文章编号:1008-0570(2007)11-1-0221-02

基于计算机的作者身份识别的研究

The Study on Authorship Recognition Based on Computer

(沈阳化工学院)李国强 李瑞芳 魏立峰
LI GUOQIANG LI RUIFANG WEI LIFENG

摘要:本文利用相关系数检验法,在基于相关性分析的基础上,讨论客观量化地分析文学作品作者身份的技术应用,并在此基础上进行了可信度的论证,初步得出验证未知作者身份的可行性。

关键词:计算语言学;语言风格;相关系数;可信度
中图分类号:TP3-05 文献标识码:A

Abstract:The correlation coefficient method is used in this paper. Based on the analysis of correlation, the main of goal this study has been achieved, that is, to use scientific ,objective ,quantitative methods of statistical analysis to determine whether it is feasible in the authorship. And show the faith degree.

Key words:computional linguistics,language style,correlation coefficient, faith degree

技术创新

中国文学作品研究对于计算机工作者来说是一个新领域。基于计算语言学的理论、方法,可以加强对中文语言现象进行较深层次的研究与探索,进行信息数据挖掘,同时也会促进文学研究的深入发展,推动中文信息处理的发展。

本文从统计学的角度出发,遵照客观的量化分析,讨论将未知作者身份的作品依据统计学的方法模型化,探索利用统计学的方法,进而得出有争议作品的真实作者身份的最大可能性或最小可能性。

1 语言风格概念

语言风格是指作者在作品中体现出的个人特征,如思想情操、语言文字等各个方面。尤其是语言文字方面,每个作者选词、用词的方式不尽相同,这也是个人风格所在。这种风格也可以在各位作者的作品的统计特征上表现出来,换句话说,风格可以在数量上有所体现,是可以量化的,这是计算语言学的一个研究领域,即计算风格学。

关于风格的要素具体由什么组成的观点各有不同,难以统一。有人赞同风格是作者的记号,是个人的特征模式;有人认为风格是有意识或无意识选择的结果;有人认为风格是用同一种语言传达近乎同一个意思的两个言辞, 他们的语法结构不同,这就是不同的风格。无论这些答案是否完整,大家人都认可风格是一个作者区别于其他作者的本质特征。

在这种统计方法中,最棘手的问题是选样。简单地说,选样的基本点就是以最小的努力获得整部文本的最准确的信息。选样的第二个标准是尽可能的决定评估可信度的客观性,即选样的内容决定评估的准确性。

虚字是句子的应用特征,能与作家们造句时的心性吻合,而使文句展现特色,不仅表现出作家们的造句技巧,而且,还将他们的写作个性自然的流露了出来。这种在不知不觉中自然流露出来的写作个性,基本上应该属于作者的真正写作特征。

结合文学作品的用词习惯和汉语虚词字典,可以对虚字进行抽样,并对其进行划分,如划分为:句尾虚字、白话虚字、文言虚字、转折虚字等类。同时,形成向量空间,每一类虚字以向量的形式存在。

2 作者身份识别方法与技术

2.1 风格的量化

文学作品是语言文字的组合, 语言是人类用来表达和交流思想的工具,文字是记录语言的符号,语言风格是以语言文字的形式表现出来的,在计算机应用技术中,要通过计算机的程序化,使其以客观、量化的数字形式表现出来,便于计算机的运算。检索流程图如下:

经过检索、统计,我们可以采集到基本数据,生成向量空间,作为实验样本。

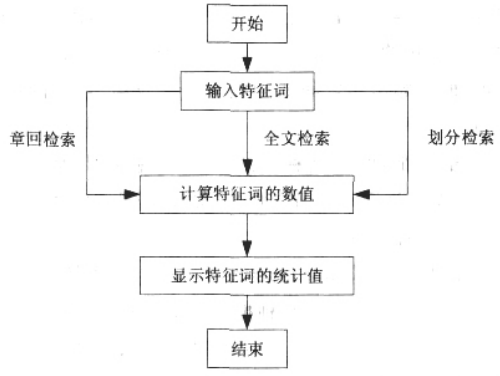

图 1 检索流程图

2.2 相关性分析

当对 n 个指标变量进行相关性测试时, 用相似系数来衡量变量之间的相关程度,一般地,若 $r_{\alpha\beta}$ 表示变量 $x_\alpha x_\beta$ 之间的相似系数,应满足:

1) $|r_{\alpha\beta}| \le 1$ 且 $r_{\alpha\alpha}=1$;

2) $r_{\alpha\beta}=\pm 1$ 当且仅当 $x_\alpha=cx_\beta (c\neq 0)$;

3) $r_{\alpha\beta}=r_{\beta\alpha}$。

$r_{\alpha\beta}$ 的绝对值越接近于 1,说明变量 $x_\alpha x_\beta$ 的关联性越大。

李国强: 讲师 硕士在读

基金项目:辽宁省教育厅科学研究计划(20040291)

相似系数中最常用的是相关系数。Pearson Correlation Coefficient(PCC)方法是一种统计学方法,它可以定量地衡量变量之间的相关关系,在计算语言学中被普遍应用,主要用于文本分类、信息抽取等方面的研究。通过对特征词的数字化,进行相关度检测,进而对文本信息进行分析、研究。

在本研究中,采用这一数学理论。计算方法步骤如下:

1) 求 $\sum X, \sum Y, \sum X^2, \sum Y^2, \sum XY$

2) 按公式求取相关系数 r

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

公式中变量说明如下:

N:样本数据组数

X,Y:做相关分析的两变量

$\sum X^2$, $\sum Y^2$:两变量的平方和

$\sum XY$;两变量对应值乘积之和

程序框图如下:

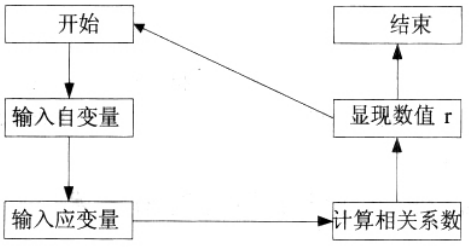

图 2 相关系数计算程序图

结合检索采集到的基本数据,按照上面的算法描述。可以采用 VC++等软件设计系统,经过程序具体实现,进行相关性检测。

假设Ⅰ为未知作者的作品文本,Ⅱ、Ⅲ为两部已知的同一作者的作品。如果Ⅰ Ⅱ、Ⅰ Ⅲ、Ⅱ Ⅲ的相关系数值相近,甚至相等,假设值约等于 M。那么Ⅰ变化,Ⅱ也进行相应的 M 程度的变化;同理Ⅰ变化,Ⅲ也进行相映的 M 程度的变化;Ⅱ变化,Ⅲ也进行相映的 M 程度的变化。即Ⅰ≈Ⅱ≈Ⅲ。这里所说的变化是指作者在不同文章中的虚词样本使用的情况,这里的 M 值是指相关系数值。

Ⅰ、Ⅱ、Ⅲ三部作品之间的相近程度比较高,他们的系数近乎相等。则我们可以初步推测:Ⅰ≈Ⅱ≈Ⅲ,即三部作品的写作风格有一定的相似性,按此种方法分析,可初步认为作品Ⅰ与作品Ⅱ、Ⅲ的作者为同一作者。反之,则不为同一作者。

2.3 显著性的假设检验

参数估计与假设检验是统计的两个必要组成部分,它们都是通过样本数据对总体进行某种推断。根据样本数据计算出来的相关系数 r,只是抽取样本相对总体相关系数 ρ的估计值。所选取的样本能否代表总体特征,得出的结果能否体现整部著作的风格取向,需要验证。因此,需要对相关系数 r 进行检验,估计值 r 是否可靠,是否接近或等于 ρ即检验相关系数的显著性。

样本相关系数 r 的分布取决于样本容量 n 和总体相关系数 ρ的值。当总体相关系数 ρ=0 时,相关系数 r 的分布随样本容量的增加而趋于自由度为 n-2 的 t 分布,即统计量: $t = \frac{\sqrt{n-2}\, r_{xy}}{\sqrt{1-r_{xy}^2}}$

服从自由度为 n-2 的 t 分布。因此,可以用 t 分布对两个变量总体之间是否相关进行假设检验。

根据所研究问题的性质不同,显著水平 α的取值也不同,一般多规定 α=0.1、α=0.05、α=0.01,分别表示中等显著、显著和高度显著。在本文中只需测试相关性是否显著,即相关性的真实性,所以建议一般选取 α=0.05 作为显著水平。

通过计算机程序实现,得到 t 值。如果所有的 t 值都大于临界值 $t_{\frac{\alpha}{2}}(n-2)$。则可以初步得出:用公式根据样本数据计算出来的相关系数 r,是从中抽取样本的总体之间的相关系数 ρ的估计值。即使由于抽样误差的存在,样本数据对总体进行的相关性推断依然成立,估计值 r 是可靠的,接近或等于 ρ即检验相关系数的显著性,假设性推断成立。反之,则假设性推断不成立。

3 结束语

在本研究中采用了计算机技术与文学研究相结合的方式,利用计算机程序对文学著作的文本进行统计分析,通过这样的嫁接和融合,在文学研究和计算机科学技术的边缘探索具有研究意义的领域,其方法具有一定的代表性。本文作者的创新点是采用了客观、定量的统计分析方法来推测作者身份未知的作品是谁所作,可以得出其具有最大可能性。当然,这只是初步探索,将来还再进行更深入的研究,这样将能够得出的更能使人信服的结论。

参考文献

[1]李敏。数据挖掘在辅助决策系统的应用研究[J],微计算机信息,2004 ,5 :96- 97
http://reddream.net/Article/ShowArticle.asp?ArticleID=3361
[2]Bing C. Chan. The Authorship of The Dream of the Red Chamber Based on a Computerized Statistical Study of Its Vocabulary [M], HK: Joint Publishing Co.,1986,63- 70
[3]范金城,梅常林。数据分析,北京:科学出版社[M],2002,205- 210
[4] 姚文辉。相关系数的一种简易计算程序[J],甘肃环境研究与监测,2003,16 (4):363- 365
[5] 刘顺忠。数理统计理论、方法、应用和软件设计[M],武汉:华中科技大学出版社,2005,13- 33

作者简介:李国强(1977-), 男,辽宁人,工作于沈阳化工学院计算机科学与技术学院,讲师,硕士在读,主要从事计算机应用技术的研究。

Biography:Li Guo- qiang (1977-),man (The Han nationality), Liaoning, Shenyang Institute of Chemical Technology, Lecturer, Master is in read, Major in the computer applied technique.

(110142 沈阳 沈阳化工学院)李国强 李瑞芳 魏立峰

通讯地址:(110142 沈阳 沈阳化工学院计算机科学与技术学院)李国强

(收稿日期:2007.8.13)(修稿日期:2007.10.15)