

关于常用字覆盖率统计算法的研究

阿不都克里木·玉素甫^{1,2}, 杨 琴^{2,3}, 王亮亮¹

(1. 新疆教育学院 现代教育技术中心, 新疆 乌鲁木齐 830043;

2. 新疆教育云技术与资源实验室, 新疆 乌鲁木齐 830043;

3. 新疆教育学院 信息科学与技术学院, 新疆 乌鲁木齐 830043)

摘 要:对常用字在教育资源电子文本中的覆盖率、使用率、字频统计算法进行了研究,并根据算法通过计算机语言开发常用字覆盖率统计分析系统。统计分析系统可以对文本中所使用的常用字进行统计分析,即可以统计常用字覆盖率、文本汉字数、常用字字频、常用字使用率等,并根据统计数据以饼形图的方式显示。为了了解常用字在文本中的覆盖率和使用情况,通过常用字覆盖率统计分析系统对一些电子文本进行了统计分析,并得出相应的结果。结果表明常用字在文本中的覆盖率和使用率相当高,即581个常用字在文本中的覆盖率平均在68.9%以上,1 000个常用字在文本中的覆盖率平均在81.4%以上,2 500个常用字在文本中的覆盖率平均在96%以上,并且常用字在不同统计对象文本中的使用频度也会有所不同。

关键词:常用字;统计算法;覆盖率统计;使用率统计;字频统计

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2020)05-0201-05

doi:10.3969/j.issn.1673-629X.2020.05.038

Research on Statistical Algorithms of Coverage Rate of Commonly Used Chinese Characters

ABUDUKELIMU Yusufu^{1,2}, YANG Qin^{2,3}, WANG Liang-liang¹

(1. Modern Education Technology Center, Xinjiang Education Institute, Urumqi 830043, China;

2. Xinjiang Laboratory of EducationCloud Technology and Resources, Urumqi 830043, China;

3. School of Information Science and Technology, Xinjiang Education Institute, Urumqi 830043, China)

Abstract: The coverage, usage and frequency statistics arithmetic of commonly used Chinese characters in electronic texts has researched. According to the arithmetic, a statistical analysis system for coverage of commonly used Chinese characters has been developed by computer language, which can make statistical analysis of commonly used Chinese characters in text. It can count the coverage rate of commonly used Chinese characters, the number of text Chinese characters, frequency of commonly used Chinese Characters, utilization rate and so on, and display them in a pie chart according to statistical data. In order to understand the coverage and usage of commonly used Chinese characters in texts, some electronic texts are analyzed by the coverage statistical analysis system of commonly used Chinese characters, and the corresponding results are obtained. It is showed that the coverage and usage of commonly used Chinese characters in texts are quite high, that is, the coverage of 581 commonly used Chinese characters in texts averages over 68.9%, that of 1 000 commonly used Chinese characters in texts averages over 81.4%, and that of 2 500 commonly used Chinese characters in texts averages over 96%, and that the frequency results of common Chinese characters used in different statistical object texts are different.

Key words: commonly used Chinese characters; statistical algorithm; coverage statistics; utilization statistics; frequency statistics

0 引 言

常用字是现代汉语中经常用到的字,即字频和使用度最高的字。随着社会的发展,常用字的使用频率也在不断的变化。而常用字最基本的选字原则就是根

据字的使用频度,选取使用频度高的字。除此之外还有其他的选字原则,如:根据字的使用分布选取分布均匀的字,选取构字能力和构词能力强的字,根据汉字的实际使用情况斟酌取舍等。目前计算机选字主要采用

收稿日期:2019-05-17

修回日期:2019-09-18

网络出版时间:2019-12-18

基金项目:新疆维吾尔自治区重点实验室开放课题(2019D04024)

作者简介:阿不都克里木·玉素甫(1980-),男(维吾尔族),工程师,硕士,研究方向为计算机信息处理与软件工程、计算机自然语言处理。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20191218.1110.008.html>

统计字频的方法,并以字频高低逐一排序。字频指的是汉字的使用频率,即某个汉字在抽样统计资料里出现的次数在统计总字数中所占的比例。字频统计对识字教学、字书编纂以及汉字的机械处理和信息处理等都十分重要。1988年1月,国家语委和国家教委联合发布《现代汉语常用字表》^[1],共收录常用字3 500个,其中常用字2 500个,次常用字1 000个。1988年3月发布《现代汉语通用字表》^[1]共收录通用字7 000个(包括《现代汉语常用字表》的3 500字),这两种表都是以字频的高低来排序的。为了了解常用字在文本中的使用情况,以计算机信息处理的方式来获取统计信息,并且本研究作为新疆高校教育资源安全审查信息化系统研究项目的基础研究部分,主要研究了常用字在电子文本中的覆盖率统计,使用率统计和字频统计的数学算法以及计算机程序算法,并根据得出的研究方法研发常用字覆盖率统计分析系统,最后做一个统计实验,即分别通过《现代汉语常用字表》中的频度最高的581个常用字^[2],1 000个常用字和2 500个常用字对电子文本进行统计分析,并获取覆盖率、使用率、字频统计信息,以此了解文本中常用字的使用情况。

1 覆盖率统计算法的优化

覆盖率统计的主要任务就是统计出给定文本中常用字的覆盖情况,根据统计信息结果就可以知道常用字在文本中的覆盖率或者说是比率。为此在前期研究中^[3]首先将电子文本中非汉字元素取出后,再对所剩下的汉字元素进行统计分析。但是在计算机处理中该方法还不是很实用。因为为了先抽取文本中除了汉字以外的元素,对于计算机来说需要先定义大量字符元素,以便计算机可以识别并分类。如:数字、各种符号以及其他未知符号等。这对实现计算机程序算法带来了一些困难,也有可能由于程序无法识别文字字符产生统计误差等问题。因此对前期所使用的数学公式进行优化处理,以便适用于计算机程序算法^[4-6]的实现。

1.1 覆盖率统计数学算法

覆盖率是阅读教材里被包含的在字表里的汉字与阅读教材里的全部汉字的比率^[7]。在优化后的算法中不再对文本中的非汉字字符进行统计和抽取操作,而是直接对文本中的汉字字符^[8]进行统计,这也更符合计算机的处理。具体数学表达式如式(1)所示。

$$\text{常用字覆盖率} = \frac{\text{出现次数}}{\text{文本长度}} \times 100\% = \frac{F}{L} \times 100\% = \frac{\sum_{i=1}^N \sum_{j=1}^L X(C_i, T_j)}{L} \times 100\% \quad (1)$$

其中, F 为常用字在电子文本中的出现次数, L 为文本

的长度; C 为常用字, C_i 为常用字表中下标为 i 的汉字, N 为常用字字数; T 为电子文本, T_j 为电子文本中下标为 j 的汉字。出现次数 F 主要是通过常用字和电子文本中的汉字逐一对比后获取的统计结果,即当 $C_i = T_j$ 时, $X(C_i, T_j) = 1$, 当 $C_i \neq T_j$ 时, $X(C_i, T_j) = 0$, X 函数的值将会累计计算,运算结束后作为 F 的值。

1.2 覆盖率统计程序算法的实现

(1) 程序处理流程。

根据式(1)可以通过计算机程序来实现覆盖率统计。首先将程序处理流程定义如下:

第一步:统计出文本中汉字的个数 L 。

第二步:统计常用字在文本中的出现次数 F , 具体流程如图1所示。

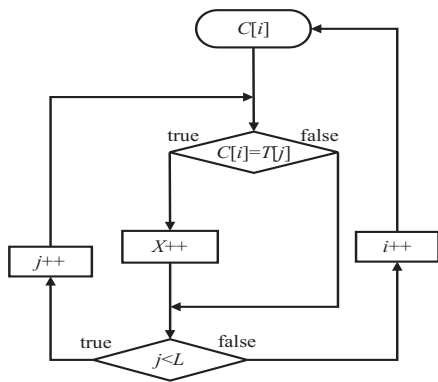


图1 常用字出现次数统计流程

该流程图中进行循环操作将对常用字和电子文本中的汉字逐一进行对比,符合条件 $C[i] = T[j]$ 时 X 的值加1,不符合时先判断 $j < L$ 的值是否为真,如果值为真 j 的值加1,值为假 i 的值加1。循环直到条件 $i < N$ 的值为假为止执行,如果值为假,那么将结束出现次数的统计,最后 X 的累计值作为出现次数 F 的值。

第三步:根据式(1)计算覆盖率。

(2) 程序算法的实现。

根据式(1)和程序处理流程,覆盖率统计核心Java^[9-11]程序算法如下:

```

int X; //统计常用字在文本中的出现次数
public int getCiShu_tongji( String text ) { //获取出现次数的函数
    X = 0; //出现次数赋值为0
    for ( int i = 0; i < N; i++ ) {
        for ( int j = 0; j < T.length(); j++ ) {
            if ( C[i].equals( T.charAt( j ) + "" ) )
                { X++; }
        }
    }
    return X; //返回出现次数的值 }

```

(3) 程序统计流程示例。

下面将通过一个简单示例来说明程序覆盖率统计的过程,首先需要有一个常用字表和文本。为了简化,只

抽取了频度最高的 14 个常用字。具体覆盖率统计示例如下:

常用字:的,一,是,不,了,在,有,人,这,上,大,来,和,我

文本:这些是不是你的?
可以算出文本长度 L 的值为 7,常用字 N 的值为 14。那么首先计算常用字在文本中的出现次数 F ,具体流程如表 1 所示。

表 1 覆盖率统计流程示例

序号		是	否	有	人	在	家	次数
1	的	0	0	0	0	0	0	0
2	一	0	0	0	0	0	0	0
3	是	1	0	0	0	0	0	1
4	不	0	0	0	0	0	0	0
5	了	0	0	0	0	0	0	0
6	在	0	0	0	0	1	0	1
7	有	0	0	1	0	0	0	1
8	人	0	0	0	1	0	0	1
9	在	0	0	0	0	1	0	1
10	上	0	0	0	0	0	0	0
11	大	0	0	0	0	0	0	0
12	来	0	0	0	0	0	0	0
13	和	0	0	0	0	0	0	0
14	我	0	0	0	0	0	0	0
总出现次数 F								5

上述表 1 所示,常用字与文本中的汉字逐一对比后的常用字出现次数 F 的值为 5,那么根据覆盖率统计公式计算结果如下:

覆盖率 = $\frac{F}{N} \times 100\% = \frac{5}{7} \times 100\% = 71.43\%$

2 使用率统计算法

常用字使用率是指电子文本中所出现的常用字在常用字中的比率。

2.1 使用率统计数学算法

通过统计电子文本中的常用字使用率,可以了解到文本中所使用的常用字使用比率,具体数学表达式如式(2)所示。

常用字使用率 = $\frac{\text{使用次数}}{\text{常用汉字数}} \times 100\% =$
$$\frac{G}{N} \times 100\% = \frac{\sum_{i=1}^N \sum_{j=1}^L Y(C_i, T_j)}{N} \times 100\% \quad (2)$$

其中, G 为文本中常用字使用次数(该值不计算重复出现的常用字), N 为常用字数, C_i 为常用字表中下标为 i 的汉字, T_j 为电子文本中下标为 j 的汉字。电子文

本中使用次数 G 是通过常用字与电子文本逐一对比后获得的结果,但与式(1)的出现次数 F 还有一定区别。 G 在统计过程中不计算重复出现的常用字,因为常用字与电子文本汉字对比时,只要有一个符合条件,它就代表该常用字已经使用,因此无需与下一个文本汉字对比。即当 $C_i = T_j$ 时, $Y(C_i, T_j) = 1$ 且 $j = L$, 当 $C_i \neq T_j$ 时, $Y(C_i, T_j) = 0$ 。 $j = L$ 表示一旦符合条件将文本 T 的下标 j 赋值为文本长度 L , 此时就会重新开始从下一个常用字进行对比,避免了重复计算,保证了统计结果的准确性。

2.2 使用率统计程序算法的实现

- (1) 程序处理流程。
根据式(2)使用率程序处理流程定义如下:
第一步:获取常用字字数 N 。
第二步:计算文本中的常用字使用次数 G , 流程如图 2 所示。

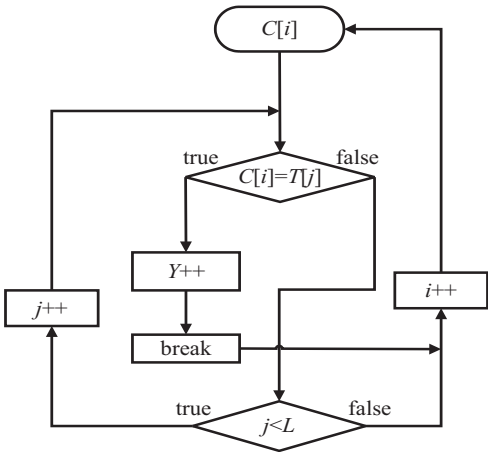


图 2 常用字使用次数统计流程

图 2 中单个常用字 $C[i]$ 在循环对比过程中如果满足条件 $C[i] = T[j]$, 首先将 Y 的值加 1, 再用 `break` 命令结束内循环, 这样就可以保证每个常用字统计结果不重复。然后 i 的值加 1, 再从下一个常用字 $C[i]$ 开始统计。

- 第三步:根据式(2)计算使用率。
(2) 程序算法的实现。

以下为使用率统计核心算法。

```
int X; //统计常用字在文本中的个数
public int getShiYong_tongji( String text) { //获取使用次数
的函数
    Y = 0; //使用次数赋值为 0
    for ( int i = 0; i < N; i++) {
        for ( int j = 0; j < T.length(); j++) {
            if ( C[i].equals( T.charAt( j) + " " ))
                Y++;
            break;
        }
    }
    return Y; //返回出现次数的值 }
```

3 字频统计算法

字频是指每个常用字在文本中的出现频度^[12-14]。

3.1 字频统计数学算法

为了计算字频,首先需要统计每一个常用字在文本中的出现次数,然后再将每个汉字的出现次数除以文本长度,具体字频统计数学表达式如式(3)所示:

常用字字频 = $\frac{\text{单个常用字使用次数}}{\text{常用字数}} \times 100\% =$

$$\frac{P_i}{L} \times 100\% = \frac{\sum_{j=1}^L X(C_i, T_j)}{L} \times 100\% \quad (3)$$

其中, P_i 为每个常用字在文本中的出现次数, C_i 为常用字表中下标为 i 的汉字, T_j 为电子文本中下标为 j 的汉字。每次对比后 $X(C_i, T_j)$ 累计值作为 P_i 的值,再计算下一个常用字 P_i 的值,即当 $C_i = T_j$ 时, $X(C_i, T_j) = 1$, 当 $C_i \neq T_j$ 时, $X(C_i, T_j) = 0$, 直到 $j < L$ 满足条件为止进行累计计算。由于该公式只表示一个常用字的频度,因此结束运算后还需再次使用公式计算下一个常用字 P_i 的值。

3.2 字频统计程序算法的实现

(1) 程序处理流程。

字频统计程序流程如图 3 所示。

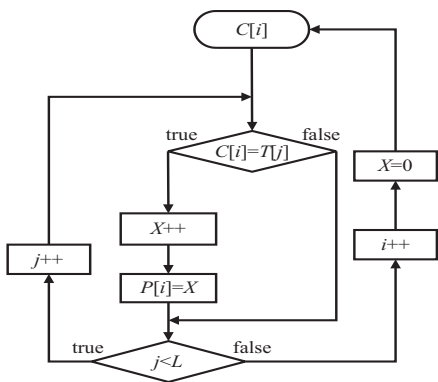


图 3 常用字字频统计流程

在此流程中首先还是要对单个常用字 $C[i]$ 进行逐一对比,如果满足条件 $C[i] = T[j]$, X 的值加 1 并将该值赋给负责存储每个常用字频度的数组 $P[i]$, 然后判断下一个条件 $j < L$, 如果为真 j 的值加 1 并继续对比同一个常用字与下一个文本汉字,循环直到满足条件 $i < N$ 为止执行。

(2) 程序算法的实现。

字频统计核心程序算法如下:

```
int X; //统计每个常用字在文本中的个数
int[] P=new int[ N]; //该数组用于获取下标为 i 的常用字在文本中的使用次数。
public int[] getPinDu_tongji( String text) {
X=0; //使用次数赋值为 0。
for ( int i =0; i < N; i++) {
```

```
for ( int j =0; j < T.length(); j++) {
if ( C[ i ]. equals( T. charAt( j ) + " " ) )
{ X++; }
}
P[ i ] = X; //将使用次数 X 的值赋给数组 P
X=0; }
return X; //返回出现次数的值 }
```

4 常用字覆盖率统计分析系统

4.1 系统框架

服务器操作系统:CentOS 7;
使用编程语言:Java,JavaScript,XML^[15-16];
使用开发工具:Eclipse;
系统框架:主要采用 B/S 架构。

4.2 系统功能

系统可以根据输入的文本进行统计分析,可以统计文本中常用字的覆盖率、使用率、字频等。可根据需要选择目标常用字,即可以选常用 581、1 000、2 500 个常用字表对文本进行统计分析。图 4 为常用字覆盖率统计分析系统的字频统计功能界面。

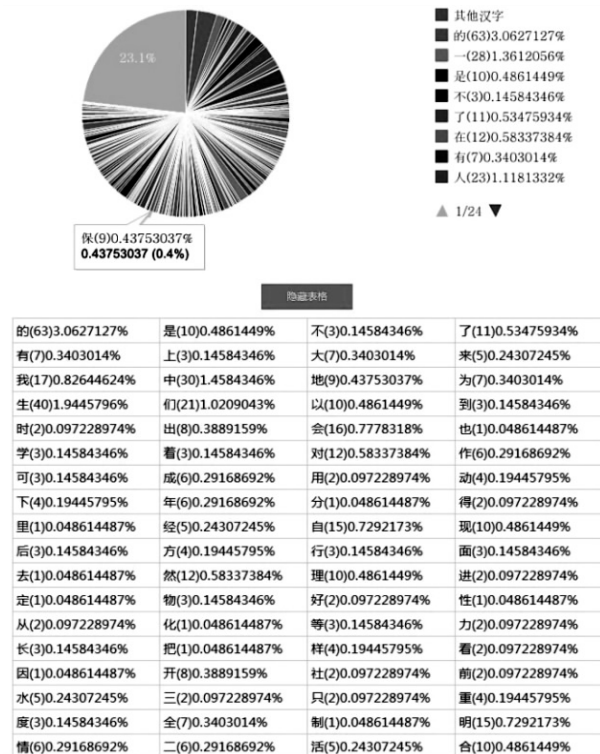


图 4 常用字在文本中的字频统计

5 常用字统计实验

为了测试系统,以四大名著和新华网、人民网共 116 篇文章作为统计对象,分别统计分析了字频最高的 581 个常用字、1 000 个常用字和 2 500 个常用字在这些统计对象中的覆盖率、使用率以及字频,具体统计结果如表 2 所示。

表2 常用字统计分析

统计对象	字数	581 个常用字		1 000 个常用字		2 500 个常用字	
		覆盖率/%	使用率/%	覆盖率/%	使用率/%	覆盖率/%	使用率/%
红楼梦	730 537	72.8	99.1	84.6	98.5	96.0	94.9
三国演义	486 106	58.5	97.8	72.7	97.3	94.3	91.6
水浒传	705 811	67.1	98.1	80.2	98.1	95.7	94.1
西游记	604 456	67.1	99.0	79.1	98.4	95.1	95.6
新华网和人民网 文章(116 篇)	214 068	78.9	100	90.2	99.2	98.8	85.0
	平均值	68.9	98.8	81.4	98.3	96.0	92.2

那么再来看一下统计对象中常用字字频的情况。具体统计结果如表3所示。
在统计结果中只抽取了使用频度最高的前10个汉字,

表3 字频统计

序号	《红楼梦》	《三国演义》	《水浒传》	《西游记》	新华网、人民网文章 (116 篇)
1	了(21 179)2.899%	不(6 731)1.384%	道(10 433)1.478%	道(11 152)1.844%	的(6 290)2.938%
2	的(15 712)2.150%	人(5 120)1.053%	了(11 459)1.623%	不(9 001)1.489%	一(2 538)1.185%
3	一(12 136)1.661%	大(4 173)0.858%	一(10 029)1.420%	一(8 149)1.348%	国(2 290)1.069%
4	来(11 420)1.563%	来(3 287)0.676%	来(9 798)1.388%	了(7 780)1.287%	中(1 948)0.909%
5	道(11 027)1.509%	下(2 780)0.572%	人(8 828)1.250%	那(7 530)1.245%	人(1 722)0.804%
6	人(10 531)1.441%	于(2 778)0.571%	不(8 351)1.183%	我(7 253)1.199%	在(1 540)0.719%
7	是(10 133)1.387%	中(2 694)0.554%	个(6 577)0.931%	是(6 685)1.10%	是(1 394)0.651%
8	说(9 695)1.327%	为(2 646)0.544%	上(5 754)0.815%	来(6 036)0.998%	大(1 352)0.631%
9	我(9 155)1.253%	而(2 498)0.513%	去(5 496)0.778%	他(5 802)0.959%	了(1 286)0.601%
10	他(7 730)1.058 1%	可(2 310)0.475%	大(4 644)0.657%	个(5 752)0.951%	发(1 266)0.591%

从表3中可以看出,根据不同的统计对象常用字的使用频度也会有所不同。

6 结束语

对常用字在教育资源电子文本中的覆盖率统计,使用率统计,频度统计相关的统计算法进行了研究,并结合相关程序算法,以计算机程序的方式来实现一个常用字覆盖率统计分析系统,并通过统计分析系统对四大名著和新华网、人民网116篇文章中所使用的常用字进行了统计分析。结果表明常用字在文本中的覆盖率和使用率相当高,即581个常用字在文本中的覆盖率平均在68.9%以上,1 000个常用字在文本中的覆盖率平均在81.4%以上,2 500个常用字在文本中的覆盖率平均在96%以上,并且常用字在不同统计对象文本中的使用频度也会有所不同。因此常用字不管是在生活中还是在工作中都无处不在,对人们的学习、生活、工作起着至关重要的作用。

参考文献:

[1] 孙曼均. 汉字应用水平测试用字的统计与分级[J]. 语言文

字应用,2004(1):63-70.

[2] 王永强. 常用汉字581[M]. 北京:语文出版社,2006.
[3] 阿不都克里木·玉素甫,王亮亮,覃其益. 基于常用581个汉字的双语点读学习系统[J]. 计算机与现代化,2016(2):91-93.
[4] 宋娟. Java常用算法手册[M]. 第3版. 北京:中国铁道出版社,2016.
[5] 邓洁,桂改花. 计算机数学:算法基础 线性代数与图论[M]. 北京:人民邮电出版社,2016.
[6] SEDGEWICK R,WAYNE K. 算法(第4版)[M]. 谢路云,译. 北京:人民邮电出版社,2012.
[7] 张卫国. 阅读:覆盖率、识读率和字词比[J]. 语言文字应用,2006(3):102-109.
[8] CHEN Lin, PERFETTI C, FANG Xiaoping, et al. Reading Pinyin activates sublexcial character orthography for skilled Chinese readers[J]. Language, Cognition and Neuroscience, 2019,34(6):736-746.
[9] QIU Dong, LI Bixin, LEUNG H. Understanding the API usage in Java[J]. Information and Software Technology, 2016, 73:81-100.
[10] WANG Lulu, LI Jingyue, LI Bixin. Tracking runtime concur-

(下转第210页)