宋词研究的新视角：文本关联与时空可视分析

张　玮[1)], 谭思危[1)], 刘　凯[1)], 石　磊[1)], 陈思明[2,3)], 陈　为[1)*]

[1)] (浙江大学 CAD&CG 国家重点实验室　杭州　310058)
[2)] (Rheinische Friedrich-Wilhelms-Universität Bonn　Bonn　53113)
[3)] (Fraunhofer Institute for Intelligent Analysis and Information Systems　Sankt Augustin　53757)
(chenwei@cad.zju.edu. cn)

摘　要：传统的历史与文学研究通常是案例式地分析具体的文学作品与词人的背景. 文学作品的内容、情感与含义和作者的生平经历、师承流派、家国时代背景息息相关，传统的分析方法难以直观地了解以上各项内容的关联. 以宋词作为文学作品代表进行分析，提出了一套基于文本关联与时空可视分析的方法. 它支持文学研究者从词人的生平轨迹和不同生活年代的背景进行分析与对比，探索不同年代、不同经历的词作在文本主题上的相关性与独特性. 并且，通过对宋词文本的音律情感可视化与文本意象的关联分析，让文学研究者拥有多维度的视角去了解宋词文本的特性，不仅仅是分析文本本身，而是支持词作与词人的年谱和时代背景进行交叉验证. 与宋代文学研究专家合作，通过对他们进行用户实验研究，证明了该方法可以有效地帮助专家对词人的文学生涯及作品进行深入分析，并探索出新的研究视角，为今后的研究提供了更广阔的分析基础.

关键词：文学可视分析；宋词可视化；文本关联；时空可视分析
中图法分类号：TP391. 41　　**DOI:** 10.3724/SP.J.1089.2019.17970

A New Perspective on the Study of Literature (Songci): Text Correlation and Spatio-Temporal Visual Analytics

Zhang Wei[1)], Tan Siwei[1)], Liu Kai[1)], Shi Lei[1)], Chen Siming[2,3)], and Chen Wei[1)*]

[1)] (*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou*　310058)
[2)] (*Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn*　53113)
[3)] (*Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin*　53757)

Abstract: The work of traditional historical and literary studies analyzes the specific literary works and background of writers in a case-by-case way. The content, emotion, and meaning of a literary work is closely related to its author's life experience and the background. However, it is difficult to understand the above-mentioned connections by traditional methods, which may not address on the overall pictures. In this work, we propose a tailored text correlation and spatio-temporal visual analytics methods for classical literature, and we take Songci (a representative literature style in Song Dynasty in China) as an example. Those methods allow literary researchers analyze the trajectories of poets and the relevance and uniqueness of topic of literatures in different time period and their living background. Furthermore, through visual analysis of the rhythm, intentions, and expressions of emotions in the text, literary researchers can form multi-dimensional understanding. We cooperated with literary experts of Song Dynasty and conducted user study with them, which proves that the method we proposed can effectively help experts make an in-depth

收稿日期: 2019-07-01；修回日期: 2019-07-23. 基金项目：国家自然科学基金重点项目(61772456, 61761136020). 张　玮(1988—)，女，硕士研究生，主要研究方向为可视化；谭思危(1997—)，男，硕士研究生，主要研究方向为可视化；刘　凯(1999—)，男，在校学生；石　磊(1997—)，男，在校学生；陈思明(1989—)，男，博士，研究员，主要研究方向为可视化；陈　为(1976—)，男，博士，教授，博士生导师, CCF 会员，论文通讯作者，主要研究方向为可视化与可视分析.

analysis of the literary career and work of the poets. Moreover, our method can provide them a new insight in a comparative way.

Key words: literature visual analytics; visualization of Songci; text correlation; spatio-temporal visual analytics

宋词所代表的宋代文学，是中国文学史上璀璨的明珠之一. 宋词蕴含着古典化的审美、情感和精神，从中焕发出来的诗性、理性与人性历久而弥新. 然而，学术界对宋词的研究方法，主要是案例式地分析具体词作与词人的背景环境. 以郁玉英等[1]详细阐述宋词第一名篇《念奴娇·赤壁怀古》经典化的进程为例，研究者从词作内在的审美力量、词人超脱的品征意识与时代潮流的契合度等方面探析其经典化的原因.

案例式分析能够具体、深入地探索词作的文学特征，但在与文学研究者的讨论与合作中发现，对文学的研究需要关联分析. 这里的关联分析主要有 3 个层面，一是词的文本本身中意象、音韵层面的关联，二是不同词作之间的关联，三是词作背后词人的人生轨迹与家国天下、时代背景的关联. 这是当前文学研究所欠缺的、急需却又面临技术挑战的方面，也是本文需要解决与探索的方向. 首先对这 3 个层面的关联分析的需求，从可行性(存在性)、必要性以及挑战性 3 个方面进行论证.

(1) 宋词本身包含多种文学属性，如词名、词牌、情绪、意象及风格流派等，是一个高维的、极具发掘潜能的文化价值体系. 其意象、音韵等关联天然存在，这为探索关联性提供了基础. 而且，宋词文本中的意象、音韵等元素都与作者想表达的情感息息相关. 因此，研究者需要一种直观的方式进行音韵与情感的关联分析和探索，但宋词的多维特征给关联分析带来了较大的挑战.

(2) 在诗词创作中，作品的主题往往是相关联的. 同一词人在同一时期的词作，其主题会有一定程度上的相关性和独特性，词作主题的整个变化过程是符合作者人生经历的. 另外，不同词人的词作之间也存在一定的关联性，具体体现在文本内容的化用和作者之间的“答赠”上. 探索其词作之间的相关关系，找到相关关系形成的原因、意象运用的演变和不同词人风格的对比具有很大的难度，需要算法筛选提供关联的建议，并结合研究者的专业知识进行研究.

(3) 词人一生中经历过的所有事件，构成了其创作作品的缘由和内在情感. 尤其是在文官治国的宋代，一个文学家如果与政治事件相关，他的人生轨迹会和整个国家命运有一定的关联. 因此，本文分析宋词文本，还需要关联时代大背景下作者与他人政治、朋友、文学关系，以及某个具体的政治事件对词人造成的影响. 在文学研究中，分析文学作品的外部数据，是学者比较棘手的地方，这涉及复杂的时空数据的分析与整理. 同时相比于白话文中直白的情感抒发，诗词中通过意象抒发情感，表达非常含蓄，往往要结合大量外部资料进行分析，所以现成的情绪分类方法的效果并不好. 如果研究者独自查阅纸质文献，或通过数据库进行对比排查，海量文本数据会是另一个巨大的挑战.

为了支持以上 3 个层面的宋词文本关联分析，本文提出了一套可视分析方法，包含了词人人生轨迹与时代背景的时空可视分析、文本与意象比较可视化及音律和文本关联可视分析 3 大部分. 该可视分析方法让文学研究者拥有多维度的视角，更有效地对词人的文学生涯及作品进行综合分析. 在时空可视分析中，文学专家能对词人的人生起伏有直观的概览，观察词人轨迹和创作词作的地点，进行目的性的选择与词的比较，进而通过设计的词云对比视图，呈现不同时期词的文本与意象之间的相似与异同. 针对用户感兴趣的词，本文进一步提供音韵、意象的细节分析. 应用“五度标记法”将宋词音律可视化，同时用间隔编码意象，用颜色编码情绪，让学者以多维度的视角研究宋词文本的特性. 通过寻找相似的意象用法和相关的词人进行进一步分析，同时映射到时空历史场景与时代背景中，完成分析的闭环. 在这个过程中，文学专家能够对词作的意象和相关性进行深入的分析，获得新的研究视角.

1 相关工作

本节从 3 个主要方面回顾了一些相关工作，即文本可视分析、时空可视分析及文学作品分析与可视化.

1.1　文本可视分析

文本可视分析包括文本预处理和可视化 2 个主要步骤. 在文本预处理阶段, 与英文文本相比, 中文文本的分词面临更大的困难[2]. 尽管目前分词①②领域已取得不错的成果[3-4] , 但这些成果主要针对现代汉语, 若应用在文字、词汇和语法等诸多方面与现代汉语有所不同的古代汉语上则稍显不足[5]. 《汉籍电子文献》借助隐马尔可夫模型对先秦时期部分文献进行分词与词性标注[6], 石民等借助条件随机场模型对《左传》进行分词与词性标注[7]. 本文以古文字典为中间媒介, 使用 Jieba 分词器对古文文本进行分词处理, 并取得了较为满意的结果.

文本可视分析作为一种高效的数据分析方法, 已得到广泛应用, 可视化技术的发展也已较为成熟. InfoSky[8]是一种经典的静态可视化方法, 它将文档集中的文档投影成银河系中的恒星, 展现各个主题之间的层次和结构关系. TagCloud[9]是一种经典的基于词频的文本可视化分析方法, 它以一定规律(如词频)将关键词排列, 其大小则表示词的重要性或频度. ThemeRiver[10-13]是一种经典的以静态的方式表示时序动态信息的文本可视化方法, 它以河流为隐喻, 将时间比作从左到右流淌的河水, 用不同的色带代表不同的主题. 王臻皇等[14]使用微博数据, 支持交互地编辑改进主体模型.

本文的工作中使用了词云等通用文本可视化形式, 但与已有工作不同之处在于, 本文侧重于文本的相关性和文本背后词人的相关性分析.

1.2　时空可视分析

时空可视分析方法针对带有地理信息的时变数据, 设计一个自然、直观的交互式系统, 帮助人们更全面、准确地理解复杂空间信息的时间变化规律. 至今时空可视分析手段已得到广泛应用.

张金秋等[15]设计了一个基于点和热度图的交通数据可视分析平台, 兼顾通用性和专业性, 为交通时空数据提供多样化、可视化的分析手段. VAiRoma[16]基于维基百科文本构建相关的时空可视化, 让研究者更深入地了解罗马历史和历史人物. VAUD[17]是一个可视化分析系统, 通过迭代探索 来分析城市事件. Li 等[18]提出一种语义、空间、时间的 SpaceTimeCube, 帮助用户理解语义在空间和时间维度上的变化.

与以往工作不用的是, 本文将词人人生起伏结合其行迹进行呈现; 同时支持时间的筛选与对比, 对不同时期的作品进行时空和文本层面的比较.

1.3　文学作品分析与可视化

在文学研究中, 对文学作品尤其是经典作品的文本分析, 是一种很重要的研究内容和形式. 刘俐俐③从白先勇的《游园惊梦》文本的案例式分析着手, 详细阐述其互文性的分析方法. 宋词在中国文学史上享有很高的地位, 许多研究人员从不同角度进行研究. 在宏观上, 王兆鹏[19]以时间为线索, 结合时代环境变化和抒情范式变革, 将宋词的发展划分为 6 个阶段. 何阿珺④以李清照为例, 深入分析了人生遭遇对词人词风的影响, 并重点分析了其常用意象“酒”、“花”的丰富意蕴.

然而, 可视化分析作为一种高效的多维复杂数据分析手段, 至今尚未广泛应用到宋词研究领域中[20]. 但是在目前数字人文领域, 已经有非常多优秀的历史信息化的项目, 例如, 中国历史人物传记数据库⑤(China biographical database project , CBDB)提取整理了近百万历史人物时空事件和关系信息, 张建立等[21]将 GIS 技术应用到文学作品的空间分析中, 为文学研究提供了新的视角. 基于用户认知模型, 胡华华[22]采用主题河流图展现宋词风格流变、采用基于地理位置的空间分布图展现词人的行迹信息等, 以便从宏观上整体把握宋词风格流变的历史进程.

与以往工作不同, 本文将词人的人生地理轨迹和宋朝的政治历史环境, 与词作文本数据的意象、情绪、音律等信息进行统一编码, 构建一个结合文学内外部数据、对文本关联进行深层分析的可视分析系统, 在可视分析领域和宋词研究领域都具有一定的贡献.

2　数据与任务

本节首先描述数据集. 在此基础上, 阐述了用历史分析方法描述分析任务的方法和步骤, 总结了实现分析任务的设计要求.

① https://github.com/fxsjy/jieba
② http://lib.cqvip.com/Qikan/Article/Detail?id=28744089
③ http://lib.cqvip.com/Qikan/Article/Detail?id=8772838850484856484949484853
④ http://lib.cqvip.com/Qikan/Article/Detail?id=28759067
⑤ https://projects.iq.harvard.edu/cbdb

2.1 数据抽象

本文数据源如表 1 所示，包括复旦中华古诗词数据库[①]、唐宋文学编年地图[②]以及 CBDB. 复旦中华古诗词数据库是主要数据来源，包含从周朝至宋朝的所有留存古诗词数据，共有 20 万篇，本文对这些数据进行文本分析. 唐宋文学编年地图是中南民族大学王兆鹏教授的课题，包含唐宋共 70 位文学家详细的年谱、轨迹、作品以及作品创作时间地点，本文用这些数据作为宋词的外部信息. CBDB 包含中国古代历史人物的个人信息、社会血缘关系和结构化的年谱事件，本文同样将这些数据作为宋词研究的外部信息.

表 1 宋词数据

维度	属性	规模
时间	宋词的写作时间	公元 900—1300 年
地点	宋词的写作地点	约 300 个
人物	宋词的作者 人物之间的关系	约 3 000 位
文本	宋词的内容 宋词的词牌名	约 25 000 首
背景事件	宋词的背景事件	约 80 万

2.2 需求分析

在为期半年的研究过程中，与 3 位中国古代文学专家进行合作，并定期与他们举行会议，确定设计需求，迭代系统设计，并评估最终的系统.

在需求调研阶段，中国古代文学专家讲解了他们的日常工作流程，包含大量的文献检索和交叉对比验证工作，这会耗费研究者大量的时间. 其次，专家们提到，文学研究中比较难的是处理文学外部的数据，如在整个历史背景下词人和他人的关系、国家政治决策对他是否产生影响，以及他所经历的事件和人生与其作品之间的联系等. 同时，在对宋词文本进行研究时，其中的音律、情绪、意象无法直观地呈现. 这些问题阻碍了专家们进行更深入的研究，因此本文确定了 3 项分析任务.

(1) 呈现词人的生平经历以及时代背景，支持对词人生平和家国时代背景的关联分析与交叉验证. 文学家对一个词人一生所经历的兴衰很感兴趣，而这种兴衰往往是由多方面因素造成，需要结合其经历的事件及所处的历史背景来进行分析.

(2) 探索词人不同时期作品风格演变. 一个词人的作品风格随着时间地点的变化而演变. 找出该词人各个时期词作的特点并在文本层面与文本写作的时间空间层面探索其异同，有助于文学家对词人风格演变特征进行深入的探究.

(3) 在意象的相关层面，理解词的语言运用和化用，更深入地了解一首词所表达的内在情感.

文学家需要一种新颖的方法，结合多方面的信息对词文本进行更多维更深入的分析，探寻词的内在情感. 这种内在感情包括意象、情绪、音律；同时，文学家需要了解词的化用关系，进而探讨词作风格的相互影响及背后的历史环境、社会关系.

2.3 设计目标

从以上可视分析任务出发，总结可视分析系统的设计目标为以下 4 点.

(1) 词人的年谱和时空轨迹可视化. 提供词人的不同方面的资料概览，包括人生起伏、行迹、不同时期地点词作数量. 这同时也可以作为词作的分析入口.

(2) 可视化对比词人不同时期地点的词作风格. 支持时间的筛选与对比，并对不同时期的作品进行时空和文本层面的比较. 文本上需要理解相似与不同的意象和主题.

(3) 宋词文本属性可视化. 可视化编码宋词的音律、意象、情绪，让用户直观地理解音律与意象、情绪的关联.

(4) 宋词文本关联与可视化. 支持研究者查看词作中的化用关系，以及在此基础上的词人社交关系. 视图间提供交互操作，让文学家可以结合多源信息，对词作与词人年谱和时代背景交叉验证分析.

3 数据整合和模型

本节主要介绍图 1 所示数据的处理流程和建模，包括文本处理和背景事件的整理，其最终目的是帮助用户了解一个词人的文学生涯.

3.1 背景事件提取

事件描述人物在特定时间地点的动作，它们会影响词人作品表达的感情和含义. 年谱由人物相关的事件串组成，事件串的地点组成了词人的轨迹，本文采用人工制定的规则，从 CBDB 和文学编年地图中提取事件，构成词人年谱，将每首词所作的时间和地点范围内的事件作为这首词的背景事件.

① https://github.com/chinese-poetry/chinese-poetry

② https://sou-yun.cn/poetlifemap.aspx

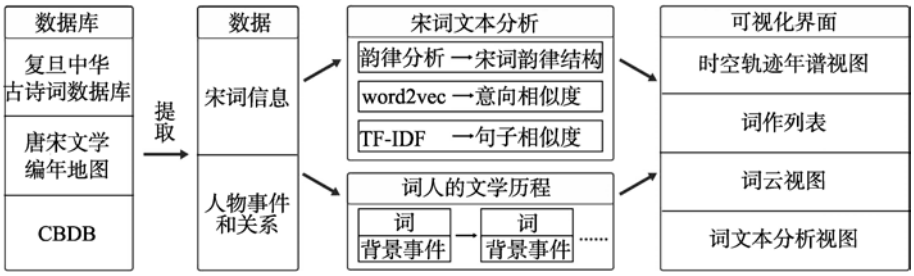

图 1　数据处理流程

3.2　意象的实体发现

意象是诗词的实体性要素, 词人往往使用作品中的意象来表达和寄托情感, 而且词人使用意象方式常常有相似性. 但是, 很多古代的实体名称并没有被收录到现在分词库的词典中, 所以诗词难以用现有的词典进行分词, 提取意象. 本文采用 2 种方法添加新词, 一种是现有的宋词词典, 如把《唐诗宋词鉴赏辞典》[5]中的词作为新词, 第二种是使用基于信息熵的新词发现算法.

根据中文词的特性, 将文本字符串中字数小于 4 的子串都作为备选单词, 先筛掉频率小于阈值的字符串, 然后用单词的左右信息熵和词内部的点互信息作为 2 个标准衡量这个单词是否为新词.

左右信息熵指备选单词的左边的字符串和右边的字符串的信息熵, 信息熵的公式为

$$H(X)=-\sum_{x\in X}p(x)\,\mathrm{lb}\,p(x) \tag{1}$$

信息熵可以反映出信息量, 所有左右信息熵可以反映一个词左右搭配的词的丰富程度, 如果搭配越丰富, 则越有可能是一个单词.

单词内部的点互信息(pointwise mutual information, PMI)为

$$\mathrm{PMI}(x,y)=\mathrm{lb}\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

PMI 越大, 表示单词内部的字越有联系. 将备选单词的 PMI 和单词的左右信息熵都归一化后, 用三者的和进行排序, 取前 600 个词作为新词, 最后将 2 种方法结合得到的新词添加到词典中, 使用 jieba[6]对宋词进行分词, 再筛去由专家指出的虚词, 得到文本的实词, 将其作为诗词包含的意象.

3.3　文本相似度计算

为了帮助用户了解宋词的流派和化用情况, 本文使用词频-逆文本频率(term frequency inverse document frequency, TF-IDF)来分析宋词的意象运用, 并计算句子之间的相似度. TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术[4], 计算文本中的词在表达文本主题时占得的权重, 并通过比重计算文本的相似性. 设文本 $\boldsymbol{Q}$ 和文本 $\boldsymbol{W}$ 中每个词 j 的 TF-IDF 值所组成的向量分别为 $\boldsymbol{Q}=(t_{11},t_{12},t_{13},\cdots,t_{1n})$ 和 $\boldsymbol{W}=(t_{21},t_{22},t_{23},\cdots,t_{2n})$, 它们之间的相似性可以用余弦距离来表示, 即令

$$l=\cos\langle\boldsymbol{Q},\boldsymbol{W}\rangle=\frac{\boldsymbol{Q}\cdot\boldsymbol{W}}{|\boldsymbol{Q}||\boldsymbol{W}|} \tag{3}$$

距离 l 越大, 相似性越大. TF-IDF 值计算公式为

$$t_{ij}=f_{ij}\times I_{ij} \tag{4}$$

其中, i 为文本的序号, j 为词的序号; f_{ij} 为词频(TF), I_{ij} 为逆文本频率(IDF), 其公式为

$$I_{ij}=\mathrm{lb}\frac{|D|}{|j:t_i\in D_j|} \tag{5}$$

$|D|$ 是语料库中文件总数; $|j:t_i\in D_j|$ 是包含词语 t_i 的文件总数.

3.4　音律提取

词的音韵和其文本的美感、情绪密切相关. 根据专家的建议, 本文采用平水韵①和词林正韵②来分析宋词的音律. 现代的文学家在研究宋词的字的音律时常从调值和韵母这 2 个点入手, 调值展现了一个字内的音高升降, 组成了整首宋词朗诵的抑扬顿挫, 这种起伏往往与作者的情绪波动有关; 韵母则和押韵有关, 影响了宋词朗读的优美性. 在平水韵中, 所有的字分成了 106 部, 词林正韵中则分为 14 部. 平水韵中每一部都对应特定的调值及韵母, 而在词林正韵中属于同一部的字之间才会押韵. 因此, 按照规则得到宋词中每个字的调值, 同时将距离小于 8 个字内韵部相同的字看成押韵.

① http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZSGY200910002005.htm

② http://lib.cqvip.com/Qikan/Article/Detail?id=670140364

4 可视分析系统

根据任务和设计目标，本文提出并实现了一个基于宋词文本关联和时空可视分析的系统，包含时空轨迹视图、文本比较视图和文本细节分析视图．可视分析的流程是从词人生平变迁的总览入手，对时空进行筛选探索并比较不同时期的词作风格变化；最后对感兴趣的词文本进行关联分析，找到其相似的词和词人背后的历史因素，在时空场景下进行交叉验证，形成可视分析的闭环．

4.1 概览．万水千山走遍——结合词人生平变迁的地理可视分析

时空可视分析视图是词人生平及其词作历史背景分析的入口，包括词人的基本信息、轨迹以及年谱信息．本文将这 3 个信息分成 3 个子视图：基本信息视图、地图、人生起伏视图．图 2 中区域基本信息视图罗列了词人的基本信息，包括姓名、别名、生卒、身份以及作品数量等．

4.1.1 地图视图

图 2 中 B 区域地图视图用于分析词人的行动模式．采用宋代行政地图，地图视图上的圆点标注词人停留的地点，圆的大小编码词人在该地作词的数量多少．同时本文希望将词人的轨迹和年谱映射在一起，方便用户将两者对照分析．由于同时连接上百个地点和几十个年份会产生视觉的遮挡，因此将词人按照地点和时间化成几个人生阶段．把事件的时间 t、地点经纬度(x, y)组成向量(x, y, t)；然后使用 k-means 将其聚成 k 个人生阶段，只将人生阶段的首尾所在的地点通过连线映射到人生起伏的时间轴上，这样用户就可以了解到词人在不同人生阶段停留的地区．使用 k-means 聚类的依据是根据人物在某地长时间停留，就会留下相对多的词作，这些词作往往反映了词人当前的人生阶段．当选择一个地点时，用户可以看到途经该地点的路径，也可以通过地点和人生起伏视图上的连线，看到地点对应的时间．

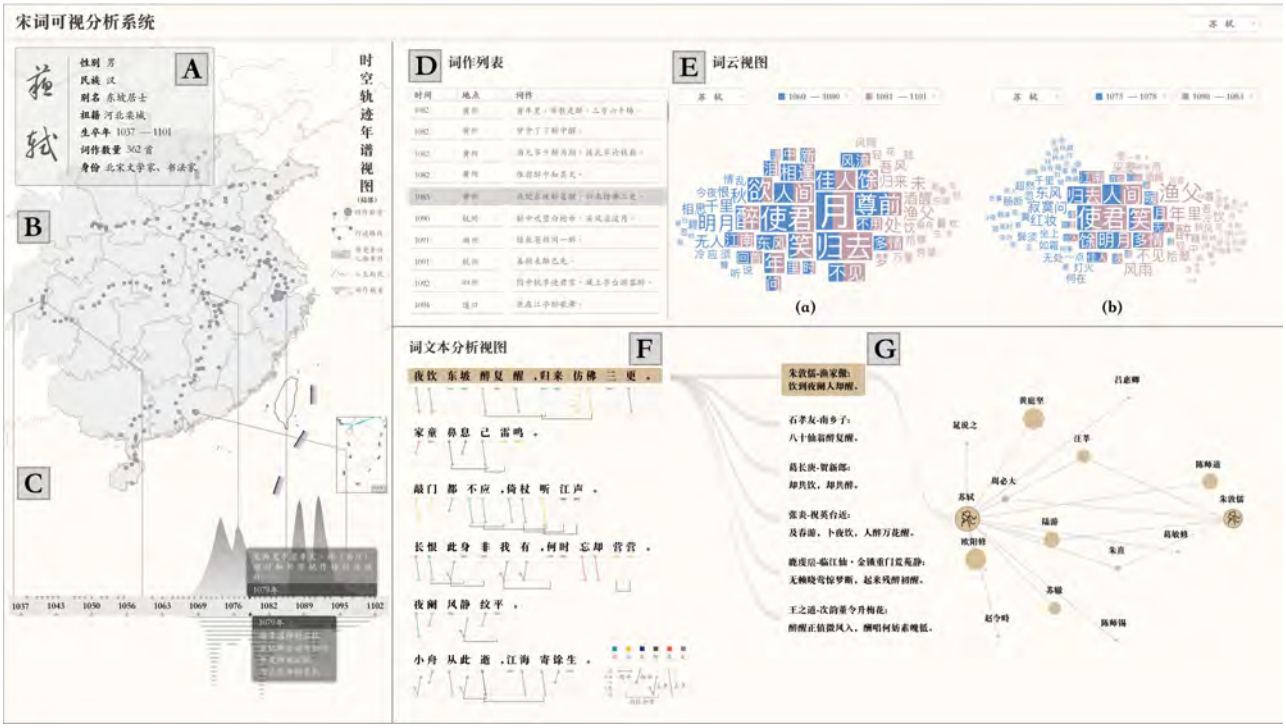

图 2 系统主界面

4.1.2 人生起伏视图

图 2 中 C 区域人生起伏视图呈现词人人生的总览，用山水画中的山以及山在水中的倒影隐喻人生起伏和词作数量变化．以坐标轴为界，上部的“山脉”折线图代表人生起伏，下部的“倒影”折线图代表词作数量变化．词作数量和人生起伏具有一定关系，因此本文的设计支持两者结合分析．人生起伏的计算参照基于词典的情感分析算法，也就是对事件标签进行人工打分然后加权求和，人工打分由合作的文学研究者完成，权重为事件标签在所有人物年谱中的逆文本频率 I_{ij}．图中上升部分往往由登科、升官决定，而下降部分由贬官、被

弹劾决定．人生起伏时间轴上部的三角代表时代背景事件，如皇帝登基、外交、政治军事等，用于支持外部事件的关联分析．用户可查看某时刻词人发生的事件和时代背景事件，同时地图上会高亮词人当时所处的地点．用户也可以选择时间段，地图也会高亮出对应的轨迹．

当用户在时空视图上选择时间段或者地区时，宋词列表会列出词人在该范围内所作的词．用户也可以选择 2 个区域进行比较，则进入探索的下一阶段：词风的对比与趋势分析．

4.2　探索对比．寻章摘句——文本意象比较

文本属性综合视图呈现词人作品的总体情况，包括宋词列表、词云视图．图 2 中 D 区域宋词列表简单地罗列了词作的信息，包括时间、地点和部分词作内容．它让文学家以他们最熟悉的方式浏览宋词内容．用户选择诗词后可以在文本视图上进行分析；同时，当用户对意象、时间或者地点进行筛选时，列表上仅显示筛选后的结果．

图 2 中 E 区域词云视图显示了用户选择 1 个或 2 个时期的作品词频，当词云中包含 2 个时期的词频集合 t_1 和 t_2 时，词云会变成类似韦恩图的分布，左侧蓝色的字为 $(t_1-t_2)\bigcap t_1$ 的词频，中间方块中的字为 $t_2\bigcap t_3$ 的词频，右侧则是 $(t_2-t_1)\bigcap t_2$ 的词频．方块的颜色比例编码了 2 个时期词频的比例．用户可以选择某一关键词，然后通过宋词列表浏览包含这个词的所有词作．视图还支持对不同词人意象运用的比较．

4.3　草木皆有情——文本细节与关联可视分析

用户选择一首词后，可以通过文本可视分析视图研究其音律、感情和语言的运用．文本分析视图主要分成 2 个部分，图 2 中 F 区域情感音律视图呈现了词本身的研究，G 区域的语言运用视图则给用户提供了探索语言运用的相似性的入口．

4.3.1　情感音律视图

情感音律视图中，宋词的内容按句子分行显示，单词之间存在空格，辅助用户断句．如图 3 所示，文本下方的短线段编码字音节的调值，即朗诵时声调的起伏．颜色编码意象的情绪，情绪是通过 word2vec[23]得到的向量值，与“喜怒哀乐思”5 个字求相似度得到，对于与情绪词都不相近的词，颜色设成灰色．现有的情绪分析算法对古文文本的分析效果并不好，但是因为意象是情绪的载体，word2vec 结果中意象围绕情绪词聚类的现象非常明显．

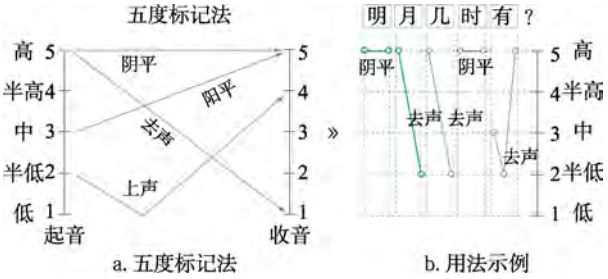

图 3　情感音律视图编码

4.3.2　语言运用视图

语言运用视图旨在帮助用户分析当前词作和其他词作中相同的语言运用，研究这种运用在环境大背景下的共性和特点．同时，本文也希望能以语言应用为入口，发现文学上的“团体”．如果用户选择其中一个句子作为基准进行深入研究，视图上会显示所有与该词句相似的句子、出处和作者，用户可以选择其中某个词句．本文希望用户能将难以理解的语言共性和社会关系结合起来分析，去探索词派和文学交流．视图采用了力导向图展现与 2 个作者共同相关的人的关系网络，关系网络中圆的大小编码该作者与基准词句作者之间相似句子的数量；初始只显示 2 人最短路径，用户选择路径上的人物后，隐藏的关系才会显示出来，用户可以选择 2 个人物，在词云视图上比较他们的词风．

5　案例分析

本文通过 2 个案例来介绍系统的可用性和效率．第 1 个案例介绍如何通过系统研究词人的文学历程，第 2 个案例介绍如何对一首词进行文学分析．

5.1　案例 **1.** 由外向内的苏轼的文学生涯分析

本文的第 1 个案例选择的是苏轼．虽然目前有很多关于苏轼的文学研究，但是这些研究都比较片面，往往是从 1~2 首诗分析入手再结合一个时期的背景进行研究，缺乏总体的依据．因此，本文希望能通过系统从外部数据入手探索苏轼的文学历程．

5.1.1　纵览苏轼

进入图 2 所示苏轼的诗词分析界面，用户首先看到苏轼的词云，左侧是苏轼创作生涯前 20 年的作品词频，右侧是后期的作品词频．可知“月”、“人间”、“东风”等是贯穿苏轼整个创作生涯的意向．进一步地，用户可以探索包含这些意向的词作，如选择“东风”后，宋词列表将展示包含“东风”的句子，用户可以选择其中一句，如图 4 所示“今岁花时深院，尽日东风，荡扬茶烟”．通过文本分析视

图进行文本分析，其颜色显示“东风”常与思念有关，且两字押韵，朗朗上口. 用户可预测“东风”常被用于寄托苏轼思念之情，但为何常用“东风”，而非“春风”等同义词，其原因是它押韵，更加符合诗词含蓄和音律上的美感.

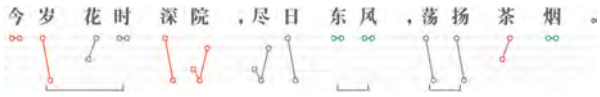

图 4 词文本分析

如图 2 的 E 区域中 a 图所示，用户进一步通过词云视图了解苏轼人生前后半段的词频变化会发现，前 20 年，“相思”、“明月”、“千里”、“风流”、“情”这些词频繁出现，显然这些词与情爱相关，勾勒了苏轼青年时期意气风发的形象；而在后 20 年，“渔父”、“酒醒”、“梦”、“扁舟”、“芳草” 等词大量出现，贯穿一生的同时又在渐渐增加“归去”2 字，都叙述着一个渴望归隐，经历风霜的寂寞形象.

研究者可以浏览苏轼的年谱和轨迹来更进一步地了解苏轼. 如图 2 中 B 区域地图所示，苏轼的路径呈“7”字形，按照人生阶段逐个浏览，苏轼早年由家乡四川入仕来到中原，定居 30 载. 在这里他起起伏伏，即使 1080 年左右跌落谷底也没有离开. 直到 1095 年前后，苏轼离开中原南下海南. 再从图 2 中 C 区域苏轼年谱视图可以看到，他前半生最得意的时候是在 1076 年. 通过点击图中的最高点，得知此时他身在山东，刚到任密州知州，正是仕途蒸蒸日上之时. 而在 1079 年左右，他的人生轨迹急转直下，令人感兴趣的是，作品数量却迅速增加. 阅读背景事件可知该年乌台诗案发生，他遭到大量攻讦，因言获罪，被贬离京. 用户想看在这时间前后他的词作风格的变化，通过图 2 的 E 区域中 b 图对比 1075—1078 年和 1080—1083 年的词频可以看到，乌台诗案之前，他的词中多有“东风”、“千里”、“问”等抒发豪迈气逸和政治豪情的文字；之后，词中多“风雨”、“醉”、“流水”等质朴意象，从中可以看出他的作品逐渐从少年的喟叹，转向中年的无奈，转向自然本真.

5.1.2 夜饮东坡醒复醉

为了更深入地分析苏轼的词风变化细节，用户选择了一首包含“醉”的词作《夜饮东坡醒复醉》，可以看到图 2 中 F 区域的视图，该词作于 1083 年，距离苏轼在乌台诗案后被发配至黄州已有数载，如图 2 中 C 区域的人生轨迹图所示，他正处在人生地位的低谷期和创作量的高峰期，这也可以侧面体现出作者怀有大量情绪需要发泄.

从音律的颜色可以发现，此词开篇多用喜、乐、思 3 种情绪，都是积极或者中性的情感词，声调节奏相对缓和，很少出现快速的起伏升降，渲染了悠长的氛围，但是实际上这却是以乐景衬哀情. “敲门都不应，倚杖听江声”这 2 句反复使用同韵字和去声字，如音乐上的重音符号般，加强了音韵上的激昂之感. 下阕开篇抒发怀才不遇之情，充满了无奈，而结于叹息之声，最后 2 句中性词与消极情绪词交杂，更加强了上阕强颜欢笑之感.

从图 2 中 G 区域语言运用视图可以看到，“夜饮东坡醒复醉”存在大量化用. “饮”、“醉”、“醒”渲染虚幻迷离的气氛，将浓浓情感融于“醒”、“醉”这 2 个重复的动作中. 用户再选择朱敦儒的《渔家傲》，进一步探索发现这个小团体存在书信往来、互相作序这样的文学交流，用户可以推断 2 人属于同一流派. 为了验证猜测，用户通过查阅外部资料，确定朱敦儒作为“洛中八俊”之一，是文学上从“苏轼时代”到“辛弃疾时代”的过渡者.

5.2 案例 2. 稻花香里说丰年——由内向外的词的深入挖掘

本文的第 2 个案例选择的是辛弃疾. 目前对辛词的研究多从感情、写作手法等入手来分析词的整体风格，而少有从音韵结构入手. 少数涉及音韵的研究也难以用更加整体的眼光来看辛词的词风. 因此，本文希望通过系统获得更多、更整体的收获.

以辛弃疾词作《西江月 · 夜行黄沙道中》为例，图 5 中 a~d 区域展示了整个分析流程.

5.2.1 文本分析

本文选择图 5 中 a 区域所示《西江月 · 夜行黄沙道中》进行分析. 观察声调发现，此词用韵非常巧妙. 首句大量押韵，读起来十分顺畅. 词的 6 个韵脚中，由图中可见“蝉”、“年”、“前”、“边”是都是阴平，“片”、“见”为上声. 根据相关音律要求，不可平仄通押；而在此词中，作者不以韵害意，偏偏打破格律，以仄韵压平韵. 尽管韵拗，但上下句相拗处一致，反而使得声调铿锵、韵味悠扬. 与此同时，图中可以看到在片内常用阴平、阳平和去声这样的声调，很少有上声这种声调快速起伏的音节出现，而且上下片又几乎对称. 所以朗读起来比较平缓，能够烘托出全词意境上的宁静悠远. 再观察音节上的颜色可以看出，这首诗中喜乐交杂.

5.2.2 背景分析

了解词的情感音律后，通过图 5 中 b 区域所示

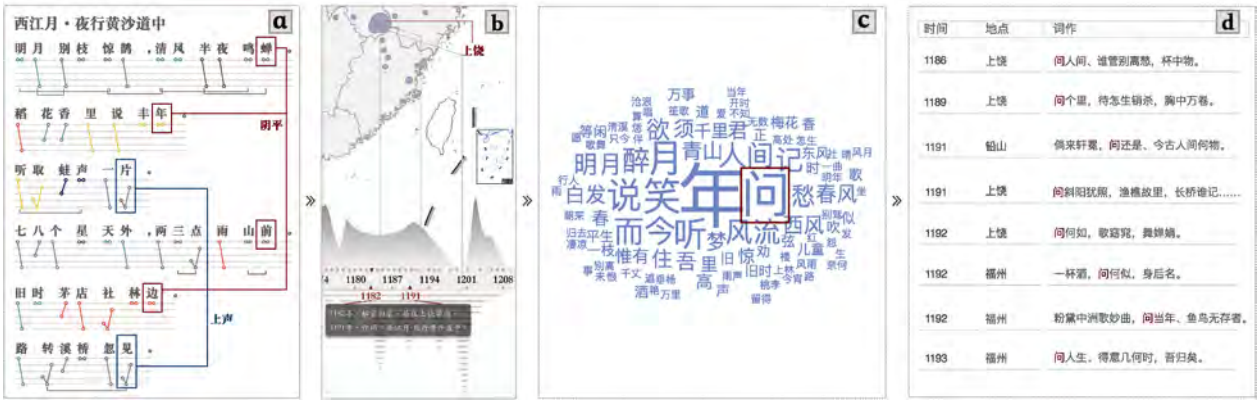

图 5　由内向外的词的深入挖掘流程图

年谱轨迹视图可知，该词作于 1191 年，即辛弃疾罢官后隐居上饶湖之时．选取这段时间里的词作，图 5c 可以看出辛弃疾常用“青山”、“春”、“月”等描述闲适的意象，以及“年”、“老去”、“白发”等感慨年纪的词．其中一个“问”字尤其突出，用户可以选择“问”字，通过图 5 中 d 区域所示宋词列表，看到辛弃疾常问，他问清山、问人间、问天上、问刘郎，也许他更想问自己，反思自己为何不被重用，只能空老去．

6　专家反馈

与 3 位中国古代文学家在为期半年的沟通合作中不断获取反馈，并进行系统迭代，最后由 2 位文学家实际操作，对系统做出最终评价．

6.1　操作过程

由于本文的设计并不复杂，专家了解各个视图的设计、交互及功能后，很快熟悉整个系统，开始实际操作．其中一位文学家发现苏轼的词中常运用与自己其他诗句相似的意象，仿佛在“自己抄自己”．她重点探索了这些相似句子在不同语境下的含义．如吟“不应有恨”和“月有阴晴圆缺”的《水调歌头》于 1076 年作于密州，当时乌台诗案未发，苏轼的人生轨迹图呈上升趋势．而 1096 年作于惠州的《三部乐》中，苏轼写道“算应无恨，安用阴晴圆缺”．专家表示结合背景知识，她能体会到苏轼对年轻时自己的调笑．在此 20 年间，他的心态由年少意气变平静旷达，词风由豪气转到清新自然．她还说，传统研究很难发现这种例子，所以一直难以找到足够的数据支撑他们研究词句化用现象，但在本文的系统中仅几分钟就能找到四五个案例，未来她希望能针对以上发现做更深入的研究．

6.2　可视设计

专家们在每个视图中都有感兴趣的设计点，对人生起伏视图，一位专家提到，“系统中包含时间地点在内的多维信息，虽然人生起伏的计算方式不能为本文的研究提供理论依据，但的确能在研究过程中提供启发和快速研究的入口”．整个系统中他们对词云比较视图和诗词的音律编码设计的印象尤其深刻，他们提到在研究一个人的词风变化时，文学家们常需花大量时间细分不同时段和地域的词作，再将它们的内容一一对比，这是一个非常烦琐的任务．而现在的词云能让他们迅速地找到时段间作品的特性和共性，而且这种特性和共性常常体现了他们在研究中需要探讨的点．在音律视图中，他们认为将调值、韵位、意象情绪和化用结合在一起的设计简洁明了，足以满足他们在文本研究上的需求．

他们也指出了一些不足之处，在对一些特殊多音字的处理上，本文没有进行编码，同时他们也希望在地图上能包含更多的信息，如经济、交通地图．

6.3　交互设计

专家们对交互流程非常认可，认为系统逻辑严密，足以支持每一步的推理论证．在实际操作过程中，他们也能灵活地使用，从各个视图中找到想要的信息，同时能将自己的专业知识与交互过程相融．他们建议在视图中添加一些调参功能，如设置词云中显示的词的数量和判断文学化用相似度的阈值．还有一位专家觉得系统中针对语句化用的分析可以更加复杂一些，同时希望能进行更多的群体性研究．

7 结 语

本节总结了与文学家合作中学到的经验和工作的局限性以及对于未来工作的展望.

7.1 与文学家合作的经验

与文学家的整个合作过程中发现，对文学领域的研究方向和研究方式最初存在一定误解，专家的需求也与预想有所不同. 通过交流发现，文学的优美性并不是最受关注的，他们更希望从诗词中提炼出情感、史实作为线索，考证当时的社会环境. 在看待数据的方式上，也有非常大的不同. 作为计算机科学的相关研究者，对大规模的数据，对模式、趋势更感兴趣，希望能把数据量化，文学家虽然也对这种研究方式及模式、趋势和团体感兴趣，但是他们都曾表达出对算法是否适合处理复杂文学数据的担忧，因为他们目前流行和认可的研究方式都是偏感性的，更善于对一个小点进行深入细致的研究.

因此这是一个有挑战的任务. 每次处理和整合数据前都要和专家确认方法的合理性，提炼出符合他们研究方法的交互逻辑. 最后在专家评估环节，他们反映这个工作的确能帮他们在研究词作时节省时间，提供全新的视野. 一位博士后提到，她从未想过竟可以通过这样的方式研究宋词.

7.2 局限和未来的工作

在词作的情感分析中，现有的交互方式无法辅助用户寻找意象和背景信息之间的联系，导致用户在浏览诗词的相关事件时会阅读到无关信息，所以在未来将使用一些信息学的方法帮助用户整理，并呈现词作和事件之间的联系.

在词的文本分析中，本文以 2 人之间语言运用相似性为入口，以此发现词作内容相近的作者所属的文学流派，但系统缺乏总体性的视图，用户必须对周围的词人单独调查，因此在体现团体性上证据略显不足. 所以未来将会思考如何实现双向关系、甚至多向关系的呈现，将所有词和作者通过散点图的形式呈现，并通过聚类、主题词展示的方式，帮助用户分析多种风格、多种团体的形成及原因.

7.3 总 结

本文提出了一种结合文本关联、人物关系与时空轨迹进行综合分析的全新的文本可视分析的思路. 通过与文学家的紧密合作，构建针对词人创作经历和作品综合可视分析的系统，并提出一种新的可比较词语使用情况的词云设计；同时通过五度标记法将宋词的音律和韵位分布可视化，通过自动化处理，减少烦琐信息的搜索整理时间，极大地提高了用户的效率. 而且，通过基于相似意象用法的文学相似性和社会关系的交叉分析，系统为文学研究者提供了一种基于统计和语言相似性的研究视角.

参考文献(References):

[1] Yu Yuying, Wang Zhaopeng. The analysis of the canonization about the 1st classic of Songci NIANNUJIAO·CHIBIHUAIGU [J]. Qilu Journal, (6): 115-121(in Chinese)
(郁玉英, 王兆鹏. 宋词第一名篇《念奴娇·赤壁怀古》经典化探析[J]. 齐鲁学刊, 2009, (6): 115-121)

[2] Yuan Hai, Chen Kang, Tao Caixia, *et al*. Research on visualization techniques based on Chinese texts[J]. Telecommunications Science, 2014, 30(4): 114-122(in Chinese)
(袁 海, 陈 康, 陶彩霞, 等. 基于中文文本的可视化技术研究[J]. 电信科学, 2014, 30 (4) : 114-122)

[3] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality[C] // Proceedings of the 26th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2013, 3111-3119

[4] Salton G, Buckley C. Term weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1987, 24(5) : 513-523

[5] Qiu Bing, Huangfu Juan. Study on the trend of ancient Chinese words based on the word automatic segmentation[J]. Microcomputer Information, 2008, 24(24): 100-102(in Chinese)
(邱 冰, 皇甫娟. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息, 2008, 24(24): 100-102)

[6] Qian Zhiyong, Zhou Jianzhong, Tong Guoping, *et al*. Research on automatic word segmentation and pos tagging for *Chu Ci* based on HMM[J]. Library and Information Service, 2014, 58(4): 105-110(in Chinese)
(钱智勇, 周建忠, 童国平, 等. 基于 HMM 的楚辞自动分词标注研究[J]. 图书情报工作, 2014, 58(4): 105-110)

[7] Shi Min, Li Bin, Chen Xiaohe. CRF based research on a unified approach to word segmentation and POS tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2): 39-45(in Chinese)
(石 民, 李 斌, 陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010, 24(2): 39-45)

[8] Andrews K, Kienreich W, Sabol V, *et al*. The InfoSky visual explorer: exploiting hierarchical structure and document similarities[J]. Information Visualization, 2002, 1(3/4): 166-181

[9] Vi é gas F B, Wattenberg M. TIMELINE: tag clouds and the case for vernacular visualization[J]. Interactions, 2008, 15(4): 49-52

[10] Byron L, Wattenberg M. Stacked graphs-geometry and aesthetics[J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1245-1252

[11] Dork M, Gruen D, Williamson C, *et al*. A visual backchannel

for large-scale events[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1129-1138

[12] Havre S, Hetzler E G, Whiteney P D, *et al*. ThemeRiver: visualizing thematic changes in large document collections[J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 9-20

[13] Cui W W, Liu S X, Tan L, *et al*. Textflow: towards better understanding of evolving topics in text[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2412-2421

[14] Wang Zhenhuang, Chen Siming, Yuan Xiaoru. Visual analysis for microblog topic model[J]. Journal of Software, 2018, 29(4): 1115-1130(in Chinese)
(王臻皇，陈思明，袁晓如. 面向微博主题的可视分析研究[J]. 软件学报, 2018, 29(4): 1115-1130)

[15] Zhang Jinqiu, Zhao Shuxu, Qu Ruitao. Visualization of traffic spatio-temporal data based on point and heatmap[J]. Journal of Lanzhou Jiaotong University, 2017, 36(3): 63-69+75(in Chinese)
(张金秋，赵庶旭，屈睿涛. 基于点与热度的交通时空数据可视化[J]. 兰州交通大学学报, 2017, 36(3): 63-69+75)

[16] Cho I, Dou W, Wang D X, *et al*. VAiRoma: a visual analytics system for making sense of places, times, and events in roman history[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 210-219

[17] Chen W, Huang Z S, Wu F R, *et al*. VAUD: a visual analysis approach for exploring spatio-temporal urban data[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(9): 2636-2648

[18] Li J, Chen S M, Chen W, *et al*. Semantics-space-time cube: a conceptual framework for systematic analysis of texts in space and time[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, (99): 1

[19] Wang Zhaopeng. On the historical developing periods of song ci[J]. Journal of Jinan University: Philosophy & Social Science Edition, 2000, 22(6): 14-23+29(in Chinese)
(王兆鹏. 论宋词的发展历程[J]. 暨南学报：哲学社会科学, 2000, 22(6): 14-23+29)

[20] Tang Jiayu, Liu Zhiyuan, Sun Maosong. A survey of text visualization[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(3): 273-285(in Chinese)
(唐家渝，刘知远，孙茂松. 文本可视化研究综述[J]. 计算机辅助设计与图形学学报, 2013, 25(3): 273-285)

[21] Zhang Jianli, Li Renjie, Fu Xueqing, *et al*. Spatial information analysis and visualization analysis of the ancient poetry[J]. Journal of Geo-Information Science, 2014, 16(6): 890-897(in Chinese)
(张建立，李仁杰，傅学庆，等. 古诗词文本的空间信息解析与可视化分析[J]. 地球信息科学学报, 2014, 16(6): 890-897)

[22] Hu Huahua. Visualization design and research of the style and sects change of Songci[D]. Harbin: Harbin Institute of Technology, 2018(in Chinese)
(胡华华. 宋词风格流变的可视化设计研究[D]. 哈尔滨：哈尔滨工业大学, 2018)

[23] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space[OL]. [2019-07-01]. https://arxiv.org/abs/1301.3781