计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190641
57(6): 1252-1268, 2020

古诗词图谱的构建及分析研究

刘昱彤 吴 斌 白 婷
(北京市智能通信软件与多媒体重点实验室(北京邮电大学) 北京 100876)
(北京邮电大学计算机学院 北京 100876)
(liuyutong@bupt.edu.cn)

The Construction and Analysis of Classical Chinese Poetry Knowledge Graph

Liu Yutong, Wu Bin, and Bai Ting
(*Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia* (*Beijing University of Posts and Telecommunications*), *Beijing* 100876)
(*School of Computer Science*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876)

Abstract Classical Chinese poetry is a precious cultural heritage. It is significant to use the rich information in classical Chinese poetry to further investigate the language, literature and historical development of Chinese culture. However, the knowledge of classical Chinese poetry is highly fragmented. It not only exists in poetry itself, but also is widely distributed in the materials which are used to explain poetry, such as annotations, translations, appreciations, etc. Our aim is to obtain the potential semantic relationship between words and expressions, and use knowledge graph to link them. By doing this, we could integrate fragmented knowledge in a systematic way, which enables us to achieve better reasoning and analysis of classical Chinese poetry knowledge. In this paper, we propose a method to construct classical Chinese poetry knowledge graph (CCP-KG). About building nodes of CCP-KG, we use the improved Apriori algorithm to generate candidate words, then check if the candidate word appears in the annotations to determine when it can be a node of CCP-KG. About building edges of CCP-KG, the semantic relationship between words is established through the annotations, then we use the artificially constructed classical Chinese poetry hierarchical structure to establish the relationship between abstract semantics. Finally, we obtain CCP-KG, which covers every aspect of classical Chinese poetry and contains multi-layer semantic links between words. Taking Tang poetry as an example, CCP-KG can be used to analysis classical Chinese poems in different dimensions. Compared with character-based data analysis, the use of CCP-KG assists literary research more in-depth from the perspective of semantics. Therefore, CCP-KG is necessary in analyzing classical Chinese poems. In addition, CCP-KG can also be applied to various tasks like reasoning and analysis in classical Chinese poetry. We conduct experiments on the tasks of determining the theme of poetry and analyzing the emotion of poetry respectively, showing the effectiveness and application value of our constructed CCP-KG.

Key words digital humanities; classical Chinese poetry; knowledge graph; data analysis; reasoning analysis

收稿日期:2019-09-02;**修回日期**:2019-11-29

基金项目:国家重点研发计划项目(2018YFC0831500);国家自然科学基金项目(U1936220,61972047)
This work was supported by the National Key Research and Development Program of China (2018YFC0831500) and the National Natural Science Foundation of China (U1936220, 61972047).

通信作者:吴斌(wubin@bupt.edu.cn)

摘 要 古诗词是中国宝贵的文化遗产.利用计算机对诗词进行辅助研究,对语言、文学、传承普及中华文化,具有重要意义.然而,关于诗词的知识是高度碎片化的,原因是互联网上的诗词知识,不仅存在于诗词本身,还分布于诗词的各种解读资料,比如诗词的注释、译文、赏析等.若以知识图谱的方式,捕捉古诗词中词语之间潜在的语义联系并将它们以知识的方式关联起来,能够将诗词碎片化的知识有条理地整合在一起,从而更好地对古诗词知识进行推理和分析.基于此,提出了一种古诗词知识图谱的构建方法.构建图谱的节点时,首先利用改进的 Apriori 算法产生诗词中的候选词,然后检验候选词是否出现在诗词注释和中文词典中,从而判断其是否构成图谱节点.构建图谱的边时,首先利用注释信息在词语之间建立语义联系,然后用人工构建的诗词分类体系在抽象的语义之间建立联系.最终得到一个内容覆盖全面且包含多层词语语义联系的古诗词图谱.古诗词图谱可用于对诗词各种不同维度的分析研究,相比于基于字的数据分析,利用古诗词图谱能够从语义的角度更加深入具体地辅助文学研究.以唐诗为例,说明了古诗词图谱在诗词分析中的必要性.此外,古诗词图谱还适用于各种关于诗词的推理和分析任务,以判定诗词题材和分析诗词情感这 2 个任务为例,证明了古诗词图谱的有效性和应用价值.

关键词 数字人文;古诗词;知识图谱;数据分析;推理分析

中图法分类号 TP391

古诗词在中国古典文学中占有极其重要的地位.随着数字人文的发展,使用计算语言学和统计学的方法辅助诗词的研究已成为一种普遍趋势.如今,关于诗词的知识是碎片化的,原因是互联网上可得到的诗词知识,一方面来自诗词本身,另一方面来自诗词的解读资料,比如诗词的注释、译文、鉴赏等.将古诗词中的词语通过语义的关联,以知识的形式联系在一起,是一种比较合理的将诗词的碎片化知识联系在一起的方式.因此,古诗词领域的知识图谱是将诗词的碎片化知识进行关联、整合的必要手段.

现有的诗词知识图谱,如“唐诗别苑”[1]和“宋代学术传承知识图谱”[2].前者仅涉及到诗人社交网络、诗人迁徙游历、作品热点地图这 3 个方面,而后者仅涉及宋朝人物师承关系.它们仅仅从几个特殊的方面来构建诗词知识图谱,但是忽略了各个部分之间的内部联系.显然,这些诗词知识图谱在内容完整性、结构合理性、条理分明性有所欠缺.关于诗词的分析工作,如唐宋文学编年地图[3]将诗人的时空轨迹可视化,也只停留在诗词写作地点和写作时间的人工整理以及对人物轨迹简单的统计分析上,并未涉及到诗词中实体的识别和实体之间语义关联的判定.为了解决以上问题,本文构建了一个内容覆盖全面且层次结构分明的古诗词图谱,通过从文本中挖掘词语的语义关联,将古诗中的词语用知识的形式组织起来.

现有的通用领域的中文知识图谱,如《大词林》、HowNet、CN-DBpedia 等,对古诗中涉及到的词汇覆盖率极低,不能采用抽取子集的方式直接构建古诗词领域的知识图谱.而常规的知识图谱构建过程对古诗词领域并不完全适用,由于诗句不遵循语法规则,无法用实体识别、实体关系抽取等常用自然语言处理技术构建.同时也无法将诗词中的词汇链接到现有百科,因为现有百科几乎没有收录古诗词的词语.

为了构建古诗词知识图谱,我们该如何找到古诗中的词语,又该如何将词语的语义之间建立联系.这项工作面临 2 个挑战:1)如何界定古诗词的词语,又如何准确地挑选出古诗词中出现的词语.由于词语构成古诗词图谱的节点,准确地将古诗词中出现的词语抽取出来是整个工作的基础.2)如何利用和融合互联网的多种诗词知识源来准确地获取词语之间的关联.针对第 1 个挑战,为了保证诗词词语抽取的准确性,一般采用人工标注的方式,但人工标注过于费时费力.因此,我们的解决方案是模拟人工标注时参考诗词注释和词典中词语解释的过程来自动获得诗词中的词语.一方面,我们利用诗词的注释,因为注释正是对不可再分的语义单元的解释,获取注释条目就精准地获取了诗词中的词语;另一方面,我们利用古汉语的新词发现方法[4-5]产生诗词中的候选词,然后在诗词注释条目和中文词典中查找,若出现,则该候选词是诗词中的词语,否则该候选词不是诗词中的词语.针对第 2 个挑战,我们融合诗词注释和中文词典中的词语解释,找到词语与词语之间的联系,再利用人工构建的古诗词分类体系建立语义之间的联系.

本文的主要贡献有 3 个方面:

1) 提出了一种古诗词图谱的构建方法,并利用

该方法构建了一个内容覆盖全面、包含多层词语语义联系的古诗词图谱.该图谱刻画了词语的多个层级,以合理的结构覆盖了诗词的各个方面;该图谱从诗词的注释和词语的词典解释入手,挖掘出词语的语义关联.

2) 古诗词图谱可以对诗词进行各种不同维度的分析.相比于基于字的浅层数据分析,利用古诗词图谱可以从语义的角度从真正的意义上辅助文学研究.以唐诗为例,展示了古诗词图谱在诗词分析上的应用,证明了古诗词图谱在诗词分析中的必要性.

3) 古诗词图谱适用于诗词的各种推理和分析任务.以判定诗词的题材、分析诗词的情感 2 个任务为例,说明了古诗词图谱的应用价值.

1 相关工作

1.1 数字人文及其在诗词中的研究进展

近年来,随着数字技术和数字媒体的不断发展,人文学科结合数字技术的研究应运而生.2011 年 Michel 等人[6]在《科学》杂志上发表了基于百万电子化图书对文化进行量化分析的论文,挖掘出 1800—2000 年的英文语言中所反映出的语言学和文化现象;2014 年 Schich 等人[7]在《科学》杂志上发表了量化分析文化中心变迁的论文,他们利用超过 15 万的名人的出生地点和死亡地点的信息,建立了跨度 2000 年的迁徙网络,用网络科学的手段分析了欧洲文化中心的变迁.这些工作为数字人文这一新兴的交叉领域学科提供了一个良好的开端.

数字人文在中国诗词上的研究,近年来涌现出大量的工作.王兆鹏[3]搭建了唐宋文学编年地图平台,将诗人的时空轨迹分布信息可视化.新华网联合浙江大学发布的“宋词缱绻,何处画人间”[8]平台,对宋词进行可视化展示,包括宋代词人游历路线、宋代词人生平及所处年代图谱、《全宋词》常见意象统计等.清华大学自然语言处理与社会人文计算实验室搭建了计算机诗词创作系统“九歌”[9],结合 seq2seq 神经网络结构和诗词韵律实现了古诗自动生成算法[10-11].但是,以上关于诗词分析的工作都是简单地对词频进行统计,尚未涉及到诗词中词语与词语之间的语义关联.

1.2 中文通用及诗词领域知识图谱的研究进展

中文通用领域的知识图谱主要有《大词林》、HowNet、CN-DBpedia,接下来分别介绍.《大词林》[12]是自动构建的基于上下位关系的大规模开放域中文知识图谱.HowNet[13]是人工标注的语言知识库.在 HowNet 中,最小的语义单位被称为“义原”.HowNet 中的“义原”大约有 2 000 个.HowNet 基于“义原”体系,共标注了数十万词汇的语义信息.CN-DBpedia[14]是复旦大学知识工厂实验室研发并维护的大规模通用领域结构化百科.CN-DBpedia 主要是从中文百科类网站(如百度百科、互动百科、中文维基百科等)的纯文本页面中提取信息,经过滤、融合、推断等操作后,最终形成高质量的结构化数据,供机器和人使用.

然而,通用领域的中文知识图谱对古诗词中涉及到的词汇覆盖率极低,不能够直接用来分析诗词的词语语义之间的联系.

关于知识图谱的构建方法,刘峤等人[15]在《知识图谱构建技术综述》一文中总结,知识图谱的构建过程包括信息抽取、知识融合、知识加工共 3 个阶段.其中,信息抽取包括实体抽取[16-18]、关系抽取[19-21]、属性抽取[22-23];知识融合包括实体链接[24-25]、知识合并[26];知识加工包括本体构建[27]、知识推理[28]、质量评估[29].可见,常规的知识图谱都是从抽取实体和关系、构建“实体-关系-实体”三元组开始的.

然而,常规的知识图谱构建过程无法适用于古诗词知识图谱的构建.原因是,诗句不遵循主谓宾的语法结构,可以从诗词中抽取出词语,但是利用现有的关系抽取方法无法提取出词语之间的关系,也无法确定词语之间表达的语义是否相关.而且,无法利用各种百科中的实体信息,因为现有的中文百科尚未收录古诗词中的词语.

诗词领域的知识图谱方面,主要的工作有:“唐诗别苑”[1],是关于全唐诗语义检索的可视化平台,实现了唐诗的语义检索功能和知识图谱的可视化(包括诗人社交网络、诗人迁徙游历、作品热点地图、诗人属性 4 个方面).“宋代学术传承知识图谱”[2],从“中国历代人物传记数据库”[30](China biographical database project, CBDB)中抽取宋代人物之间的学术传承关系和部分亲属关系,用知识图谱来探究宋代士人的师承关系.周莉娜等人[31]设计了基于唐诗知识图谱的智能知识服务平台 KnowPoetry,提供唐诗领域的知识探索、时空轨迹、语义查询等智能化知识服务.

以上提到的诗词领域的知识图谱,是以诗人、诗歌为单位的,并未从词语的角度对诗词的知识进行建模和分析.而诗词语义的基本组成单元是词语,若要对诗词进行更深层次的语义研究,对诗词词语的

语义进行知识上的关联是无法避开的一步.此外,以上工作构建的诗词知识图谱仅从多个角度包含了诗词知识,图谱涉及到的内容不够全面,图谱的组织不够有条理.

2 古诗词图谱

2.1 古诗词图谱的形式化定义

古诗词图谱 $G=(V,E)$,由点的集合 V 和边的集合 E 组成.点的集合 $V=\{V_{\text{hier}},V_{\text{desc}},V_{\text{anno}}\}$,包含3种类型的词节点,分别是古诗词分类体系中的词 V_{hier}、描述词 V_{desc}、注释条目的词 V_{anno}.边的集合 $E=\{E_{\text{hier-hier}},E_{\text{hier-desc}},E_{\text{desc-desc}},E_{\text{desc-anno}}\}$,包含4种类型的边,分别是 V_{hier} 与 V_{hier} 之间的边、V_{hier} 与 V_{desc} 之间的边、V_{desc} 与 V_{desc} 之间的边、V_{desc} 与 V_{anno} 之间的边.

2.2 古诗词图谱的构建

古诗词图谱的构建过程如图1所示,分为节点的构建、边的构建2部分.

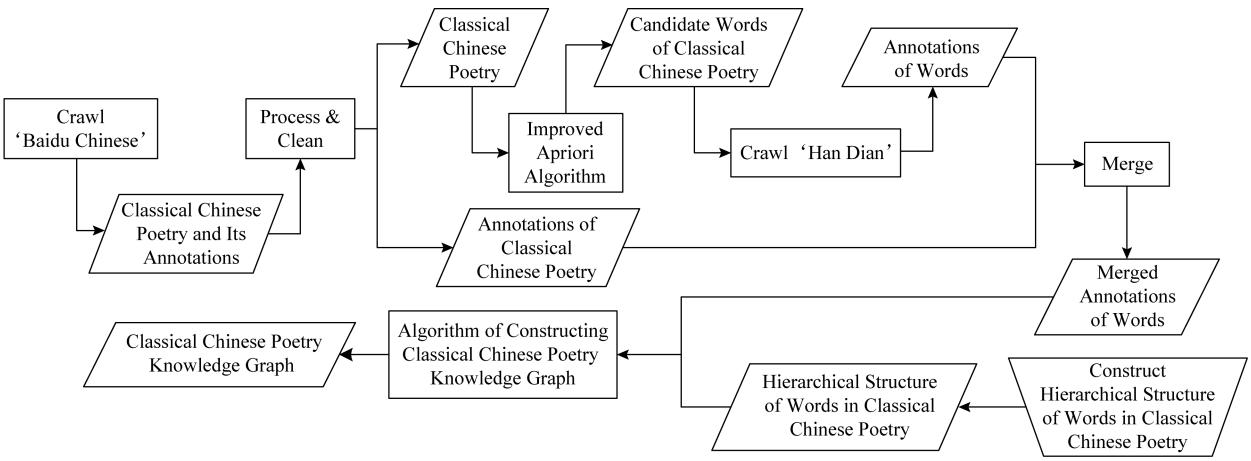

Fig. 1 The process of constructing classical Chinese poetry knowledge graph

图1 古诗词图谱构建流程图

2.2.1 古诗词图谱——节点的构建

诗词中出现的词语构成古诗词图谱的节点,所以准确地抽取出诗词中的词语非常关键.若人工标注诗词中的词语,当判断某个字串是否是词语时,参考的资料一般有2个来源:1)查看诗词的注释,若某个字串包含注释信息,那么该字串就是词语,因为注释是对不可再分的语义单元的解释.2)根据常识、使用平时的积累来判断某个字串是否是词语,其本质就是判断该字串是否是现代汉语的词语,这个过程可以用查询中文词典的方式完成.人工标注的方式过于费时费力,所以本文模仿人工标注的方式来自动获取诗词中的词语.首先利用古汉语新词发现算法中改进的 Apriori 算法[4-5]产生候选词,该算法可以不遗漏地挖掘出长度在 $2\sim K$ 之间的所有可能词语;然后搜索候选词是否出现在诗词注释条目和中文词典中,若出现则判定该候选词为诗词中的词语,这一步模拟了人工标注的过程.总的来说,这种方式全自动地、准确地得到了诗词中的词语,即图谱的节点.接下来对改进的 Apriori 算法[4-5]进行详细说明.

算法 1. 改进的 Apriori 算法.

输入:原始语料 D;

输出:候选新词的集合.

/* 产生 1-频繁项集 */

① 1-候选项集 $C_1=\{c_1,c_2,\cdots\}$;

② 统计原始语料中单个字的出现频次,得到(字,频次)的二元组集合 $S_1=\{(c_1,f_1),(c_2,f_2),\cdots\}$;

③ for (c_i,f_i) in S_1

④ if $f_i>$支持度

⑤ 把 c_i 加入 1-频繁项集 L_1;

⑥ end if

⑦ end for

⑧ 得到 1-频繁项集 L_1;

/* 产生 2-频繁项集 */

⑨ for c_i in L_1

⑩ for c_j in L_1

⑪ 把 c_i+c_j 加入 2-候选项集 C_2;

⑫ end for

⑬ end for

⑭ 统计原始语料中 C_2 的每个字符串 s_i 出现的频次,得到(字符串,频次)的二元组集合 $S_2=\{(s_1,f_1'),(s_2,f_2'),\cdots\}$;

⑮ for (s_i, f_i') in S_2
⑯ if $f_i' >$支持度
⑰ 把 s_i 加入 2-频繁项集 L_2;
⑱ end if
⑲ if $f_i' <$低频阈值
⑳ 把 s_i 加入低频项集 M;
㉑ end if
㉒ end for
㉓ 得到 2-频繁项集 L_2;
㉔ for $k=3$ to K /* 产生 k-频繁项集 */
㉕ for $p_1p_2\cdots p_{k-1}$ in L_{k-1}
㉖ for $q_1q_2\cdots q_{k-1}$ in L_{k-1}
㉗ if $p_2p_3\cdots p_{k-1}=q_1q_2\cdots q_{k-2}$
㉘ 把$p_1p_2\cdots p_{k-1}q_{k-1}$加入 C_k;
㉙ end if
㉚ end for
㉛ end for
㉜ 统计原始语料中 C_k 的每个字符串 s_i' 出现的频次,得到(字符串,频次)的二元组集合 $S_k=\{(s_1', f_1''), (s_2', f_2''), \cdots\}$;
㉝ for (s_i', f_i'') in S_k
㉞ if $f_i'' >$支持度
㉟ 把 s_i' 加入 k-频繁项集 L_k;
㊱ end if
㊲ if $f_i'' <$低频阈值
㊳ 把 s_i' 加入低频项集 M;
㊴ end if
㊵ end for
㊶ 得到 k-频繁项集 L_k;
㊷ end for
㊸ return $L_2 \cup L_3 \cup \cdots \cup L_k \cup M$.

2.2.2 古诗词图谱——边的构建

古诗词图谱的边刻画了古诗词中词语的语义联系,所以利用互联网存在的多种知识源来准确地获取词语之间的关联非常重要.现阶段关于诗词词语语义相似度的工作[32-33]只能挖掘出频繁共现的词,并认为它们是相似的.但是,我们的目标是准确地将具有相同主题的词相关联.比如西晋名将羊祜的典故有多种表达:"羊公碑"、"岘山"、"征南"、"泪碑","羊公",这些词语之间是具有语义联系的.利用完全无监督的方法很难获取到这样的语义联系,而诗词的注释和词典的词语解释是对语义的权威释义,利用词语的权威解释可以准确有效地挖掘出词语的语义联系.因此,本文基于诗词注释和中文词典中的词语解释,在相互关联的词语之间建立了联系,但是,语义和语义之间仍然是不相关的.为了进一步在不同的语义之间建立联系,我们设计了古诗词分类体系,以多层分类结构将不同的语义有条理地组织在一起,从而形成完整的古诗词图谱.

如图 1 所示,构建古诗词图谱的边可大致分为 3 个步骤:

1) 获取和处理注释信息.爬取"百度汉语"网站,获取诗词注释.爬取"汉典"网站,获取诗词候选词的注释.然后将 2 部分注释信息分别处理清洗,之后合并.

2) 构建古诗词分类体系.构建关于古诗词的分层体系,得到 4 棵根节点分别是"时间"、"地点"、"景物"、"人"的树.

3) 构建古诗词图谱.将注释条目的词通过描述词关联到诗词分类体系的词,得到古诗词图谱.

2.2.2.1 获取和处理注释信息

关于诗词的注释,我们爬取"百度汉语"① 网站来获取.百度汉语的注释示例如图 2(a)所示.在该例中,"黄鹤楼"、"孟浩然"、"广陵"、"故人"、"烟花"、"唯见"是所需要的词语,而"碧空尽"、"天际流"这样的词并不是我们所需要的."碧空尽"中的"碧空","天际流"中的"天际"也是所需要的.因此,首先,我们用正则表达式把所有的注释处理成(词,解释)的二元组形式,图 2(a)的例子就包含了 14 个这样的二元组(除了图 2(a)展示的 11 个,还包含"尽"、"碧空"、"天际").然后把所有诗的相同词的注释合并,以特殊符号"|"分隔.

关于候选词的解释,我们爬取"汉典"② 网站来获取.汉典中的词语解释示例如图 2(b),对于每个词,我们爬取的是"解释"和"国语辞典"2 个部分.

最后,将百度汉语和汉典这 2 个渠道获得的词语解释合并.

2.2.2.2 构建古诗词分类体系

董乃斌在《中国文学叙事传统论稿》一书中提到"诗歌和其他各种文学作品一样,都是要通过'叙述'来表达的.时间、地点、景物、人物、事件,包括作者内心的感情,统统都得由作者或口头或书面地叙述出来.抒情其实也是一种叙述,不过所叙的是感情或

① https://hanyu.baidu.com/
② https://www.zdic.net/

情绪,且主要是作者本人的感情或情绪而已."[34]

具体而言,人处于特定的时间、地点,受到眼前景物的触发,再结合自身的背景和经历,会产生不同的心情、抒发不同的感情和人生感悟.所以本文设计的古诗词分类体系是围绕时间、地点、景物、人这 4 个方面展开的,如表 1 所示,受篇幅限制,只展示了部分分类体系.古诗词分类体系是由我们邀请的 3 位古汉语文学专业的专家利用 3 天时间手工构建的.由于构建古诗词分类体系涉及到诗词的常识知识,且没有现成的古诗词领域常识图谱,因此构建过程必须人工参与.构建完成古诗词分类体系,就创建了 V_{hier} 和 $E_{\text{hier-hier}}$.

(a) Baidu Chinese (b) Han Dian

Fig. 2 Example of annotations in Baidu Chinese and Han Dian

图 2 "百度汉语"和"汉典"的注释示例

Table 1 Hierarchical Structure of Words in Classical Chinese Poetry

表 1 古诗词分类体系

Level-1	Level-2	Level-3	Level-4	Level-5
时间	节日、季节	元宵、人日、端午、冬至、社日、春、夏、秋、冬……		
地点	山、水、边塞、建筑、名胜、行政区划	峰、湖、溪、河、池、西域、塞外、边关、沙漠、亭、斋、台、馆、游览胜地、名楼、河北、河南、湖北、湖南……		
景物	天气、日月星辰、植物、动物	雨、雾、霜、日、月、星、树、花、草、谷物、鸟、昆虫、哺乳动物、冷血动物	柳树、桑树、梅花、兰花、兔丝、女萝、飞蓬、粟、麦、粱、麻、黍、乌鸦、燕子、蝴蝶、蜜蜂、蚂蚁、蝉、猿猴、狗、兔、蛇、龟、蟾蜍……	
人	基本需求、基本属性、知识储备	衣、食、住、行、玩、神话、历史、心情、经历、品质、人生大事、社会地位	帽子、短衣、头巾、酒、菜、饭、肉、被子、枕头、床、轿子、马车、船、书法、下棋、长寿、祥瑞、夏朝、商朝、西周、东周、喜悦、哀伤、送别、羁旅、孝顺、守信、弱冠、登科、贫民、官宦……	黄帝、唐尧、虞舜、伯夷、叔齐、姜子牙、召伯、周文王、周武王……

2.2.2.3 构建古诗词图谱

古诗词分类体系是古诗词图谱的一部分,本节讲述如何将注释条目加入古诗词分类体系中,构成最终的古诗词图谱.算法 2 描述了该过程.算法 2 的主要思路是:找到与古诗词分类体系的叶子节点相关的若干个描述词,在叶子节点和描述词之间建立联系,在描述词之间建立联系.对某个描述词遍历所有的注释条目,若该描述词出现在某注释内容中,则在该描述词与该注释条目之间建立联系.

算法 2. 构建古诗词图谱算法.

输入:古诗词分类体系 $Trees=\{Tree^1,Tree^2,Tree^3,Tree^4\}$、词语解释集合 $word\_anno\_set=\{(word^1,anno\_list^1),(word^2,anno\_list^2),\cdots,(word^n,anno\_list^n)\}$;

输出:古诗词图谱 G.

① $G=Trees$;
② for $tree$ in $Trees$
③ for $node$ in $tree.leave\_nodes$ /* $node$ 有多个描述词:$desc^1,desc^2,\cdots,desc^m$ */
④ for i in $range(m)$ /* 对每个描述词 */
⑤ $node'=Add\_node(desc^i)$; /* 将 $desc^i$ 作为节点加入图 G */
⑥ $Add\_edge(node,node')$; /* 在 $node$ 和 $node'$ 之间构建连边 */
⑦ for $(word,anno\_list)$ in $word\_anno\_set$
⑧ if $desc^i$ in ''.$join(anno\_list)$
⑨ $Add\_edge(node',word)$;
⑩ end if
⑪ end for /* 遍历词语解释集合,若 $desc^i$ 出现在某个词语的解释中,就在 $desc^i$ 节点与词语节点之间构建连边 */
⑫ end for
⑬ for i in $range(m)$
⑭ for j in $range(m)$
⑮ $node\_i=Get\_node(desc^i)$;
⑯ $node\_j=Get\_node(desc^j)$;
⑰ $Add\_edge(node\_i,node\_j)$;
⑱ $Add\_edge(node\_j,node\_i)$;
⑲ end for
⑳ end for /* 在两两描述词之间构建连边 */
㉑ end for
㉒ end for
㉓ return G.

下面通过例子来解释这一过程.如图 3 所示,"风"是诗词分类体系中根节点为"景物"的树的一个叶子节点,它的父亲节点是"天气".与"风"这一节点有关的描述词有"狂风"、"刮风"、"大风".步骤 1,在"风"与"刮风","风"与"大风","风"与"狂风"之间构建 1 条边.这一步骤的含义是同一事物有多个描述词.步骤 2,在"刮风"、"大风"、"狂风"两两之间构建 1 条边.这一步骤的含义是事物的不同描述词是等价的、地位相同的.步骤 3,遍历所有的注释条目,分别找到这 3 个描述词出现在注释内容中的注释条目.比如,"狂风"分别出现在了"冲风"、"惊飙"、"惊风"这 3 个词的注释中,那么,就将"狂风"与"冲风","狂风"与"惊飙","狂风"与"惊风"之间各构建 1 条连边.这一步骤的含义是如果某个词出现在了某条注释中,就说明该注释条目与该词内容相关.

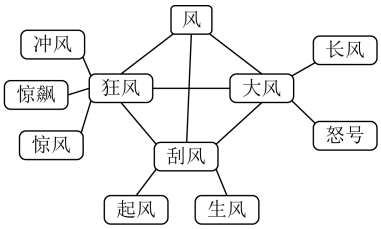

Fig. 3 Example of the process of constructing classical Chinese poetry knowledge graph

图 3 古诗词图谱构建过程举例

2.3 古诗词图谱构建的评测

以唐诗为例,构建古诗词图谱.在古诗词图谱的节点的构建过程中,关于改进的 Apriori 算法,K 设置为 3,支持度设置为 5,低频阈值设置为 2.在古诗词图谱的边的构建过程中,"百度汉语"中可以检索到的唐诗共 40 531 首,包含注释信息的共 1 330 首,通过处理、清洗、合并后共得到 7 668 条注释信息.通过改进的 Apriori 算法产生的候选词,在"汉典"中能够找到解释的共有 18 589 个.将"百度汉语"和"汉典"这 2 个渠道获得的词语解释合并后,共得到 21 907 条词语和与之对应的解释.最终,构建的古诗词图谱共有 6 619 个节点、10 292 条边.其中,单个字节点有 231 个,2 字词节点有 6 116 个,3 字词节点有 259 个,4 字词节点有 13 个.

为了评测构建的古诗词图谱的性能,对诗词中涉及的古诗词图谱的词语个数进行统计,如图 4 所示.统计的范围是唐诗中的五言诗和七言诗,共

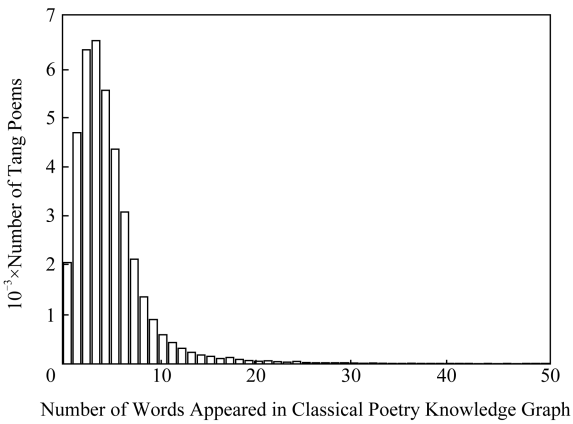

Fig. 4 Statistics of number of poems' words in classical Chinese poetry knowledge graph

图 4 诗词涉及的古诗词图谱的词语数量统计

40 210 首,从图 4 中可以看到,只有大约 2 000 首诗没有包含古诗词图谱的词语.大部分诗词涉及的古诗词图谱的词语个数在 0～10 之间.而大部分唐诗是五言或七言的律诗或绝句,也就是说大部分唐诗的字数是几十个字,因此,我们构建的古诗词图谱的词语覆盖率是可观的.

3 利用古诗词图谱对诗词进行分析

本节以唐诗为例,说明古诗词图谱在诗词分析中的作用.

3.1 不同因素对诗人情感的影响

本节探究了季节、天气、地点对诗人情感的影响.利用古诗词图谱,把与季节、天气、地点、情感有关的词分别关联起来,从更广的词的范围对“季节-情感”、“天气-情感”、“地点-情感”进行全面、具体地数据分析,从而得到更深层次的结论.

3.1.1 季节对诗人情感的影响

欲探究不同的季节对诗人的情感产生什么样的影响,我们将统计描述季节的词和描述情感的词在诗中的共现关系.

对于季节 S_i,我们需要分别统计 S_i 和诗人的 9 种情感(哀伤、失意、愁绪、喜悦、孤独、恐惧、愤怒、怨恨、惊讶)之间的共现次数,即季节 S_i 和情感 E_j 共现的诗的个数($j=1,2,\cdots,9$).然后,计算情感 E_j ($j=1,2,\cdots,9$)对季节 S_i 所占的比例:

$$E_j \text{ 对 } S_i \text{ 所占比例}=\frac{\#(S_i,E_j)}{\sum_{j'}\#(S_i,E_{j'})}\times 100\%,\quad(1)$$

其中,$\#(S_i,E_j)$表示季节 S_i 和情感 E_j 共现的诗的个数.

如果没有古诗词图谱,在统计诗词描述的季节时,我们只会检验“春”、“夏”、“秋”、“冬”这 4 个字是否出现在诗中.但是,有时诗中有关季节的表达却不止这 4 个字,比如,“霜天”、“归雁”、“玉露”、“红叶”也同样表明是秋天.在古诗词图谱中,“春”、“夏”、“秋”、“冬”是较为抽象的词语,属于古诗词分类体系中的词语,由它们向下延伸可以拓展出很多相关词语.描述情感的词同理,在古诗词图谱中,“哀伤”、“失意”、“愁绪”等也属于古诗词分类体系中的词语.利用古诗词图谱对季节和情感共现次数进行统计,可以得到更加全面、真实的结果.

图 5 展示了 4 种季节中每一种情感所占的比例.从图 5 可以看出,春、夏、秋、冬 4 种季节占比最高的情感分别是喜悦、哀伤、哀伤、哀伤.关于春天和秋天,这是符合常识的,春天万物复苏、生机勃勃,容易使人心情愉悦并产生积极向上的情绪;而秋天萧索悲凉,万物枯败,容易使人联想到自己的衰老和不济,产生悲伤的情绪.关于夏天和冬天,我们发现,这 2 个季节的主要情感也是哀伤.而且,这 4 个季节中,哀伤、失意、愁绪、孤独这 4 种负面情感占据了非常大的比例,而恐惧、愤怒、怨恨、惊讶占据比例很小.因此,我们可以得出结论,大部分诗词表达的情感都是负面的,但是春天是容易让诗人的情感转变为积极的一个重要因素.

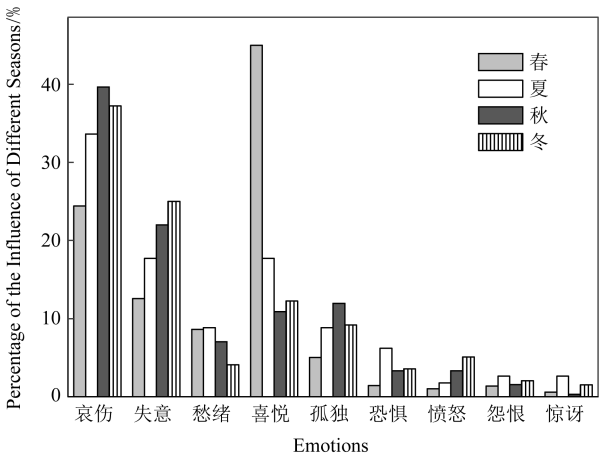

Fig. 5 The histogram of the influence of different seasons on poets' emotions

图 5 不同季节对诗人情感的影响柱形图

3.1.2 天气对诗人情感的影响

欲探究不同的天气对诗人的情感产生什么样的影响,我们将统计描述天气的词和描述情感的词在诗中的共现关系.

类似地,计算情感 E_j($j=1,2,\cdots,9$)对天气 W_i 所占的比例:

$$E_j \text{ 对 } W_i \text{ 所占比例}=\frac{\#(W_i,E_j)}{\sum_{j'}\#(W_i,E_{j'})}\times 100\%,\quad(2)$$

其中,$\#(W_i,E_j)$表示天气 W_i 和情感 E_j 共现的诗的个数.

利用古诗词图谱可以发现,描述“雪”的词还有“飘素”、“玉蕊”、“天花”、“玉英”、“玉花”等.描述“雨”的词还有“蹄涔”、“丁丁”等.描述“风”的词还有“摧折”、“长条”等.利用古诗词图谱进行统计,比仅仅对诗词统计“雪”、“雨”、“风”、“露”、“云”这 5 个字更加全面.

图 6 展示了 5 种天气中每一种情感所占的比例.从图 6 可以看出,哀伤、失意、愁绪、喜悦、孤独仍然占比很大,恐惧、愤怒、怨恨、惊讶仍然占比很小,

结合 3.1.1 节可以说明,前 5 种情感在整个唐诗集合中占比很大,后 4 种情感在整个唐诗集合中占比很小.接下来,通过比较在特定情感下 5 个代表天气柱的相对高度来解析不同天气对诗人情感的影响.

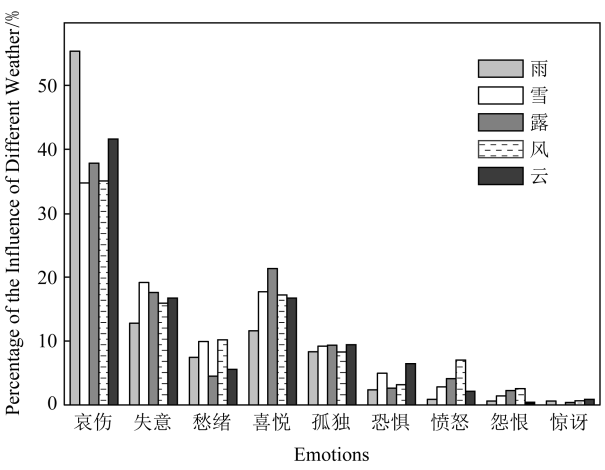

Fig. 6 The histogram of the influence of different weather on poets' emotions

图 6 不同天气对诗人情感的影响柱形图

在图 6 中,"哀伤"在"雨"这种天气中的占比显著高于"哀伤"在其他天气中的占比,"喜悦"在"雨"这种天气中的占比显著低于"喜悦"在其他天气中的占比.这说明了"雨"这种天气相比于其他天气,更加倾向于带给人哀伤的情绪,换句话说,诗人更愿意用"雨"来渲染诗歌的悲伤氛围.此外,"恐惧"在"云"中的比例显著高于"恐惧"在其他天气中的比例."愤怒"在"风"中的比例显著高于"愤怒"在其他天气中的比例.这说明了诗人更倾向于使用"云"这样的天气来表达"恐惧"的情绪,更倾向于使用"风"这样的天气来表达"愤怒"的情绪.

3.1.3 地点对诗人情感的影响

欲探究不同的地点对诗人的情感产生什么样的影响,我们将统计描述地点的词和描述情感的词在诗中的共现关系.

类似地,计算情感 $E_j(j=1,2,\cdots,9)$ 对地点 P_i 所占的比例:

$$E_j \text{ 对 } P_i \text{ 所占比例}=\frac{\#(P_i,E_j)}{\sum_{j'}\#(P_i,E_{j'})}\times 100\%, \quad (3)$$

其中,$\#(P_i,E_j)$ 表示地点 P_i 和情感 E_j 共现的诗的个数.

诗词中关于地点的描述十分复杂,没有统一的规则.比如典型的描述"边塞"的词语有"轮台"、"碛西"、"楼兰"、"龙城"、"北庭"等.若要对地点进行覆盖面全的统计,利用古诗词图谱是必要的.在古诗词图谱中,"边塞"属于古诗词分类体系中的词语,由"边塞"可以拓展到"西域"、"塞外"、"新疆"、"边关"、"塞北"、"沙漠"等,然后可以进一步拓展直到底层节点为止.通过这样的层级结构,完整而有条理地覆盖了所有地点词汇.对于"山"、"水"、"建筑"、"名胜"也是如此.

图 7 展示了 5 种地点中每一种情感所占的比例.在图 7 中,"哀伤"在"边塞"中所占比例相比于"哀伤"在其他地点中所占比例是最高的,这说明了边塞更容易让人产生悲伤的情绪,原因是边塞总是和战争、戍边联系在一起."愁绪"在"水"中所占比例相比于"愁绪"在其他地点中所占比例是最高的,这说明古人总是将"水"和"愁绪"联系在一起,比如李白的著名诗句"抽刀断水水更流,举杯消愁愁更愁"."孤独"在"山"中所占比例相比于"孤独"在其他地点中所占比例最高,"山"带给人的感觉确实是孤独的.

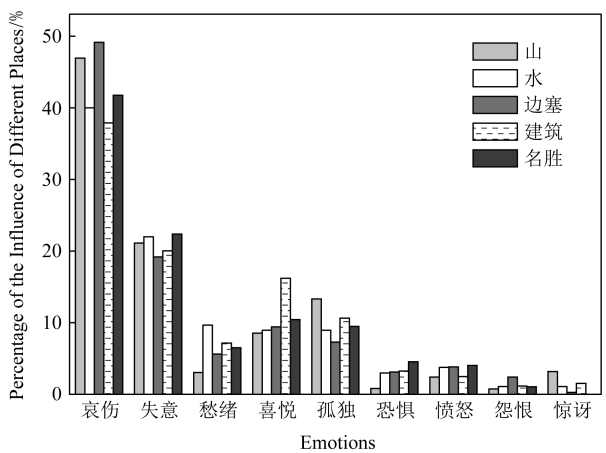

Fig. 7 The histogram of the influence of different places on poets' emotions

图 7 不同地点对诗人情感的影响柱形图

3.2 不同时期对诗人写作风格相似度的影响

唐朝分为初唐、盛唐、中唐、晚唐 4 个时期,本节将首先描述 2 个诗人之间写作风格相似度的度量方式,然后对同一时期诗人写作风格的相似度和相邻时期诗人写作风格的相似度进行度量,探究不同时期对诗人写作风格相似度的影响.

3.2.1 2 个诗人之间写作风格相似度的度量

我们对写作风格的定义有 2 个维度,分别是题材和情感.关于题材,有"送别"、"羁旅"、"战争"、"田园"、"爱情"、"怀人"、"被贬"、"咏史"、"思乡"、"山水"10 种类别.关于情感,有"喜悦"、"愤怒"、"哀伤"、

“愁绪”、“孤独”、“恐惧”、“惊讶”、“怨恨”、“失意”9 种类别.利用古诗词图谱,可以更加全面、准确地统计诗词中与题材和情感有关的词语.

每位诗人可表示为 $\boldsymbol{V}_{\text{poet}}=(\boldsymbol{V}_{\text{theme}},\boldsymbol{V}_{\text{emotion}})$.其中,$\boldsymbol{V}_{\text{theme}}$和$\boldsymbol{V}_{\text{emotion}}$分别是题材向量和情感向量.题材向量可表示为 $\boldsymbol{V}_{\text{theme}}=(per^{1}_{\text{theme}},per^{2}_{\text{theme}},\cdots,per^{10}_{\text{theme}})$,它的每一个维度:

$$per^{i}_{\text{theme}}=\frac{\#\text{words about }theme^{i}}{\sum_{i=1}^{10}\#\text{words about }theme^{j}}$$

表示该诗人所写的诗中与 $theme^{i}$ 题材有关的词语占全部与题材有关的词语的比例.情感向量可表示为 $\boldsymbol{V}_{\text{emotion}}=(per^{1}_{\text{emotion}},per^{2}_{\text{emotion}},\cdots,per^{9}_{\text{emotion}})$,它的每一个维度:

$$per^{i}_{\text{emotion}}=\frac{\#\text{words about }emotion^{i}}{\sum_{i=1}^{9}\#\text{words about }emotion^{j}}$$

表示该诗人所写的诗中与 $emotion^{i}$ 情感有关的词语占全部与情感有关的词语的比例.

然而,得到每位诗人写作题材和情感的占比,对于分析时期的特点没有太多价值,更有价值的是将占比从大到小排序.因此,每位诗人的表示由 2 个向量转换成了 2 个有序列表,即 $\boldsymbol{L}_{\text{poet}}=(\boldsymbol{L}_{\text{theme}},\boldsymbol{L}_{\text{emotion}})$,其中 $\boldsymbol{L}_{\text{theme}}=(theme^{1},theme^{2},\cdots,theme^{10})$,$\boldsymbol{L}_{\text{emotion}}=(emotion^{1},emotion^{2},\cdots,emotion^{9})$.通过度量 2 个有序列表的相似度,我们可以度量 2 个诗人写作风格的相似度.我们使用 RBO(rank-biased overlap)[35] 来度量有序列表的相似度.

3.2.2 同一时期诗人写作风格的相似度

利用古诗词图谱的信息,可以将时间和诗人写作风格这 2 个因素有效地联系起来.而且,古诗词知识图谱包含的信息有多种维度且非常全面,足以支撑对同一时期诗人写作风格的相似性的探究.

对同一时期的两两诗人之间计算相似度后取平均,可以得到这一时期写作风格的相似程度.初唐、盛唐、中唐、晚唐,这 4 个时期写作风格的相似程度如图 8 所示.

图 8 中,这 4 个时期诗人写作风格的相似程度整体呈现上升趋势,究其原因,是源于对前一时期和同一时期著名诗人写作风格的效仿.而从中唐到晚唐,写作风格相似程度降低,说明诗坛的彻底没落导致对同一时期的效仿现象减少.

具体来说,初唐时期写作风格的相似程度是非常低的,也就是说,诗人的写作风格在初唐是多元化的,之后到盛唐、中唐,诗人的写作风格越来越趋于一致.晚唐时,相比于中唐时期,写作风格的相似程度有所下降.结合当时的背景来看,这个结论是比较合理的.初唐是唐诗的起步发展时期,写作风格是多元化的,而后面的时期诗人的写作就有了可参照的对象.盛唐时期是中国诗词的最高峰,在这个时期之后,很多诗人都在效仿盛唐时期诗人的写作风格.比如杜甫就是一个典型的被后世效仿和追随的对象.由于盛唐之后诗坛的没落,在同时期十分出色的诗人也会被效仿.比如中唐时期,元和新体的 2 个诗派分别以白居易和韩愈为首,这 2 位诗人就是被追随的对象.所以,从初唐到中唐,写作风格的相似程度呈现出上升的趋势.然而到了晚唐,诗坛的彻底没落,会导致诗人没有可以追随的同时期的诗人,所以写作风格的相似程度会呈现出下降的趋势.

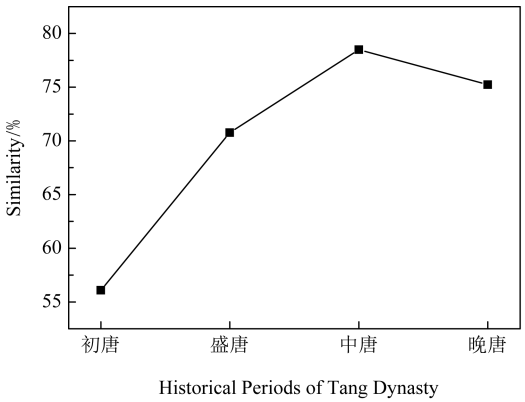

Fig. 8 Trend of the similarity of writing style in different historical periods of Tang Dynasty

图 8 唐朝不同时期写作风格相似度的变化趋势图

3.2.3 相邻时期诗人写作风格的相似度

利用古诗词图谱中蕴含的信息,可以将时间和诗人写作风格这 2 个因素关联起来.而且,古诗词知识图谱覆盖面全、具有多种维度的特点,使得对相邻时期诗人写作风格的相似度进行分析成为可能.

对 2 个相邻时期的两两诗人之间计算相似度后取平均,可以得到这 2 个相邻时期的写作风格的相似度.图 9 展示了初唐-盛唐、盛唐-中唐、中唐-晚唐的相似度.

从图 9 可以发现,随着时间的推移,相邻 2 个时期写作风格的相似度越来越高.这说明随着时间的推移,效仿前一个时期写作风格的现象越来越显著.

利用古诗词图谱对诗词进行数据分析,相比于只统计单字,可以得到更深层次、更有意义的结论.

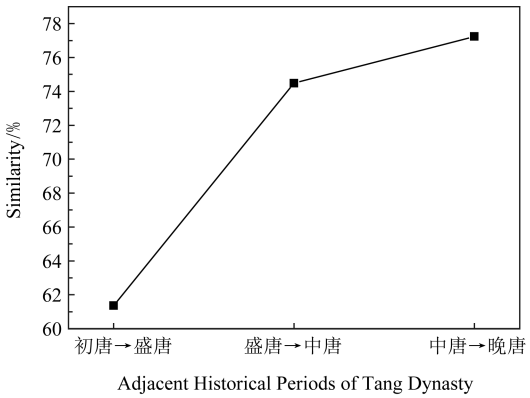

Fig. 9 Trend of similarity of writing style in the adjacent historical periods of Tang Dynasty

图 9 唐朝相邻时期写作风格相似度的变化趋势图

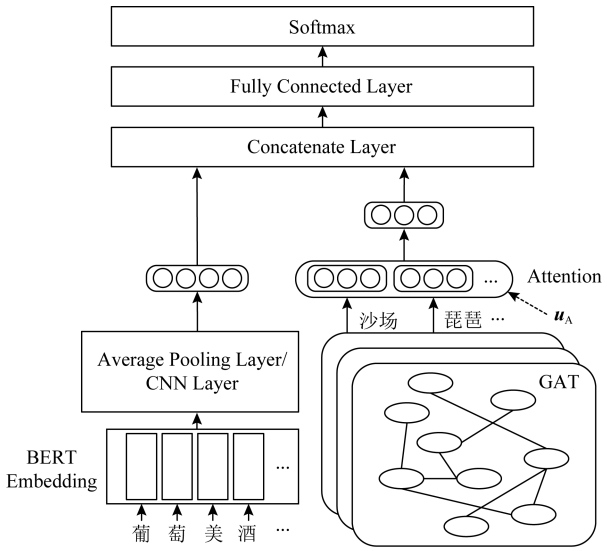

Fig. 10 The model integrating classical Chinese poetry knowledge graph

图 10 融合古诗词图谱的模型

4 古诗词图谱用于推理和分析任务

古诗词图谱可以适用于各种关于诗词的推理和分析任务,下面在判定诗词题材和分析诗词情感的 2 个任务上进行验证.

4.1 分类模型

针对判定诗词题材和分析诗词情感的 2 个任务,模型结构如图 10 所示.模型由 Embedding 层、CNN 层/平均池化层、图注意力层、注意力层组成.首先,将诗词内容用 BERT 编码,然后通过 CNN 层/平均池化层得到诗词的更抽象的表示;将古诗词图谱用图注意力网络编码,学习到每个节点的表示,然后抽取出诗词包含的图节点,用注意力机制把它们的表示做加权和.最后,将诗词的这 2 部分表示拼接起来,接 softmax 分类器,用交叉熵损失函数来计算损失.

4.1.1 Embedding 层

本文使用谷歌开源的 BERT[36] 中文预训练模型初始化字向量.对于每首唐诗,将标题和内容输入 BERT 模型,最大长度设为 150.由于 BERT 中文预训练模型输出的字向量的维度是 768,所以每首诗经过 BERT 预训练模型后,得到的表示的维度为 (150,768).

4.1.2 CNN 层/平均池化层

BERT 预训练模型输出字的表示将通过 CNN 层或平均池化层.接下来分别描述 CNN 层和平均池化层.

卷积神经网络(convolutional neural network, CNN)[37] 可用于捕捉文本的 n-gram 特征.

假设 $\boldsymbol{x}_i \in \mathbb{R}^k$ 表示句子第 i 个字的字向量,维度为 k.则长度为 n 的句子,向量表示为

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n, \tag{4}$$

其中,$\oplus$ 表示拼接操作.更一般地,$\boldsymbol{x}_{i:i+j}$ 指的是 $\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_{i+j}$ 的拼接.

对于卷积层,假设过滤器 $\boldsymbol{w} \in \mathbb{R}^{hk}$,每次以大小为 h 的窗口滑过句子时都会产生一个新的特征.比如,当过滤器滑到 $\boldsymbol{x}_{i:i+h-1}$ 时,产生的特征 c_i 为

$$c_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h-1} + b), \tag{5}$$

其中,$b \in \mathbb{R}$ 为偏置项,f 是激活函数,本文使用 ReLU(rectified linear unit)作为激活函数.所以,当过滤器滑过句子,也就是经历了所有窗口 $\{\boldsymbol{x}_{1:h}, \boldsymbol{x}_{2,h+1}, \cdots, \boldsymbol{x}_{n-h+1:n}\}$ 时,产生特征图 $\boldsymbol{c} \in \mathbb{R}^{n-h+1}$:

$$\boldsymbol{c} = (c_1, c_2, \cdots, c_{n-h+1}). \tag{6}$$

对于池化层,采用最大池化,对每个特征图捕捉最重要的特征:

$$\hat{c} = \max\{c_1, c_2, \cdots, c_{n-h+1}\}. \tag{7}$$

以上描述了通过 1 个过滤器获得 1 个特征的过程.在常见的 CNN 模型中,用不同尺寸的多个过滤器获得多个特征.

平均池化层采用 bert-as-service① 的平均池化策略,即每首通过 BERT 模型编码的诗,对第 1 个

① https://github.com/hanxiao/bert-as-service

维度(最大序列长度)取平均,这样每首诗就表示成768维的向量,使得不同长度的诗全都编码成相同长度的向量.

4.1.3 图注意力层

图注意力网络(graph attention network, GAT)[38],能够很好地学习图结构数据的节点表示.只需要提供图结构和节点的初始特征,给予适当的监督信号,就能够自动学习其他节点对当前节点的重要程度,从而对当前节点进行更好的表示.图注意力网络的核心是图注意力层.接下来,对图注意力层的原理进行详细介绍.

图注意力层的输入是节点的初始特征,$H=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N\}$,$\boldsymbol{h}_i\in\mathbb{R}^F$,$N$ 是节点个数,F 是每个节点的特征数.图注意力层的输出是节点的新的特征,$H'=\{\boldsymbol{h}'_1,\boldsymbol{h}'_2,\cdots,\boldsymbol{h}'_N\}$,$\boldsymbol{h}'_i\in\mathbb{R}^{F'}$,$F'$是每个节点新的特征数.

首先,对每个节点的特征进行线性变换,通过权重矩阵 $\boldsymbol{W}\in\mathbb{R}^{F'\times F}$,将 $\boldsymbol{h}_i$ 变为 $\boldsymbol{W}\boldsymbol{h}_i$,使维度由 F 维变为 F'维.

然后,计算两两节点之间的注意力分数 e_{ij}:

$$e_{ij}=a(\boldsymbol{W}\boldsymbol{h}_i,\boldsymbol{W}\boldsymbol{h}_j)=LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\|\boldsymbol{W}\boldsymbol{h}_j]). \tag{8}$$

接下来,通过 $softmax$ 函数将注意力分数 e_{ij} 进行规范化,从而得到注意力权重 α_{ij},计算过程为

$$\alpha_{ij}=softmax(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})}. \tag{9}$$

最后,节点 i 的表示 $\boldsymbol{h}'_i$ 是它的所有邻居节点表示的加权和再通过 ELU(exponential linear unit)[39] 激活函数得到的:

$$\boldsymbol{h}'_i=\sigma\Big(\sum_{j\in N_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\Big). \tag{10}$$

之后,还可以进一步扩展为多头图注意力层.把 K 个图注意力层得到的节点表示进行拼接,就得了新的节点表示:

$$\boldsymbol{h}'_i=\Big\|_{k=1}^{K}\sigma\Big(\sum_{j\in N_i}\alpha_{ij}^{k}\boldsymbol{W}^{k}\boldsymbol{h}_j\Big). \tag{11}$$

将古诗词图谱的每个节点用 BERT 模型初始化特征,然后输入多层 GAT,得到新的节点表示.

4.1.4 注意力层

假设一首诗包含的图节点的集合是 $N=\{n_1,n_2,\cdots,n_m\}$.从 GAT 层的输出取出这些节点的特征 $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m\}$,通过注意力机制,对诗中提到的不同节点的重要度进行学习,以产生合理的表示.注意力层的公式为

$$\boldsymbol{u}_i=\tanh(\boldsymbol{W}_{\mathrm{A}}\boldsymbol{f}_i+\boldsymbol{b}_{\mathrm{A}}), \tag{12}$$

$$\beta_i=\frac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\mathrm{A}})}{\sum_i\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\mathrm{A}})}, \tag{13}$$

$$\boldsymbol{v}=\sum_i\beta_i\boldsymbol{f}_i, \tag{14}$$

这一层的参数是 $\boldsymbol{W}_{\mathrm{A}}$,$\boldsymbol{b}_{\mathrm{A}}$,$\boldsymbol{u}_{\mathrm{A}}$.

4.1.5 损失函数

把 CNN 层/平均池化层输出的句子表示与层级注意力网络输出的表示通过 Concatenate Layer 结合在一起,然后经过 Fully Connected Layer 将维度映射成分类标签的个数 C,然后利用 Softmax 层进行归一化,得到预测向量 $\hat{\boldsymbol{y}}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_C)$.假设真实向量 $\boldsymbol{y}=(y_1,y_2,\cdots,y_C)$,若样本的真实标签对应的下标为 m,则 $y_m=1$,$\boldsymbol{y}$ 的其他元素都为 0.得到预测向量后,用交叉熵损失函数来计算损失:

$$Loss=-\sum_{i=1}^{C}y_i\cdot\mathrm{lb}(\hat{y}_i). \tag{15}$$

4.2 实验设置

4.2.1 数据集

为了评测判定诗词题材和分析诗词情感这 2 个任务的实验效果,我们设计了一些规则自动地对诗词进行类别的标注.关于判定诗词题材任务,标注了“战争”、“送别”、“闺怨”、“怀古”、“田园”、“山水”、“思乡”、“怀人”、“咏物”、“悼亡”共 10 种类别.关于分析诗词情感任务,标注了“悲伤”、“愁绪”、“孤独”、“怨恨”、“思念”共 5 种类别.关于这 15 种类别的标注规则,详见表 2.数据集的规模详见表 3.

4.2.2 评价指标

在涉及的所有分类实验中,我们使用准确率(precision, P)、召回率(recall, R)和 F 值(F-measure)作为每个类别的评价指标,使用宏平均值(macro-average)作为每种分类方法最终的评价指标,计算方法为

$$P=\frac{TP}{TP+FP}, \tag{16}$$

$$R=\frac{TP}{TP+TN}, \tag{17}$$

$$F=\frac{2\times P\times R}{P+R}, \tag{18}$$

$$P_{\mathrm{macro}}=\frac{1}{n}\sum_{i=1}^{n}P_i, \tag{19}$$

$$R_{\mathrm{macro}}=\frac{1}{n}\sum_{i=1}^{n}R_i, \tag{20}$$

$$F_{\mathrm{macro}}=\frac{2\times P_{\mathrm{macro}}\times R_{\mathrm{macro}}}{P_{\mathrm{macro}}+R_{\mathrm{macro}}}, \tag{21}$$

Table 2 Rules of Constructing Datasets

表 2 构建数据集的规则描述

Dataset	Category	Rules
Theme	战争	统计与战争有关的词集合 $W_{战争}$={沙场,都护,凉州,…}和字集合 $C_{战争}$={虏,蕃,战,…}.假设一首诗中出现在 $W_{战争}$中词的个数为 W_{count},出现在 $C_{战争}$中字的个数为 C_{count}.若 $W_{count}\geqslant 2$ 或 $C_{count}\geqslant 5$,则判定该诗类别为“战争”.
	送别	若“送”或“别”出现在某诗的标题中,则判定该诗类别为“送别”.
	闺怨	统计与闺怨诗的标题有关的词集合 $Title_{闺怨}$={春怨,宫怨,玉阶怨,…}.若某诗的标题中含有 $Title_{闺怨}$中的词,则判定该诗类别为“闺怨”.
	怀古	统计与怀古诗的标题有关的词集合 $Title_{怀古}$={怀古,览古,咏史,…}.若某诗的标题中含有 $Title_{怀古}$中的词,则判定该诗类别为“怀古”.
	田园	统计与田园诗的标题有关的词集合 $Title_{田园}$={田家,田园,闲居,…}.若某诗的标题中未出现“寄”、“赠”、“送”、“呈”,且含有 $Title_{田园}$中的词,则判定该诗类别为“田园”.
	山水	若某诗的标题中未出现“赠”、“答”、“怀”、“忆”,且标题的第一个字为“游”或“登”或“望”,且标题中出现了“寺”、“山”或“峰”,则判定该诗类别为“山水”.
	思乡	统计与思乡有关的词集合 $W_{思乡}$={故乡,他乡,游子,…}.若某诗的标题中未出现“送”、“别”,且诗句中含有 $W_{思乡}$中的词,则判定该诗类别为“思乡”.
	怀人	统计与怀人有关的词集合 $W_{怀人}$={知音,故人,思君,…}.若某诗的标题中出现“怀”或“忆”或“寄”,且诗句中含有 $W_{怀人}$中的词,则判定该诗类别为“怀人”.
	咏物	统计与咏物有关的词集合 $W_{咏物}$={柳,松,菊,…}.假设某诗的标题 $title$ 中含有 $W_{咏物}$中的某个词 $word$,且满足以下的 3 种形式之一,$title=word$,$title$=“咏”+$word$,$title$=“题”+$word$,则判定该诗类别为“咏物”.
	悼亡	若某诗标题的第 1 个字为“哭”,则判定该诗类别为“悼亡”.
Emotion	悲伤	统计与悲伤有关的词集合 $W_{悲伤}$={流涕,沾衣,垂泪,…}和字集合 $C_{悲伤}$={悲,凄,泣,…}.假设一首诗中出现在 $W_{悲伤}$词的个数为 W_{count},出现在 $C_{悲伤}$中字的个数为 C_{count}.若 $W_{count}\geqslant 1$ 或 $C_{count}\geqslant 3$,则判定该诗类别为“悲伤”.
	愁绪	统计与愁绪有关的词集合 $W_{愁绪}$={旅愁,客愁,愁思}.若一首诗中出现了 $W_{愁绪}$中的词,则判定该诗类别为“愁绪”.
	孤独	统计与孤独有关的词集合 $W_{孤独}$={孤舟,孤客,孤灯,…}和字集合 $C_{孤独}$={孤,独,…}.假设一首诗中出现在 $W_{孤独}$中词的个数为 W_{count},出现在 $C_{孤独}$中字的个数为 C_{count}.若 $W_{count}\geqslant 1$ 或 $C_{count}\geqslant 1$,则判定该诗类别为“孤独”.
	怨恨	统计与怨恨诗的标题有关的词集合 $Title_{怨恨}$={怨,宫词,闺,…}.若某诗的标题中出现 $Title_{怨恨}$中的词,则判定该诗类别为“怨恨”.
	思念	统计与思念诗的标题有关的词集合 $Title_{思念}$={忆,寄,怀,…},与思念有关的词集合 $W_{思念}$={相思,故人,旧游,…}.若某诗的标题中出现 $Title_{思念}$中的词,且诗句中出现 $W_{思念}$中的词,则判定该诗类别为“思念”.

Table 3 Statistics of Datasets for Two Tasks

表 3 2 个任务的数据集信息

Labels	Dataset for Predicting Theme of Classical Chinese Poetry											Dataset for Predicting Emotion of Classical Chinese Poetry					
	战争	送别	闺怨	怀古	田园	山水	思乡	怀人	咏物	悼亡	Total	悲伤	愁绪	孤独	怨恨	思念	Total
Train	369	524	301	321	240	228	659	412	316	125	3 495	440	213	263	301	315	1 532
Dev	22	27	18	22	14	10	30	13	18	16	190	55	26	32	37	39	189
Test	33	42	28	41	20	14	31	7	23	16	255	55	28	34	39	40	196
Total	424	593	347	384	274	252	720	432	357	157	3 940	550	267	329	377	394	1 917

其中,TP(true positive)表示将正类预测为正类的个数,TN(true negative)表示将负类预测为负类的个数,FP(false positive)表示将负类预测为正类的个数,FN(false negative)表示将正类预测为负类的个数,n 是类别个数.

4.2.3 超参数设置

在验证集上对超参数进行网格搜索,确定下来的超参数如表 4 所示.在判定诗词题材的模型中,使用平均池化层;用 2 层 GAT 来学习古诗词图谱的节点表示.在分析诗词情感的模型中,使用 CNN 层;

Table 4 Hyper-Parameters of the Models

表 4 模型的超参数

Hyperparameters	Model-1	Model-2
GAT Heads	[4,4]	[4]
GAT Features	[32,32]	[32]
Filters		128
Kernel Sizes		[2,3,4]
Learning Rate	0.005	0.005
Dropout Rate	0.5	0.5
Batch Size	64	512
Epoch Size	100	100

用1层GAT来学习古诗词图谱的节点表示.GAT Heads表示每一层GAT的头的个数,GAT Features表示每一层GAT输出的特征个数,Filters表示过滤器个数,Kernel Sizes表示不同卷积核的尺寸.

4.3 实验结果与分析

4.3.1 剥离实验

1) Rand.最基础的模型是"随机初始化字向量+平均池化层/CNN层+softmax分类器",用"Rand"简写.

2) +BERT.为了探究BERT的影响,将随机初始化字向量替换为BERT,变成"BERT+平均池化层/CNN层+softmax分类器",用"+BERT"简写.

3) +BERT+GAT.为了探究古诗词图谱的影响,将其加入模型,变成"BERT+平均池化层/CNN层+GAT+softmax分类器",用"+BERT+GAT"简写.

在判定诗词题材任务上,加入BERT后,模型效果提升了14.61%,再加入GAT后,模型效果提升了5.72%.在分析诗词情感任务上,加入BERT后,模型效果降低了1.86%,加入GAT后,模型效果提升了4.13%.实验结果如表5所示,C(compare)列表示模型相比于Rand模型效果提升的大小.

Table 5 Results of Ablation Experiments

表5 模型的剥离实验 %

Methods	Predicting Theme of Classical Chinese Poetry				Predicting Emotion of Classical Chinese Poetry			
	P_{macro}	R_{macro}	F_{macro}	C	P_{macro}	R_{macro}	F_{macro}	C
Rand	62.30	58.26	56.36		92.66	92.22	92.22	
+BERT	71.73	72.57	70.97	+14.61	90.97	90.12	90.36	−1.86
+BERT+GAT	**78.10**	**75.87**	**76.69**	+20.33	**95.29**	**93.93**	**94.49**	+2.27

Note: The boldfaced numbers emphasize the performance of our model.

实验结果表明:

1) BERT在判定诗词题材任务上对模型效果有极大提升,而在分析诗词情感任务上甚至起到了反作用,出现这种现象的原因是,BERT可以有效捕捉上下文的信息,包括文本的n-gram信息,在判定诗词题材任务上,若使用"随机初始化字向量+平均池化层",无法捕捉到字与字之间的联系,而加入BERT,就加入了字的上下文信息,所以加入BERT有极大提升.而在分析诗词情感任务上,最初的模型是"随机初始化字向量+CNN层",CNN层本身就可以捕捉文本的n-gram信息,再加入BERT,BERT中包含的上下文信息就可能变成噪声,影响模型的效果.

2) 古诗词图谱在2个任务上对模型效果均有较大提升,证明了古诗词图谱可以为诗词的推理和分析任务提供知识,对诗词的理解有帮助.

4.3.2 与其他方法的对比

本文选择的对比方法如下,实验结果如表6所示:

1) NB.胡韧奋等人[40]对唐诗题材自动分类的研究采用的方法之一.我们基于scikit-learn①,以TF-IDF作为文本特征,实现了多项式朴素贝叶斯分类器.

2) SVM.胡韧奋等人[40]对唐诗题材自动分类的研究采用的方法之一.我们基于scikit-learn,以TF-IDF作为文本特征,实现了线性核支持向量机.

3) CNN.用1层CNN自动学习文本特征[37],然后接softmax分类器.共有2个变种:CNN-rand,CNN-pretrain.对于前者,字向量被随机初始化,且在训练的过程中可被微调;对于后者,使用在4万首唐诗上预训练产生的字向量,不可微调.

4) GRU.用1层GRU自动学习文本特征[41],然后接softmax分类器.共有2个变种:GRU-rand,GRU-pretrain.对于前者,字向量被随机初始化,且在训练的过程中可被微调;对于后者,使用在4万首唐诗上预训练产生的字向量,不可微调.

5) TextGCN.TextGCN[42]模型将文本语料建模成由文档节点和词节点组成的异质图,然后利用图卷积神经网络(GCN)做半监督文本分类.

实验结果表明:

1) 我们的模型在判定诗词题材和分析诗词情感这2个任务上分别超过最好的基线模型2.64%和0.45%,证明了我们提出的模型的有效性.

2) 传统的朴素贝叶斯和支持向量机分类器在这2个任务上具有可比的效果,比一些基于神经网络的模型效果还要好.

3) CNN-pretrain的效果在这2个任务上比

① https://scikit-learn.org/stable/

CNN-rand 高出 2.98%和 1.33%,GRU-pretrain 的效果在这 2 个任务上比 GRU-rand 高出 3.84%和 5.17%,说明字向量的初始化对模型训练非常重要,使用与训练数据相同领域的大规模语料预训练字向量,能够大幅提升模型效果.在判定诗词题材任务中,GRU-rand 和 GRU-pretrain 的效果分别高于 CNN-rand 和 CNN-pretrain 0.57%和 1.43%;在分析诗词情感任务中,CNN-rand 和 CNN-pretrain 的效果分别高于 GRU-rand 和 GRU-pretrain 25.97%和 22.13%.可以看出,对不同的任务和数据,CNN 和 RNN 有不同的偏好,在判定诗词题材任务上,RNN 的表现更好,在分析诗词情感任务上,CNN 的表现更好.

4) TextGCN 在这 2 个任务上是表现最差的模型,由此可以看出,TextGCN 对于诗词这样特殊的短文本无法达到很好的分类效果.

Table 6 Performance Comparison of Different Methods

表 6 不同方法的性能比较 %

Methods	Predicting Theme of Classical Chinese Poetry			Predicting Emotion of Classical Chinese Poetry		
	P_{macro}	R_{macro}	F_{macro}	P_{macro}	R_{macro}	F_{macro}
NB	75.38	71.81	72.03	71.97	68.27	68.16
SVM	74.03	74.64	72.96	81.90	78.22	79.08
CNN-rand	76.88	72.13	69.64	93.38	92.59	92.71
CNN-pretrain	78.19	72.98	72.62	94.42	93.89	94.04
GRU-rand	84.28	68.75	70.21	67.36	66.5	66.74
GRU-pretrain	77.53	76.96	74.05	76.42	73.05	71.91
TextGCN	73.74	68.91	69.31	64.94	63.64	63.97
Our Model	**78.1**	**75.87**	**76.69**	**95.29**	**93.93**	**94.49**

Note: The boldfaced numbers emphasize the performance of our model.

5 总结与展望

本文提出一种古诗词知识图谱的构建方法,并使用该方法得到一个内容覆盖全面、结构层次分明、包含词语语义联系的古诗词知识图谱.利用古诗词图谱,可以从各种不同的维度上更好地分析诗词.古诗词图谱对于诗词的数据分析是不可或缺的,它能够从语义的角度有效地辅助文学研究.另外,古诗词图谱能够适用于古诗词的各种推理和分析任务上,为这些任务提供必要的知识,从而使机器更好地理解诗词.

该工作的不足之处在于构建的古诗词图谱囊括的知识仍然比较局限,目前包含的知识仅包含诗词注释和中文词典的词语解释.

未来工作中,我们将进一步优化古诗词图谱的构建过程,尽可能地降低人工参与的程度.进一步扩大古诗词图谱的规模,使其囊括更加全面的古诗词知识,比如历朝历代对著名诗人诗词的解读文字等.并且,将古诗词图谱扩展到更加广泛的应用场景,比如在诗歌生成中引入古诗词知识,使其更加符合人的意愿进行创作.再比如利用古诗词图谱中的地名信息,分析古人一生的足迹并对古代的文化中心、历史名胜进行探究等.

参 考 文 献

[1] National Engineering Laboratory for Cyberlearning and Intelligent Technology. Garden of Tang Poetry [EB/OL]. [2019-08-29]. http://poem.studentsystem.org (in Chinese)
(互联网教育智能技术及应用国家工程实验室. 唐诗别苑 [EB/OL]. [2019-08-29]. http://poem.studentsystem.org)

[2] Department of Information Management of Peking University. The knowledge graph of the academic genealogy in Song Dynasty [EB/OL]. [2019-08-29]. http://dh.kvlab.org/cbdb_kg/ (in Chinese)
(北京大学科学评价研究组. "宋代学术传承"知识图谱 [EB/OL]. [2019-08-29]. http://dh.kvlab.org/cbdb_kg/)

[3] Wang Zhaopeng. Chronological map of Tang and Song literature [EB/OL]. [2019-08-29]. https://sou-yun.cn/PoetLifeMap.aspx (in Chinese)
(王兆鹏. 唐宋文学编年地图 [EB/OL]. [2019-08-29]. https://sou-yun.cn/PoetLifeMap.aspx)

[4] Xie Tao, Wu Bin, Wang Bai. New word detection in ancient Chinese literature [C] //Proc of Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conf on Web and Big Data. Berlin: Springer, 2017: 260-275

[5] Liu Yutong, Wu Bin, Xie Tao, et al. New word detection in ancient Chinese corpus [J]. Journal of Chinese Information Processing, 2019, 33(1): 46-55 (in Chinese)
(刘昱彤, 吴斌, 谢韬, 等. 基于古汉语语料的新词发现方法[J]. 中文信息学报, 2019, 33(1): 46-55)

[6] Michel J B, Yuan Kuishen, Aiden A P. Quantitative analysis of culture using millions of digitized books [J]. Science, 2011, 331(6014): 176-182

[7] Schich M, Song Chaoming, Ahn Y Y, et al. A network framework of cultural history [J]. Science, 2014, 345(6196): 558-562

[8] State Key Lab of CAD&CG in Zhejiang University, Xinhuanet Co LTD. Use Song Ci to draw the world [EB/OL]. [2019-08-29]. http://www.xinhuanet.com/video/sjxw/2018-09/07/c_129948936.htm?tdsourcetag=s_pctim_aiomsg (in Chinese)
(浙江大学CAD&CG国家重点实验室, 新华网股份有限公司. 宋词缱绻, 何处画人间[EB/OL]. [2019-08-29]. http://www.xinhuanet.com/video/sjxw/2018-09/07/c_129948936.htm?tdsourcetag=s_pctim_aiomsg)

[9] Guo Zhipeng, Yi Xiaoyuan, Sun Maosong, et al. Jiuge: A human-machine collaborative Chinese classical poetry generation system [C] //Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 25-30

[10] Yi Xiaoyuan, Sun Maosong, Li Ruoyu, et al. Automatic poetry generation with mutual reinforcement learning [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 3143-3153

[11] Yang Cheng, Sun Maosong, Yi Xiaoyuan, et al. Stylistic Chinese poetry generation via unsupervised style disentanglement [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 3960-3969

[12] HIT-SCIR. BiCilin [EB/OL]. [2019-08-29]. http://www.bigcilin.com (in Chinese)
(哈工大社会计算与信息检索研究中心. 大词林[EB/OL]. [2019-08-29]. http://www.bigcilin.com)

[13] Dong Zhendong, Dong Qiang. HowNet-a hybrid language and knowledge resource [C] //Proc of Int Conf on Natural Language Processing and Knowledge Engineering. Piscataway, NJ: IEEE, 2003: 820-824

[14] Bo Xu, Xu Yong, Liang Jiaqing, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system [C] //Proc of Int Conf on Industrial, Engineering and Other Applications of Applied Intelligent System. Berlin: Springer, 2017: 428-438

[15] Liu Qiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600 (in Chinese)
(刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600)

[16] Ling Xiao, Weld D S. Fine-grained entity recognition [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2012: 94-100

[17] Jain A, Pennacchiotti M. Open entity extraction from Web search query logs [C] //Proc of the 23rd Int Conf on Computational Linguistics. New York: ACM, 2010: 510-518

[18] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C] //Proc of the 2016 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 260-270

[19] Chen Liwei, Feng Yansong, Zhao Dongyan. Extracting relations from the Web via weakly supervised learning [J]. Journal of Computer Research and Development, 2013, 50(9): 1825-1835 (in Chinese)
(陈立玮, 冯岩松, 赵东岩. 基于弱监督学习的海量网络数据关系抽取[J]. 计算机研究与发展, 2013, 50(9): 1825-1835)

[20] Zhu Jun, Nie Zaiqing, Liu Xiaojiang, et al. StatSnowball: A statistical approach to extracting entity relationships [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 101-110

[21] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1535-1545

[22] Guo Jianyi, Li Zhen, Yu Zhengtao, et al. Extraction and relation prediction of domain ontology concept instance, attribute and attribute value [J]. Journal of Nanjing University: Natural Sciences, 2012, 48(4): 383-389 (in Chinese)
(郭剑毅, 李真, 余正涛, 等. 领域本体概念实例、属性和属性值的抽取及关系预测[J]. 南京大学学报: 自然科学版, 2012, 48(4): 383-389)

[23] Wang Yu, Tan Songbo, Liao Xiangwen, et al. Extended domain model based named attribute extraction [J]. Journal of Computer Research and Development, 2010, 47(9): 1567-1573 (in Chinese)
(王宇, 谭松波, 廖祥文, 等. 基于扩展领域模型的有名属性抽取[J]. 计算机研究与发展, 2010, 47(9): 1567-1573)

[24] Li Yang, Wang Chi, Han Fangqiu, et al. Mining evidences for named entity disambiguation [C] //Proc of the 19th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1070-1078

[25] Han Xianpei, Le Sun, Jun Zhao. Collective entity linking in Web text: A graph-based method [C] //Proc of the 34th Int ACM Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 765-774

[26] Mendes P N, Mühleisen H, Bizer C. Sieve: Linked data quality assessment and fusion [C] //Proc of the 2nd Int Workshop on Linked Web Data Management at Extending Database Technology. New York: ACM, 2012: 116-123

[27] Liu Xueqing, Song Yangqiu, Liu Shixia, et al. Automatic taxonomy construction from keywords [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1433-1441

[28] Lao Ni, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 529-539

[29] Tan C H, Agichtein E, Ipeirotis P, et al. Trust, but verify: Predicting contribution quality for knowledge base construction and curation [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 553-562

[30] Harvard University. China biographical database project [EB/OL]. [2019-08-29]. https://projects. iq. harvard. edu/chinesecbdb (in Chinese)
(哈佛大学. 中国历代人物传记资料库[EB/OL]. [2019-08-29]. https://projects.iq.harvard.edu/chinesecbdb)

[31] Zhou Lina, Hong Liang, Gao Ziyang. Construction of knowledge graph of Chinese Tang poetry and design of intelligent knowledge services [J]. Library and Information Services, 2019, 63(2): 24-33 (in Chinese)
(周莉娜, 洪亮, 高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作, 2019, 63(2): 24-33)

[32] Hu Junfeng, Yu Shiwen. Word meaning similarity analysis in Chinese ancient poetry and its applications [J]. Journal of Chinese Information Processing, 2002, 16(4): 40-45 (in Chinese)
(胡俊峰, 俞士汶. 唐宋诗中词汇语义相似度的统计分析及应用[J]. 中文信息学报, 2002, 16(4): 40-45)

[33] Lee J, Luo Mengqi. Inducing word clusters from classical Chinese poems [J]. International Journal of Asian Language Processing, 2018, 28(1): 13-30

[34] Dong Naibin. Comments on the Narrative Tradition of Chinese Literature [M]. Shanghai: Oriental Publishing Center, 2017 (in Chinese)
(董乃斌. 中国文学叙事传统论稿[M]. 上海: 东方出版中心, 2017)

[35] Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings [J]. ACM Transactions on Information Systems, 2010, 28(4): No.20

[36] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171-4186

[37] Kim Y. Convolutional neural networks for sentence classification [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751

[38] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks [J]. arXiv preprint, arXiv: 1710.10903, 2017

[39] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units [J]. arXiv preprint, arXiv: 1511.07289, 2015

[40] Hu Renfen, Zhu Yuchen. Automatic classification of Tang poetry themes [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2): 262-268 (in Chinese)
(胡韧奋, 诸雨辰. 唐诗题材自动分类研究[J]. 北京大学学报: 自然科学版, 2015, 51(2): 262-268)

[41] Tang Duyu, Qin Bing, Liu Ting. Document modeling with gated recurrent neural network for sentiment classification [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1422-1432

[42] Yao Liang, Mao Chengsheng, Luo Yuan. Graph convolutional networks for text classification [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 7370-7377

Liu Yutong, born in 1996. Master candidate in Beijing University of Posts and Telecommunications. Her main research interests include computational social science, natural language processing, text mining and machine learning.

Wu Bin, born in 1969. Professor and PhD supervisor in Beijing University of Posts and Telecommunications. Member of CCF. His main research interests include graph-based data mining, complex networks, computational social science, intelligent information processing, and business intelligence.

Bai Ting, born in 1992. Assistant professor in Beijing University of Posts and Telecommunications. Her main research interests include recommender systems, graph representation and data mining.