结果 1”;第二,读取二字串频次统计结果,把满足 AA 重叠格式的串挑选出来,并把它们存储到 AA 重叠式汉字串表中;第三,读取三字串频次统计结果,挑选出满足 AAA、AAB、ABB、ABA 重叠格式的三字串,分别将这些三字串及其频次信息存到 4 类重叠格式汉字串表中;第四,读取四字串频次统计结果,挑选出满足 AAAA、AAAB 等重叠格式的四字串,将这些四字串和频次信息分别存到这 14 类重叠格式汉字串表中。

表 4 二字串、三字串、四字串重叠式比率

类型	AA	AAA	AAB	ABB	ABA
数量	4219	1078	257317	272572	114773
比率	0.10%	0.00%	0.59%	0.62%	0.26%
类型	AAAA	AAAB	AABA	ABAA	ABBB
数量	490	3329	1482	1482	4349
比率	0.00%	0.00%	0.00%	0.00%	0.00%
类型	ABAB	AABB	ABBA	AABC	ABAC
数量	12510	15486	2366	822675	324189
比率	0.01%	0.01%	0.00%	0.70%	0.28%
类型	ABCA	ABBC	ABCB	ABCC	
数量	355314	881156	422025	816633	
比率	0.30%	0.75%	0.36%	0.70%	

有了以上 19 类重叠格式汉字串表,遇到用户提交的关于二字串、三字串、四字串重叠式检索请求,无需再访问图 1 中数据量巨大的“统计结果 1”,而是直接检索数据量较小的重叠式汉字串表,从而重叠式检索效率会得到显著提高。值得注意的是,对于表 4 中所列的 19 种重叠式,汉语研究者的关注程度不一致。例如对四字串重叠式研究中,他们对 AABB、ABAB 高度重视,对 AABC、ABCA 等不大重视。如果想节省存储空间,对不受关注的重叠式也可以不预先计算。

Cici 不仅能高效支持二字串、三字串、四字串的重叠式检索,还支持五字串以及更长串重叠式检索,但检索速度要慢一点。当用户提交的重叠式长度为 5 或大于 5 个汉字时,Cici 则即时检索图 1 中的“频次统计结果 2”。其检索过程为数秒种,比针对二字串、三字串、四字串的重叠式检索速度慢许多。

### 4.2 中英文混合重叠式检索

有些汉语研究者在查询语料库前,已有比较细致的想法,从而输入中英文混合的汉字串重叠式。例如要讨论“白”与其他汉字搭配的三字串重叠式,可输入“白 AA”,会得到“白花”、“白胖胖”等数十个汉字串。从表 4 可知,如果输入“ABB”,Cici 会反馈二十七万多个检索结果,用户不得不耗费大量时间从这二十七万多条检索结果中去发现以“白”为首字的三字串。

获得用户提交的中英文混合重叠式后,Cici 先分析该检索请求在表 4 中所对应的重叠模式;然后读取本文 4.1 节预先生成的重叠式汉字串表,进行汉字串匹配计算;最后,将满足要求的汉字串按频次大小排序后反馈给用户。同一重叠类型的汉字串频次差异大。以 AABB 类型四字串为例,在 15486 个 AABB 类型的汉字串中,90%的频次低于 10 次,仅 1%的频次高于 100 次。高频汉字串通常是最受关注的对象,有些频次过低的汉字串,如“三三俩俩”、“三三二二”等是因误用或方言原因所致,缺乏研究价值。Cici 允许用户自定义反馈结果的最低频率,例如不低于 100;以及最大反馈结果数,如最高频的 200 条重叠串等。这些设置功能,提高了汉语言研究者的重叠式检索效率。

需要指出的是:从语料中提取的 N-gram 汉字串未必是

语言学家期望的重叠串,例如从“武汉大学学生生活丰富多彩”一句中提取的四字串“学学生生”其实不是正确的检索结果。由于汉语研究者会检查所引用的语料,这类错误最终会被排除,并不会对语言研究带来负面影响。

## 5 带通配符的汉字串检索

汉语语法研究中,“X”有时是词,有时是短语,有时是非词非短语,因此严格来说,X表示任意汉字串。但从研究“X以上”的文献[9]所列语料来看,X是词的比率最高。文献[10]研究“非X不可”格式的历史演进和语法化,该文所列语料中X大多是词。因此,根据汉语词长的规律来设置X对应汉字串的长度最合适。汉语中,单字词、二字词、三字词和四字词占绝大多数,因此Cici默认的X长度在1-4之间。Cici也允许用户指定“X”所对应汉字串的长度。

默认情况下,Cici自动将用户输入的“X”转化为\*、\*、\*、\*、\*、\*、\*、\*这4种通配符(\*表示一个任意的汉字),然后分别对不同长度的N-gram频次信息进行检索。例如将“非X不可”转变为4个模糊查询:“非\*不可”、“非\*\*不可”、“非\*\*\*不可”、“非\*\*\*\*不可”;然后分别在四字串、五字串、六字串和七字串中进行检索;最终将4种不同长度N-gram串频次检索结果合并,反馈给用户。由于现代汉语词典(第5版)收录汉语词平均长度为1.9个汉字<sup>[17]</sup>,Cici优先把“X”为单字或双字的高频检索结果反馈给用户。“非X不可”与“X以上”两个检索式在执行过程中存在差异,其原因与X所在位置有关。“X以上”中的X位于串首位置,“非X不可”中的首字符却是确定的汉字。

对于首字符确定的带通配符的汉字串检索,Cici直接以以汉字串Unicode编码索引中去读取首字所对应的N-gram串频次统计结果文件,然后在首字所对应的汉字串的第一个和最后一个中查找。以“非X不可”的检索过程为例,先查询如表3所列的索引表,找到以“非”为首字的汉字串频次信息统计结果文件,然后对全体以“非”为首字的四字串、五字串、六字串和七字串执行汉字串匹配操作。

当通配符X位于检索式的串首位置时,Cici不得不查询全体N-gram串频次统计结果。以对“X以上”的检索过程为例,由表1可知,1.08GB的语料库包含43697074个不同的三字串和116935223个不同的四字串。由于通配符位于串首位置,当X长度设置为1时,则必须做43697074次三字串配置操作;当X长度设置为2时,则必须做116935223次四字串配置操作。不同的三字串、四字串数量众多,因此Cici根据汉字串频次索引,优先查询、反馈高频汉字串频次信息。Cici还允许用户限定最低频次或反馈结果条数,避免对数量极大但频次却极低的N-gram汉字串进行检索。

在实现带通配符的检索中,我们观察到一个很有趣的事实:其实很多汉语研究者并不十分清楚自己的检索需求。以“X以上”为例,有些研究生认为X仅代表词,有些研究生认为X是词或短语,但对非词非短语的情况几乎无人指出。他们看到文献[9]考虑了非词非短语的语例时,才意识到这也是该考虑的。因此,Cici将X等价于非定长汉字串的做法,对汉语研究的帮助起到了意想不到的效果。中文语料库检索软件开发者在做需求分析时,不仅要听取汉语研究者的意见,还得亲自阅读一些汉语研究论文,这样才会更清楚汉语研究者的检索意图,发现被他们忽视的检索需求。

## 6 涉及词性的检索

文献[11-14]的研究都期望语料库检索系统能支持对汉语词性搭配的检索。汉语研究中常用词性英文单词的首字母或前几个字母表示词性,例如V表示动词,N表示名词,A或Adj.表示形容词。虽然语料库语言学研究者一直呼吁尽早构建支持词汇搭配、词性搭配查询的语料库,但至今为止我们都没发现GB规模、支持词性搭配的中文语料库检索系统。以下两个因素可能是造成这一现象的关键原因。

第一,要支持词性搭配检索,就必须执行自动分词、词性标注处理。以往的中文熟语料库开发大多采用“机标人助”的方式来开发<sup>[18]</sup>,即先使用自动分词和自动词性标注软件处理生文本,然后再组织汉语领域的专家对标注结果进行人工校对。当语料规模仅为几百万字时,投入一定的经费和人力尚可执行;但对GB规模生语料的自动标注结果进行人工校对,仅凭数额极低的语料库开发经费是不可能实现的。

第二,如何索引GB规模语料库的词性标注结果。汉语中的词性不超过20种,每篇语料一般都超过1500字,几乎各种词性都会在语料中出现。汉语研究者对名词、动词、形容词、副词等的搭配最感兴趣,汉语中的句子大多是主谓句,一般都包括名词、动词、形容词、副词等。如果直接以词性为特征进行索引,无论是对每篇语料进行索引,还是对每个句子进行索引,都存在输入一个词性搭配检索式,输出极多检索结果的现象,用户还是得耗费大量时间进行手工筛查,检索效率低。

既不能人工校对GB规模中文语料的词性标注结果,又很难建立关于词性的索引,因此不得不寻求其他方法。观察文献[11-14],可以注意到一个有价值的规律:检索式“V一把”、“别V着”、“A不到哪里去”、“N们”中都包含了确定的汉字。只要有确定的汉字,就可以快速检索语料库的N-gram串频次统计信息。汉字串未必是词,但词一定是汉字串。因此,可以先把包含检索式中汉字的汉字串找出来;然后再判断词性标识符所对应位置的汉字串是否是词;若是一个词,判定其词性是否与检索式中的词性要求一致;如果一致,则将该汉字串作为备选串供汉语研究者选择。根据以上思路,Cici提出如图4所示的算法实现了支持词性搭配的中文语料库检索。

- 步骤1:将检索式中的词性标识符用通配符X代替;
- 步骤2:执行带通配符的汉字串检索,获得包括用户输入检索式中确定汉字的备选串;
- 步骤3:对备选串即时做自动分词和词性标注,把不符合词性要求的N-gram串过滤,把符合要求的汉字串反馈给用户作为有效串;
- 步骤4:用户选择有效串,再读取包含有效串的语料文本,仅把包含该有效串的句子进行词性标注,过滤词性不符合检索要求的结果;
- 步骤5:把全体符合要求的生语料反馈给用户。

图4 涉及词性的中文语料检索算法

在带通配符的汉字串检索中,通配符X在汉字串中的位置影响检索实现策略和查询响应时间。在关于词性搭配的检索中,词性标识符位置还会对检索结果的正确性产生直接影响。不妨对比“别V着”和“V一把”两个词性搭配检索操作,二者都是对动词搭配的检索,但因V位置不同,二者的检索过程有差异。

对于“别V着”这种词性标识符V居中,首、尾是确定性汉字的检索式,X为不同长度时所得到的备选串不会重复。

对“V一把”这种词性标识符居串首或居串末的检索,当X的长度设置为1、2、3、4时,可能出现检索结果冗余。不妨观察以下特例:假设语料中存在4个句子:“耍一把”、“我要一把”、“给我耍一把”、“你给我耍一把”。X长度为1时,备选串“耍一把”频次为4;X长度为2时,“我要一把”频次为3;X长度为3时,“给我耍一把”频次为2;X长度设置为“4”时,“你给我耍一把”频次为1。当X长度为1时,语料库检索系统已经分析了全部4个句子,因此不必分析X长度为2、3、4的情况。由此例可知,当词性标识符位于检索式的串首或串尾位置时,反馈检索结果时需要排除冗余。而该现象在对“别V着”类的检索时则不会出现。

表5列出了利用图4所示算法,在1.08GB语料库中分别检索“V一把”和“别V着”,关于步骤2产生的“备选串”、步骤3产生的“有效串”的数据特征。表5中“备选串数量”、“备选串总频次”两项是图4中步骤2产生的备选汉字串的特征;“有效串数量”、“有效串总频次”两项是步骤3产生的有效串汉字串的特征。对比“有效串数量”和“备选串数量”,发现经过步骤3对备选串自动分词和词性标注、词性匹配筛选后,80%以上的备选串被过滤了。这样用户看到的有效串较少,但正确率却提高了。总频次是语料库中包含的这些汉字串总数,对比“备选串总频次”和“有效串总频次”可知,需要从生语料中找到相关句子做词性标注、词性匹配检测的工作量明显降低,从而提高了检索效率。

表5 涉及动词检索的数值特征示例

检索式	特征	单字	双字	三字	四字
V一把	备选串数量	1867	12042	22462	24937
	备选串总频次	49151	45131	37092	30968
	有效串数量	633	673	15	22
	有效串总频次	17645	4989	53	33
别V着	备选串数量	323	664	578	608
	备选串总频次	2776	1578	854	703
	有效串数量	264	167	0	0
	有效串总频次	2346	375	0	0

表5中的单字、双字、三字、四字是指词性V所代表汉字串的长度。例如“合作一把”中的“合作”为双字。从表5可知:当词性标识符V对应为三字串或四字串时,备选串数量很多,但有效串很少。V对应的汉字串长度为1或2时,则需要重点分析。汉语研究者只需要从数十个或数百个有效串中选择自己感兴趣的汉字串,Cici再读取生语料文本中的句子做进一步分析。表5中有效串的正确率对最终检索结果的查全率和查准率影响较大。ICTCLAS是著名的自动分词和词性标注软件,其正确率超过90%<sup>[19]</sup>,Cici使用它自动对步骤2产生的备选串进行分词和词性标注。

对表5中产生的有效串,我们进行了人工判断,结果如下。对单字或二字串,ICTCLAS的词性标注准确率超过90%,但对三字串和四字串则错误率较高。二字串只有少量错误,例如“被高举着”中的动词短语“高举”被标注为动词。但当V是三字串或四字串时,错误率很高。有些错误甚至让人费解,例如把“他妈的一把年纪”中的“他妈的”标注为动词。汉语研究者在观察有效串时,会运用汉语知识进行分析,错误的有效串并不会被选中。由于GB规模的语料库能为用户提供较多关于词性搭配的检索结果,例如对“V一把”的检索,表5中步骤2产生了16万多个(V分别为单字、双字、三字、四字串时)备选串,不可能人工逐一判断,因此对涉及词性搭配

的检索,查准率比查全率重要。

表6列出了利用图4所示算法,在1.08GB语料库中分别检索“A不到哪里去”和“N们”的数据特征。由于“不到哪里去”是一个确定性的5字串,它使得满足条件的备选串很少。相反,“们”只有一个字,备选串很多。要想提高检索速度,用户应尽量输入较长的确定性汉字串。表5和表6还展示:无论是动词、形容词还是名词,能找到的三字词或四字词有效串都很少。例如“N们”有488840个4字备选串,但实际上只有“共产党员”、“共青团员”两个有效串,有效串数量不足备选串数量的十万分之一。本文仅示例了动词、形容词、名词检索,实际上只要是ICTCLAS标注了的词性,Cici都支持相关词性检索。

表6 涉及形容词、名词检索的数值特征示例

检索式	特征	单字	双字	三字	四字
A不到哪里去	备选串数量	101	222	453	504
	备选串总频次	754	736	694	620
	有效串数量	32	24	0	0
	有效串总频次	498	48	0	0
N们	备选串数量	1607	21485	188181	488840
	备选串总频次	2512490	1570267	1350210	1007638
	有效串数量	673	2985	951	2
	有效串总频次	236582	206790	13078	22

目前我们仅知道“工智检索通”系统提供了涉及词性的检索功能,但我们没查阅到关于它如何实现词性检索功能的论文,也不能获得它的语料库。两个检索系统分别对不同的语料库进行检索,很难比较其查全率。尽管可以通过输入相同的检索式在两个系统中进行检索,计算各自的查准率,但是每个检索结果正确与否,需要汉语研究者阅读原始文本才能做出判断。从表5和表6所列的“有效串总频次”数量来看,即使是判断一个检索式的全部结果,也可能要耗费许多时间。例如,由表5可知,检索“V一把”可获得22720个(1-4字有效串频次之和)有效语料。这意味着要阅读两万多个句子,才能计算出关于这一检索请求的准确率。客观比较两个检索系统的查准率,必须测试大量涉及不同词性的检索式。以我们目前所拥有的实验条件,难以对数万规模的句子进行人工判断。鉴于以上原因,此处无法定量比较Cici与同类系统在涉及词性检索的效率差异。

**结束语** 本文根据中文语料库统计与检索系统Cici的开发和应用经验,讨论了大型中文语料库检索技术。根据汉语研究者的思考过程,Cici灵活运用汉字N-gram串频次统计结果,把中文语料库检索划分为先检索N-gram串频次统计结果,再让汉语研究者自主选择检索词,最后做确定性汉字串生语料检索。该方式既提高了检索速度,也提高了语料检索的准确率,还降低了检索软件开发难度。

中文语料库统计与检索系统Cici将在以下两方面继续改进:第一,以互联网网站方式向用户提供开放的中文语料检索服务。目前Cici是单机版软件,需用户自主操作,而多数汉语研究者操作软件能力比较低。汉语研究者已在因特网上检索北大语料库CCL多年,很熟悉网络语料库检索习惯,因此Cici应尽快建成在线语料库检索系统。第二,把语料规模扩展到5GB,并优化各类语料比率,使其来源更广,从而使检索结果能得到更多汉语研究者的支持。

## 参考文献

[1] Ruslan M. The Oxford Handbook of Computational Linguistics

[M], Beijing: Foreign Language Teaching and Research Press, 2009

[2] 邱晗,周强. 自动获取大规模的汉语紧密组合词汇关联对[J]. 清华大学学报:自然科学版,2011(9):28-33

[3] 余一骄,刘芹. 面向超大规模的中文文本 N-gram 串统计[J]. 计算机科学,2014(4):263-268

[4] 罗瑜昕. 用统计的方法看“京派”与“海派”小说语言风格差异[J]. 现代语文:学术综合版,2012(4):137-141

[5] 陈功. 语料库检索的模式、问题及启示[J]. 当代外语研究,2011(10):10-14

[6] 任海波. 现代汉语 AABB 重叠式词构成基础的统计分析[J]. 中国语文,2001(4):302-308

[7] 崔四行. 从 ABAB、AABB 重音模式的句法功能看汉语的韵律形态[J]. 语言教学与研究,2012(5):63-69

[8] 蒋向勇,白解红. 汉语 ABB 式网络重叠词语的认知解读[J]. 外语研究,2013(3):30-34

[9] 邢福义. “X 以上”格式在现代汉语中的演进[J]. 语言研究,2010(1):1-10

[10] 洪涛,董正存. “非 X 不可”的历史演化和语法化[J]. 中国语文,

2004(3):253-261

[11] 邵敬敏. 说“V 一把”中 V 的泛化与“一把”的词汇化[J]. 中国语文,2007(1):14-19

[12] 李广瑜. 否定祈使句式“别 V 着”刍议[J]. 语言教学与研究,2013(1):48-55

[13] 吴为善,夏芳芳. “A 不到哪里去”的构式解析、话语功能及其成因[J]. 中国语文,2011(4):326-333

[14] 张谊生. “N”+“们”的选择限制与“N 们”的表义功用[J]. 中国语文,2001(3):201-211

[15] <http://ccl.pku.edu.cn:8080>

[16] <http://democlip.blcu.edu.cn:800>

[17] 王惠. 词义·词长·词频—《现代汉语词典》(第 5 版)多义词计量分析[J]. 中国语文,2009(5):120-130

[18] 张宝林. 汉语中介语语料库建设的现状与对策[J]. 语言文字应用,2010(3):129-138

[19] Zhang Hua-ping, Yu Hong-kui, Xiong De-yi, et al. HHMM-based Chinese Lexical Analyzer ICTCLAS [C]//Proceedings of 2nd SIGHAN Workshop Affiliated with 41st ACL, 2003:184-187

(上接第 197 页)

6 种应用程序的实验表明,本文方法可以有效降低磁盘能耗。在下一步研究工作中,我们将研究多应用背景下的磁盘空闲周期增加问题。

## 参 考 文 献

[1] 易会战,罗兆成. 基于动态电压调节的高性能业务系统能耗优化[J]. 华中科技大学学报:自然科学版,2013(1):25-29

[2] 张凯,陈书明,王耀华,等. 面向通用 HPC 的高性能 DSP 设计权衡[J]. 计算机学报,2013,36(4):790-798

[3] 张帅,宋凤龙,王栋,等. 多核结构片上网络性能-能耗分析及优化方法[J]. 计算机学报,2013,36(5):988-1003

[4] Hadjipaschalis I, Poullikkas A, Efthimiou V. Overview of current and future energy storage technologies for electric power applications[J]. Renewable and Sustainable Energy Reviews, 2009,13(6):1513-1522

[5] Li K, Kumpf R, Horton P, et al. A quantitative analysis of disk drive power management in portable computers[C]//USENIX winter, 2002:279-291

[6] Gurumurthi S, Sivasubramaniam A, Kandemir M, et al. DRPM: dynamic speed control for power management in server class disks[C]//30th Annual International Symposium on Computer Architecture, 2003. IEEE, 2003:169-179

[7] Son S W, Chen G, Kandemir M, et al. Exposing disk layout to compiler for reducing energy consumption of parallel disk based systems[C]//Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming. ACM, 2005:174-185

[8] Son S W, Kandemir M, Choudhary A. Software-directed disk power management for scientific applications[C]//19th IEEE International Parallel and Distributed Processing Symposium, 2005. IEEE, 2005:4-13

[9] 李元章,孙志卓,马忠梅,等. S-RAID5:一种适用于顺序数据访问的节能磁盘阵列[J]. 计算机学报,2013,36(6):1290-1302

[10] 刘靖宇,郑军,李元章,等. 混合 S-RAID:一种适于连续数据存储的节能数据布局[J]. 计算机研究与发展,2013,50(1):37-48

[11] 孟涛,刘浩,胡宏扬. 基于预读策略的节能数据访问技术[J]. 计

算机工程,2012,38(8):44-46

[12] 李建敦,彭俊杰,张武. 云存储中一种基于布局的虚拟磁盘节能调度方法[J]. 电子学报,2012,40(11):2247-2254

[13] Weissel A, Beutel B, Bellosa F. Cooperative I/O: A novel I/O semantics for energy-aware applications [J]. ACM SIGOPS Operating Systems Review, 2002,36(SI):117-129

[14] Papathanasiou A E, Scott M L. Energy efficient prefetching and caching[C]//Proceedings of the 2004 USENIX Annual Technical Conference, Berkeley, CA, USA, 2004:255-268

[15] Pinheiro E, Bianchini R. Energy conservation techniques for disk array-based servers[C]//Proceedings of the 18th annual international conference on Supercomputing. ACM, 2004:68-78

[16] Son S W, Kandemir M, Choudhary A. Software-directed disk power management for scientific applications[C]//19th IEEE International Parallel and Distributed Processing Symposium, 2005. IEEE, 2005:4b-4b

[17] Son S W, Kandemir M. Energy-aware data prefetching for multi-speed disks[C]//Proceedings of the 3rd Conference on Computing Frontiers. ACM, 2006:105-114

[18] Thakur R, Groppe W, Lusk E. Data sieving and collective I/O in ROMIO[C]//The Seventh Symposium on the Frontiers of Massively Parallel Computation, 1999 (Frontiers '99). IEEE, 1999:182-189

[19] Liao W, Coloma K, Choudhary A, et al. Collective caching: application-aware client-side file caching[C]//14th IEEE International Symposium on High Performance Distributed Computing, 2005 (HPDC-14). IEEE, 2005:81-90

[20] Calder P C, Jacobsen C, Skall Nielsen N, et al. Nutritional benefits of omega-3 fatty acids[M]//Food enrichment with omega-3 fatty acids. 2013:3-26

[21] Whaley J. Joeq: A virtual machine and compiler infrastructure [C]//Proceedings of the 2003 Workshop on Interpreters, Virtual Machines and Emulators. ACM, 2003:58-66

[22] Palmer A J, Brandt A, Gozzoli V, et al. Outline of a diabetes disease management model: principles and applications [J]. Diabetes research and clinical practice, 2000,50:S47-S56

[23] Ligon W, Ross R. PVFS: Parallel virtual file system[M]//Beowulf cluster computing with Linux. MIT Press, 2001:391-429