文章编号:1000-5862(2015)02-0117-07

中国古诗统计建模与宏观分析

钱 鹏[1],黄萱菁[2*]
(1. 复旦大学中国语言文学系,上海 200433;2. 复旦大学计算机学院,上海 201203)

摘要:利用自然语言处理技术处理文学文本是计算语言学领域近年来的热门话题. 该文结合点态互信息量与频率阈值,自动发现中国古诗词汇. 基于构建的诗歌词典,利用启发式的正向最大匹配算法,对中国古诗作分词处理. 采用主题模型对分词后的诗歌文本进行统计建模,并在此基础上进行了主题演变和诗人群体风格网络的探索性分析. 基于全唐诗语料的实验结果表明:主题模型可以给出具有较好解释力的中国古诗统计模型,验证已有的文学史研究,并在传统的文本细读的研究范式之外,对中国诗学提供了全新视角的宏观刻画、描述与阐释.

关键词:中国古诗;统计建模;分词;主题模型
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.02.02

0 引言

中国诗歌区别于其他文学形式的最重要的特点就在于其丰富的形象、多变的主题以及特殊的修辞结构. 诗歌主题、体裁的多样性发展经历了文学观念、文化语境等历史因素的变化. 诗歌意象的模式与类聚、新意象的创造与突围,也代表了一批诗人在探索诗歌语言形式和诗歌意象道路上、竭力突破已有范式的反常创造. 本文通过考察大量诗歌文本,以唐代诗歌作为切入点,纵横数百年的发展历史,探索用计算和定量的研究范式发现意象聚合的历史事实,勾勒出诗歌主题的历史波动,验证文学史研究者提出的题材演化与诗歌变迁的理论.

利用自然语言技术处理文学文本是计算语言学领域近年来兴起的热门话题. 以往的自然语言处理研究注重新闻类文本的分析,而后有研究开始转向文学性语言的计算分析. 然而,绝大多数工作是集中于计量风格学领域的研究,如作者识别、风格分析等[1-6],或者是单纯研究诗歌生成任务[7-9]. 近年来出现了从计算分析的角度关注文学理论的研究,如文献[10]关注了如何从小说文本中自动抽取出相关的人物社会关系网络. 文献[11]对英语文学 1800 年之后 200 多年的历史作了风格学分析,考察了文学作品的影响与时代风格的关系,英语文学语料显示,较早时间段的作家对邻近时期的作家影响比较大,但较晚阶段的影响较小. 文献[12]对英文诗歌做了分析,试图从定量分析的角度,找出优秀诗歌区别于一般诗歌的特质. 文献[13]对中国当代诗歌进行了分析,认为内地当代诗人的诗歌呈现出较高的远离传统诗歌的特征,但同时发现台湾诗人的创作中似乎保留了一些古典诗歌的风格.

诗歌在中国是一个极为重要且历史悠久的文学体裁. 纵横千年的时间跨度、数量巨大的诗人群体、卷帙浩繁的诗歌文本都使得在传统的文本阅读方式下,中国诗歌的宏观把握变得极其不易. Voigt 等[13]虽然也关注了中国现代诗歌与古典诗歌的差异,但其使用的数据量较小,提供的分析手段与评判指标仍较为粗糙. 因此,本文希望利用自然语言处理技术,对中国传统诗歌文本进行统计建模. 试图从另一个视角出发,更高效地处理中国文学大数据,尝试提供计算语言学视角下对中国诗学的宏观刻画、描述与阐释.

本文的研究目标是以全唐诗为样本进行主题建模,从宏观和计算的角度,验证文学史研究的已有结论,并尝试着发现一些新的文学史现象. 但与以往基于字的诗歌分析不同,笔者认为双音节词汇的发展决定了对诗歌文本进行分词的必要性. 因此,在现有

收稿日期:2015-02-16
基金项目:国家自然科学基金(61472088)资助项目.
通信作者:黄萱菁(1972-),女,浙江平阳人,教授,博士生导师,主要从事自然语言处理和信息检索的研究.

的基于现代汉语语料库的分词技术无法在古典文学领域取得良好适应性的现状下,本文首先进行词汇发现,构建诗歌词典,利用基于词典的正向最大匹配法,获得较好的诗歌分词结果. 在分词的诗歌文本之上,进行主题建模和探索分析.

1 词语发现

1.1 启发式方法

汉语文本的考察必然要牵涉到字与词的问题. 一方面,从形式上能够完全确定的是字的边界,但在语言学意义上,词是真正能够独立运用的最小单位. 另一方面,对于古典诗歌等古代汉语文本,尚且缺乏直接可用的词表. 因而,从以字为单位的文本中发现潜在的词语,是一项很重要的工作.

词语发现主要是基于 2 元组合点态互信息量(Point-wise mutual information)的搭配发现算法. 只不过,在原始版本的搭配发现算法中[14],待发现的是词组搭配,构成搭配的是词. 而在本文中,待发现的是词语,其基本的形式单位是字. 因此,本文借鉴统计抽词方法[15],并更改相关变量的定义,将此算法迁移至由字组词的过程中,以发现具有特殊含义的诗歌词汇. 此外,值得说明的是,汉语具有强烈的双音化倾向,大多数词是双音节形式. 因此,在词语发现的过程中,暂且只考虑 2 元模式.

定义 2 个汉字 C_1、C_2 结合成的二元模式 C_1C_2 能够被当成一个词的基本条件为

$$IsWord(C_1C_2) = \begin{cases} \text{True}, & PMI(C_1,C_2) > \theta, \\ \text{False}, & PMI(C_1,C_2) < \theta, \end{cases}$$

其中

$$PMI(C_1,C_2) = \log(P(C_1,C_2)/(P(C_1)P(C_2))) \propto \log(\#(C_1,C_2)/\#(C_1)\#(C_2)).$$

关于阈值 θ 的确定,观察了 2 元模式点态互信息量的分布(见图 1). 从图 1 的累积分布图可以大致看出信息量的分布情况. 高信息量的词汇较少,而信息量中等的 2 元组居多. 因此,选取 7 作为信息量阈值,以作为判断汉字 2 元组是否可能成词的首要条件.

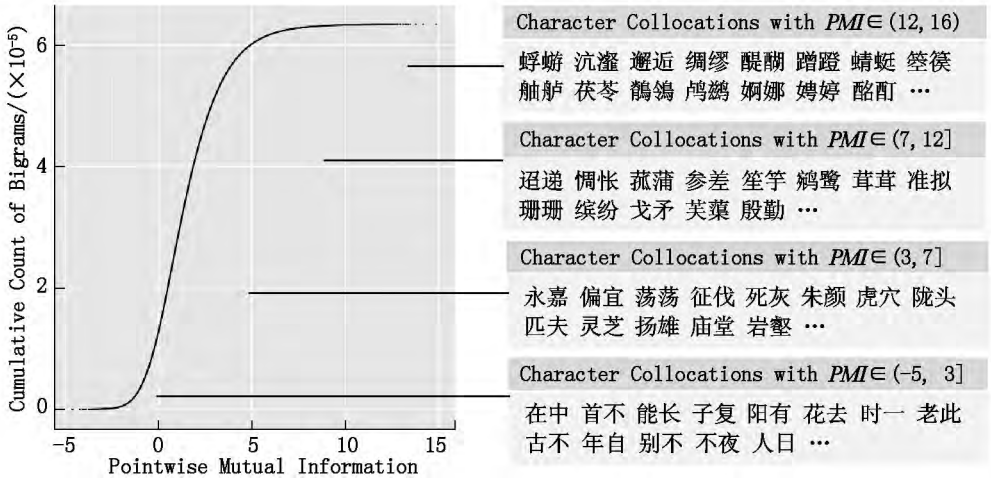

图 1 全唐诗 2 元组 *PMI* 累积分布

同时,为了过滤掉无意义的高信息量低频组合,本文也分段设置了频率限制. 不同 *PMI* 值区间的 2 元汉字组,对应不同的频率阈值. 对于 *PMI* 非常高的 2 元组,不限制其出现频率;对于 *PMI* 稍低的 2 元组,要求其出现次数必须超过一定的阈值. 基于以上条件,得到了较为理想的词表.

1.2 定性评价

经研究发现,利用 *PMI* 找出的词汇,特别是信息量高的词汇,确实具有特殊的意义. 词汇发现不仅在涉及古代汉语文本处理的工程实践中起到辅助作用,对于文学研究者来说亦能够提供帮助,对于希望了解古代文化的人也能提供良有裨益的参考. 试举一些 *PMI* 值较高的词的释义,如表 1 所示.

1.3 定量评价

为了定量评价发现的诗歌词汇,采用人工判断的方法. 按 *PMI* 逆序排列发现出的词汇,请中国语言文学专业的本科生判断所抽取的 2 元组是否为真词. 同时,又对比了单纯考虑频率的词汇发现方法,即按照频率高低逆序排列 2 元组,再人工判断,计算真词率. 评价结果如表 2 所示. 从表 2 可以看到,本文采用的点态互信息量和频率阈值相结合的方法,确实在词汇发现任务中取得了不错的效果.

表 1　诗歌词汇列举

词语	解释	词语	解释
蠓螮	彩虹的别名,一般比喻才气横溢	罘罳	古代的一种屏风
苜蓿	即三叶草	褭褭	羽毛初生的样子
婕妤	为嫔妃的等级之一.	沆瀣	“沆瀣”二字都指夜间的雾气或露水,故有成语“沆瀣一气”
蠨蛸	蜘蛛的一种,脚长. 通称蟢子		
肹蚳	散布,传播.	舢舨	形如蚱蜢的小船.《广雅·释水》:“舢舨,舟也.”
旖旎	柔和美丽		
桔槔	古代的一种原始汲水工具.	醍醐	从酥酪中提制出的油.
蟪蛄	蝉,《庄子》有“蟪蛄不知春秋”	薏苡	即薏仁米
彷徨	连绵词,表示来回走,犹豫不决	睢盱	浑朴貌; 睁眼仰视的样子
鸑鷟	古代民间传说中的五凤之一,身为黑色或紫色	舳舻	指首尾衔接的船只. 舳: 指船尾; 舻: 指船头
葡萄	从西域传来的一种水果	娉婷	形容女子姿态美好

表 2　前 N 个 2 元组中的真词比例

N		100	200	300	400
PMI + Frequency	阈值	0.960 0	0.955 0	0.947 0	0.942 5
Frequency		0.750 0	0.750 0	0.673 0	0.622 5

2　探索分析

2.1　基于词语发现的诗歌分词

处理古代汉语文本的常见方式,是将文本全部打散为独立的汉字. 这一处理方式立足于一个基本观念,即孤立的汉字承担了较为充足的语义信息,大多仍是具有较强构词能力的自由语素. 但是,到了唐代,汉语的双音化已经发展到一定的高度,汉语中产生了一系列使用频繁的双音节词汇,应以词为单位进行切分.

然而,诗歌的特点决定了现有的现代汉语分词技术难以在这一特殊领域获得良好的适应性. 因此,基于以上总结的诗歌文本的特点,采用一个启发式的分词方法,基于已获得的诗词词汇,结合诗歌文本的格律特征和古代汉语的语素 -音节对应关系,采用正向最大匹配法,对诗歌文本作简单的切分. 分词算法为

for($i = 0$; $i < len(sent)$;)
if $i == len(sent) - 1$:
在 $sent[i]$ 后切分
if $sent[i] + sent[i+1]$ 是已发现的诗歌词汇:
在 $sent[i+1]$ 后切分
$i = i + 2$
else:
在 $sent[i]$ 后切分
$i = i + 1$.

经过上述分词算法处理之后,得到的诗歌分词结果基本符合预期. 试选登部分诗歌的分词结果如表 3 所示.

表 3　分词后诗歌文本列举

篇目	作者	分词后诗歌文本
乐府答孟东野戏赠	陆长源	芙蓉 初 出 水 菡萏 露 中 花 风 吹 著 枯 木 无 奈 值 空 槎
和陈监四郎秋雨中思从弟据	王维	袅袅 秋 风 动 凄凄 烟 雨 繁 声 连 鹍鹊 观 色 暗 凤凰
赠别	杜牧	娉娉 裊裊 十 三 馀 豆蔻 梢 头 二 月 初 春 风 十 里 扬州 路 卷 上 珠帘 总 不 如
雨	杜牧	连 云 接 塞 添 迢递 洒 幕 侵 灯 送 寂寥 一 夜 不 眠 孤 客 耳 主 人 窗 外 有 芭蕉
赠十娘	张鷟	人 去 悠悠 隔 两 天 未 审 迢 迢 度 几 年 纵 使 身 游 万 里 外 终 归 意 在 十 娘 边

本文还比较了基于现代汉语语料训练的 FudanNLP 中文自然语言处理工具包[16] 在中国古诗领域的适应性和表现. 为便于定量评价,对抽样的 50 首长度不等的诗歌进行人工切分,作为参考答案. 启发式的方法取得了非常不错的效果,显著优于单纯基于现代汉语语料训练的分词模型,结果如表 4 所示.

表 4　分词结果评价

类　别	FudanNLP	启发式算法
准确率	0.641	0.851
精度	0.606	0.995
召回率	0.938	0.851
F 值	0.736	0.917

2.2　主题演变

在分词后的诗歌文本上,采用主题模型(Topic Model) 进行文学史层面的宏观分析. 以往的诗歌分

析大多基于简单的统计量[1]，难以有整体把握. 本文采用生成式模型，对诗歌中的主题分布可以进行更好的宏观建模，如图 2 所示.

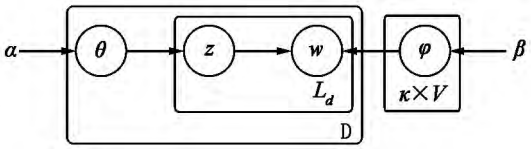

图 2 主题模型

主题模型是 Blei 等[17]提出的一种被广泛应用于文本主题建模的概率图模型. 主题模型以词袋假设作为前提条件，引入狄利克雷先验分布，在文档主题聚类等任务上均取得了不错的效果. 主题模型假定一篇文档 d 是 K 个主题上的分布，每一个主题 k 是词表上的特定分布. 文档中每一个词的产生过程，即为按照文档 d 的主题分布，从 K 个主题中随机选择一个主题；再按照该主题对应的词分布，随机选择一个词语. 这样，可以写出文本 d 中第 i 个词的产生概率为

$$p(w_i) = \sum_k^K p(w_i \mid z_i = k)\, p(z_i = k).$$

本文采用吉布斯采样算法[18]训练主题模型. 采样概率为

$$p(z_i = k \mid \overrightarrow{z_{\neg l}}, \vec{w}) = \frac{p(\vec{z}, \vec{w})}{p(\overrightarrow{z_{\neg l}}, \vec{w})} \propto \frac{n_{k,\neg i}^{(t)} + \beta}{n_{k,\neg i}^{(\cdot)} + V\beta} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha}{n_{m,\neg i}^{(\cdot)} + K \cdot \alpha}.$$

由此可以进一步得到词表 $\Phi(\varphi_{k,t})$、文档主题分布 $\Theta(\theta_{m,k})$ 参数的更新法则为

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta}{n_k^{(\cdot)} + V\beta}, \theta_{m,k} = \frac{n_m^{(k)} + \alpha}{n_m^{(\cdot)} + K\alpha}.$$

更详细的理论推导可参看文献[18]. 本文用主题模型对数万首分词后的诗歌进行建模. 从互联网上获取到了《全唐诗》文本格式语料.《全唐诗》是清代曹寅、彭定求等奉敕编纂，共 900 卷，收录唐代诗人 2 873 人的 49 403 首诗歌作品及 1 555 条全文无考的诗句. 经预处理后，得到 42 700 首不含错误字形的干净文本. 设定主题数 $K = 2$，模型参数 $\alpha = 50/K$，$\beta = 0.01$，采用吉布斯采样算法，试图发现每首诗歌中涉猎的主题以及相应的分布.

通过比较基于分词后诗歌文本的主题模型和基于单字的主题模型在测试集上的困惑度，发现启发式的分词预处理确实有助于提高模型的表达能力. 表 5 及图 3 显示了困惑度与主题数的关系，虚线为基于单字的主题模型，实线为基于分词后文本的主题模型困惑度.

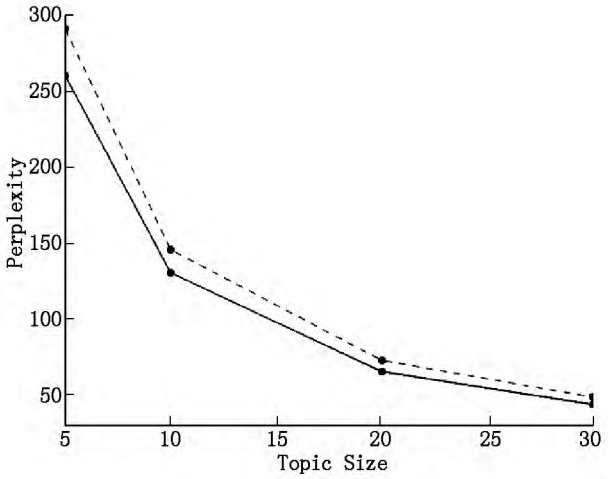

图 3 困惑度与主题数关系

表 5 字 LDA 与词 LDA 困惑度对比

主题数	5	10	20	30
词 LDA	260.3	130.3	65.2	43.5
字 LDA	291.2	145.5	72.6	48.4

考虑到主题辨识的细粒度，本文在进行探索分析时，设定主题数 $K = 2$，训练出相应的模型. 表 6 选列了部分主题的关键词以及人工标注的主题类别.

表 6 主题词及人工标注的主题类别

类别	编号	主题词
清寒	topic 6	月 夜 声 寒 秋 风 清 梦 听 愁 蟋蟀 朦胧 唧唧 螻蛄 淅淅
秾艳	topic 8	花 春 红 香 绿 新 翠 金 芳 丝 鸳鸯 芙蓉 婵娟 鹦鹉 玳瑁
边塞	topic 11	马 城 将 汉 边 行 军 塞 黄 胡 匈奴 慷慨 邯郸 骠骑 霍嫖
颂祀	topic 16	天 神 圣 方 德 文 皇 臣 乐 灵 乾坤 煌煌 宇宙 巍巍 股肱

此外，值得说明的是，《全唐诗》编者基本遵照作者生活年代的次序对诗歌进行整理编排. 诗歌在全唐诗语料库中的位置也可以在一定程度上反映诗歌所处的时代. 因此，将训练后得到的每篇诗歌的 $\vec{\theta}$ 值，按位置次序，每 100 篇进行一次区间归并，得到 427 个时间点(Time Tick)，并绘制出 $\vec{\theta}$ 的每个主题分量在 Time Tick 上的分布以及对应的平滑曲线. 本文认为一首诗歌的某一主题分量如果过低，则其文本实际上并未准确体现该主题. 所以，在绘图的过程中，只绘制主题分量大于一定阈值的数据点，以消除噪声点，使主题在时间轴上的变化趋势更为明显. 经多次尝试，设定阈值为 0.07(接近主题分量的期望值 $1/K$).

图 4 选取了 2 幅有代表性的散点图,反映的是某一主题分量在时间轴上的分布. 本文发现主题随时间的变化确实与文学史研究中已有的结论相符合.

主题 16 是颂祀主题,从散点图和平滑曲线的走势可以看到,唐初出现了一个极大的峰值. 这与明代文学批评家胡震亨《唐音癸签》中的叙述非常一致: "有唐吟业之盛,导源有自. 文皇英姿间出,表丽缛于先程; 玄宗材艺兼该,通风婉于时格. 是用古体再变,律调一新; 朝野景从,谣习浸广. 重以德、宣诸主,天藻并工,赓歌时继. 上好下甚,风偃化移,固宜于喁遍于群伦,爽籁袭于异代矣. "当是时,国运昌盛,不少台阁文人参与到诗歌的创作中来,加上君王好写诗,喜道教,自然会突出"海宇颂皇仁"的主题.

主题 8 是秾艳主题,从图 4 可以看到,平滑曲线在晚唐末期有一个突然的上升. 这一结果和文学史研究的结果完全对应. 初唐时期,陈子昂引领"风雅兴寄",诗文创作开始转变隋末遗留的秾艳之风. 中唐时期的新乐府及诗文复古运动,同样压制了个体情感的表达. 而晚唐五代的思潮有所变化,朝政的暮年之感使得诗人的关注转向个体主体性方面[19],在创作上更多地表现个体的情感,突出凄艳的风格,孕育着词的产生. "词为艳科",表现女性情感的作品在那一时间段更为集中,因而晚唐时期秾艳主题格外突出.

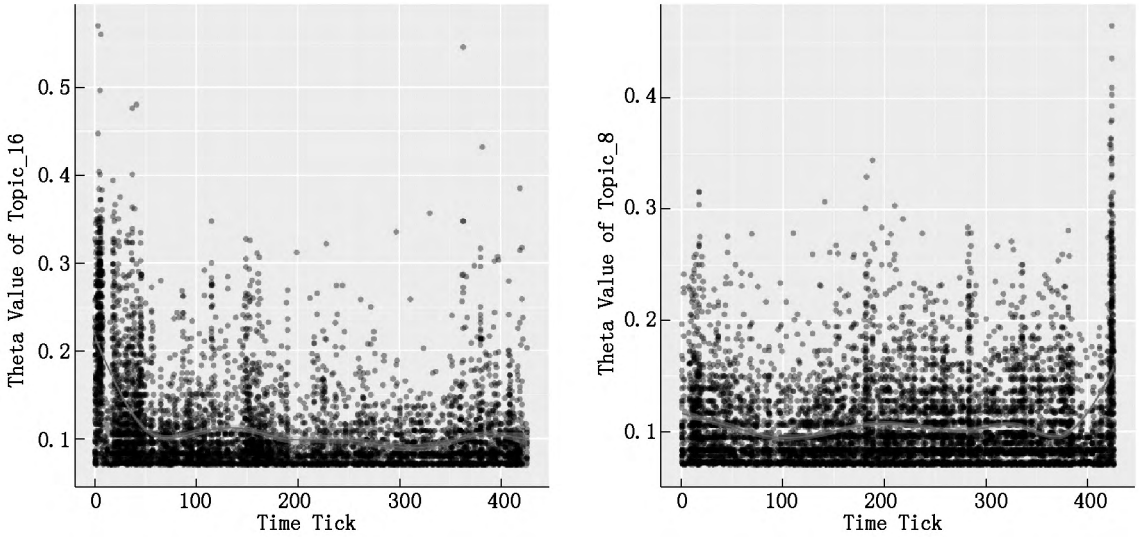

图 4　主题-时间分布图

2.3　诗人群体分析

试图通过诗歌主题分布来比较不同诗人之间的相似度. 对每一个诗人,将其所写过的诗歌的 $\vec{\theta}$ 值相加求算术平均,这样对于每一个诗人都可以计算得到相应的 $\overrightarrow{\theta_A}$. 定义 2 个诗人 A_1,A_2 之间的距离为

$$Distance(A_1, A_2) = \sqrt{\sum_{k=0}^{K} (\theta_{A_1}^k - \theta_{A_2}^k)^2}.$$

给定一个诗人 A_1,可以通过对 A_1 与其他诗人的距离进行排序,发现诗歌创作方面相近的诗人群,构建一个相似诗人群体检索系统.

表 7 是该系统给出的部分结果,A_1 列是待查询的诗人,A_2 列是按照距离递增排序后的前 5 位最接近的诗人. 这一结果与文学批评研究结果也比较接近. 例如,与李白最相似的是李颀与顾况. 这种相似不仅有文本上的直觉作为依据,从诗人的交游历史和诗学追求上亦可佐证[19-20]. 李颀与李白都崇奉道教,李颀深受李白诗风的影响. 顾况的诗歌则具有很强的多样性. 以往的研究过分注重顾况与新乐府的关系,但事实上,顾况的新乐府作品还不到其诗作总量的 1/10,最新的文学史著作开始注意到他与李白相近的风格特征. 而这恰恰与系统检索结果相符. 元稹和白居易在文学史上历来被合称为"元白诗派". 检索结果也显示,与白居易最相似的是元稹.

除此之外,经研究还发现,系统给出的结果亦能对传统的文学史研究作出较好的补充. 例如,对诗人王之涣进行检索,可以看到最相似的诗人是翁绶. 盛唐和晚唐诗人之间的相近,或许帮助研究者更好地理解诗风的流变和复归.

在检索结果的基础上,以诗人作为网络节点,检索出的诗人配对关系作为网络中的边,构建出一个诗学风格影响的网络. 本文发现,该网络的节点度统计量亦基本符合幂率分布(见图 5) . 节点度较高的诗人,意味着他们的诗学风格在同辈之间获得了更高的影响力. 网络分析结果显示,诗人李商隐获得了最高的网络节点度数. 这一新的发现超越了传统文

本分析和阅读直觉的限度,拓展了文学研究理念. 当然,由统计模型发现的相关结论是否能够获得文学史研究者的普遍支持与认同,还有待于更细致的解读.

表 7 相似诗人检索结果

A_1	A_2	$Distance(A_1,A_2)$	A_1	A_2	$Distance(A_1,A_2)$
	李颀	0.067 8		韩琮	0.066 9
	顾况	0.072 0		刘兼	0.071 9
李白	陈陶	0.075 3	温庭筠	成彦雄	0.072 5
	欧阳詹	0.075 3		薛涛	0.072 7
	王昌龄	0.077 1		李绅	0.077 5
	元稹	0.066 9		翁绶	0.114 2
	王建	0.067 0		朱琳	0.117 6
白居易	王绩	0.068 3	王之涣	李约	0.121 5
	张籍	0.072 3		郑锡	0.138 0
	薛能	0.077 1		纪唐夫	0.139 9

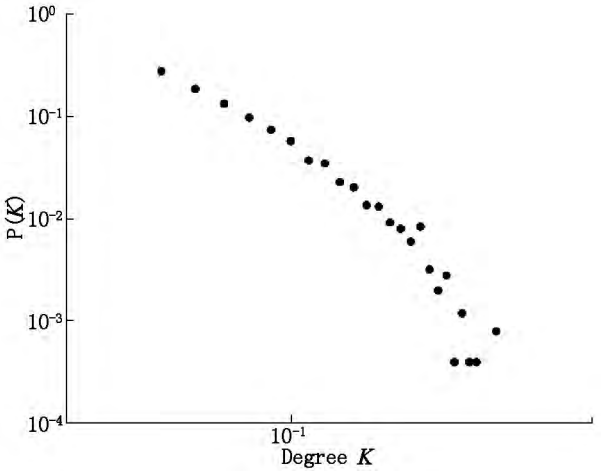

图 5 诗人群体风格网络节点度分布(对数坐标)

3 结论

本文结合点互信息量(*PMI*)与分段频率阈值,较好地构建出古诗词表. 基于发现出的诗歌词汇,用正向最大匹配法进行启发式分词. 在分词的基础之上,采用主题模型进行统计建模与宏观分析. 统计模型不仅为处理古典文献提供了技术手段,更为中国文学研究提供了新的切入点和宏观把握的方式.

在后续工作中,希望进一步完善模型,使其能够具有更好的表达能力和泛化能力,并能够对诗歌中的骈偶结构加以建模,以期在诗歌研究、诗人关系挖掘以及更为复杂的诗歌生成任务(如联句、次韵)中取得更好的表现.

4 参考文献

[1] 胡俊峰,俞士汶. 唐宋诗之计算机辅助深层研究 [J]. 北京大学学报: 自然科学版, 2001, 37(5): 727-733.

[2] 年洪东, 陈小荷, 王东波. 现当代文学作品的作者身份识别研究 [J]. 计算机工程与应用, 2010, 46(4): 226-229.

[3] 武晓春, 黄萱菁, 吴立德. 基于语义分析的作者身份识别方法研究 [J]. 中文信息学报, 2006, 20(6): 61-68.

[4] 张运良, 朱礼军, 乔晓东, 等. 基于句类特征的作者写作风格分类研究 [J]. 计算机工程与应用, 2009, 45(22): 129-131.

[5] McFarland, Daniel A, Christopher D, et al. Differentiating language usage through topic models [J]. Poetics, 2013, 41(6): 607-625.

[6] Stamatatos, Efstathios. A survey of modern authorship attribution methods [J]. Journal of the American Society for Information Science and Technology, 2009, 60(3): 538-556.

[7] 周昌乐, 游维, 丁晓君. 一种宋词自动生成的遗传算法及其机器实现 [J]. 软件学报, 2010(3): 427-437.

[8] He Jing, Zhou Ming, Jiang Long. Generating Chinese classical poemswith statistical machine translation models [C] // Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[9] Zhang Xingxing, Mirella Lapata. Chinese poetry generation with recurrent neural networks [C] // Proceedings of EMNLP, 2014: 670-680.

[10] Elson, David, Nicholas Dames, et al. Extracting social networks from literary fiction [C] //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 138-147.

[11] Hughes, James M, Nicholas J Foti, et al. Quantitative patterns of stylistic influence in the evolution of literature [J]. Proceedings of the National Academy of Sciences, 2012, 109(20): 7682-7686.

[12] Kao, Justine, Dan Jurafsky. A computational analysis of

style, affect, and imagery in contemporary poetry [C] // NAACL Workshop on Computational Linguistics for Literature,2012.

[13] Voigt, Rob,Dan Jurafsky. Tradition and modernity in 20th century Chinese poetry [C] // NAACL Second Workshop on Computational Linguistics for Literature,2013.

[14] Church, Kenneth, William Gale, et al. Using statistics in lexical analysis [C] //Uri Zernik Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale, NJ: Lawrence Erlbaum,1991:15-164.

[15] 苏劲松, 周昌乐, 李翼鸿. 基于统计抽词和格律的全宋词切分语料库建立 [J]. 中文信息学报, 2007, 21(2): 52-57.

[16] Qiu Xipeng, Qi Zhang , Huang Xuanjing. FudanNLP: A toolkit for Chinese natural language processing [C] // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,2013: 49-54.

[17] Blei, David M, Andrew Ng, et al. Latent dirichletallocation [J]. Journal of Machine Learning Research, 2003 (3):993-1022.

[18] Grififths, Thomas L MarkSteyvers. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004,101(1): 5228-5235.

[19] 章培恒, 骆玉明. 中国文学史新著 [M]. 上海: 复旦大学出版社, 2011.

[20] Luo Yuming. A concise history of Chinese literature [C]. Koninklijke Brill NV, Leiden: Netherlands, 2011.

The Statistical Modeling and Macro-Analysis of Chinese Classical Poetry

QIAN Peng[1], HUANG Xuanjing[2*]

(1. Department of Chinese Language and Literature, Fudan University, Shanghai 201203,China;
2. School of Computer Science, Fudan University, Shanghai 201203,China)

Abstract: Modeling literary texts with natural language processing technology has become a popular topic of computational linguistics in recent years. the vocabulary of Chinese classical poetry by combining point-wise mutual information (*PMI*) method and frequency threshold has been extracted. Based on the extracted poetic vocabulary,a heuristic forward maximum matching algorithm to segment the poems has been used. In order to model the poetry,latent dirichlet allocation (topic model), based on which we also put forward explorative analysis of the literature evolution and poet network has been used. The experiments on the corpus of All-Tang poetry indicate that topic model is an explanatory statistical model of the Chinese classical poetry. While proving the existed evolution theory of Chinese literature, the statistical model also provides insightful macro-analysis from a new perspective, in addition to the traditional methodology of Chinese literature research.

Key words: Chinese classical poetry; statistical modeling; word segmentation; topic model

(责任编辑: 冉小晓)