文章编号：1003-0077(2019)03-0079-08

基于主题模型的古典乐器诗词文本挖掘

申资卓，杨 莹，邵艳秋

（北京语言大学 信息科学学院，北京 100083）

摘 要：古代先贤将乐器按其制作材料分为八类，《周礼·春官·大师》中记载“皆播之以八音：金石土革丝木匏竹。”该文将《全唐诗》、《全宋词》中有关“八音”的诗句、词句作为研究对象，使用基于 LDA 和 NMF 的主题挖掘、基于 Author-Topic-Model 的作者相似度计算等方法。从宏观到微观，从整体诗词到具体诗人/词人，从主题的聚类、动词形容词的抽取到具体诗人词人作品相似度的计算，多维度、多层次、多角度研究了唐诗宋词中的中国古典乐器。

关键词：唐诗宋词；“八音”；主题模型

中图分类号：TP391 **文献标识码**：A

Mining of Classical Musical Instrument Poetry Based on Topic Model

SHEN Zizhuo, YANG Ying, SHAO Yanqiu

(School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: This paper presents an investigation into the BaYin, the eight classical musical instruments, depicted in the ancient Chinese poetry. Specifically, we applied LDA and NMF to establish the Author-Topic-Model based author similarity. From the corpus of Tang Poetry and Song Poetry, this paper delivers a panoramic view on the poems, poets, the topics the verbs related to Bayin.

Keywords: Tang Poems and Song Ci; BaYin; Topic Model

0 引言

中国是诗词的故乡，也是音乐的国度。“八音”是古代先贤将乐器按其制作材料分为的八类。韩愈在《送孟东野序》中写道“金石丝竹匏土革木者，物之善鸣者也。”千百年来，随着时代发展和文化交流，中国“八音”中的乐器有的传承至今，有的逐渐失传。

目前对于乐器中“八音”的研究大致可以分为以下几类：其一是对一种具体乐器的研究，马玉婷梳理了涉“磬”诗的发展脉络[1]；李萌、王少杰统计了宋词中的“笙”[2]；蒲雨潇研究了宋词中的“箫”[3]。其二是古代文学文本中“八音”的研究，王华统计分析了《诗经》中“八音”的分布情况[4]。其三是从音乐的视角对“八音”进行的研究。

1 唐诗宋词中“八音”的统计与分析

1.1 诗词中“八音”乐器词语统计分析

1.1.1 乐器词语频次统计

“八音”被认为是中国最古老的乐器分类方法。唐宋是中国古典音乐大繁荣的时期，随着音乐的不断丰富与发展，“八音”所包含的乐器种类也越来越多样。通过文献阅读，我们搜集了古代文学中“八音”的名称，如表 1 所示。

表 1 “八音”中的乐器

八音	代表乐器
金	锣；铮；钟；镛；铃；铙；铎；钹；镎于
石	磬

收稿日期：2018-06-10 **定稿日期**：2018-07-24

基金项目：教育部人文社科规划基金(17YJAZH068)；北京语言大学研究生创新基金(18YCX004)；国家自然科学基金(61872402)；北京语言大学校级项目(中央高校基本科研业务专项资金)(18ZDJ03)；北京市自然科学基金(4192057)

续表

八音	代表乐器
丝	琴(绿绮;丝桐);瑟;琵琶;筝;阮咸;箜篌
竹	笛(篴);芦管;笳;箫;龠;簧;篪;筚篥
匏	笙;竽
革	鼓
土	埙
木	柷;敔;春牍

根据表1分别将《全唐诗》(42 979首)、《全宋词》(21 050首)中包含“八音”的诗词按照字符串匹配的方式挑选出来,统计结果为图1、图2。

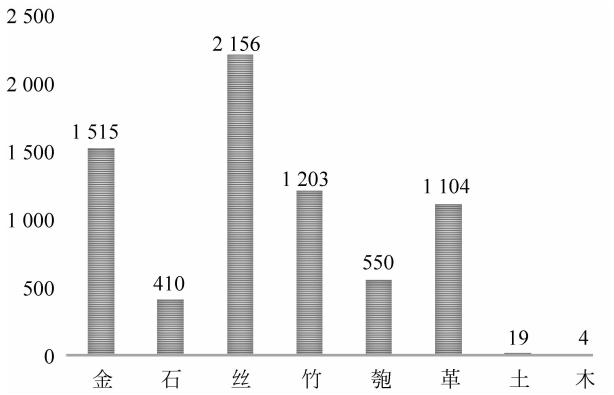

图1 八音在《全唐诗》中的频次分布

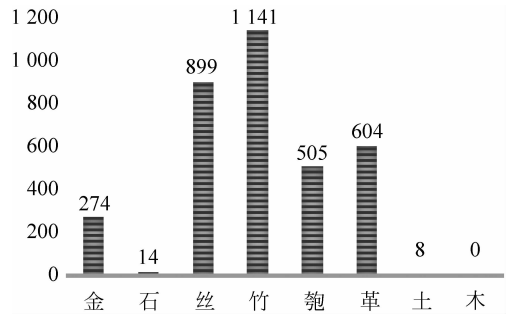

图2 八音在《全宋词》中的频次分布

虽然孤立地看每一首诗词作品都是诗人、词人的个体创作,但一个时代的作品汇集在一起,则能够反映出整个时代相关的一些情况,通过统计《全唐诗》、《全宋词》中“八音”的频次分布,可以宏观了解唐宋“八音”使用的异同。

1.1.2 唐诗中的“丝”与宋词中的“竹”

从图1、图2中可以看出,丝类乐器在唐诗中出现的频次最高,和唐诗的关系最为密切,而在宋词中,出现频次最高的乐器是竹类。

“丝”指的是弹拉弦的乐器,在“丝”类乐器中“琴”出现在1 593首唐诗中,“瑟”被写入480首唐诗中。《全唐诗》中有关“琴”、“瑟”的描写分为三类:

首先,使用于男女爱情的描写。例如“玉轸朱弦瑟瑟徽,吴娃徵调奏湘妃。”表达了“郎不归”的愁情;其次,使用于送别分离的场景。陈子昂写下“离堂思琴瑟,别路绕山川。”离别之情令人感慨歔欷;最后,使用于生活意趣的寄托。诗佛王维诗中的琴显出禅意,如“独坐幽篁里,弹琴复长啸。”诗鬼李贺笔下的琴则透着些许仙气,如“九节菖蒲石上死,湘神弹琴迎帝子。”

“竹”指的是竹制吹奏乐器。在《全宋词》中“箫”在536首词中出现,是“竹”类乐器中出现频次最高的乐器。在宋词中文人偏爱使用“箫”来造境,其一是造浑厚之境:张先的“飞槛倚,斗牛近,响箫鼓、远破重云”。“箫鼓”营造喜洋洋之境,又不乏宏大的气势;其二是造哀戚之境:柳永的“岂知秦楼,玉箫声断,前事难重偶”。箫声呜咽,寄托哀婉凄怨之情思;其三是造离愁之境:张炎的“十二小红楼,人与玉箫何处。”昔日情境,却时过境迁,玉箫声断,徒添伤感[3]。

1.1.3 唐诗宋词中的“石”

石类乐器在唐诗中出现的频次为410次,而在宋词中仅出现了14次。石类乐器在唐诗中的繁荣与唐代佛教的盛行关系密切,在宋词中的冷门原因在于宋词的文体特征。

自东汉以来,随着佛教在中国的兴盛,磬进入佛事作为法器也日渐兴盛,到唐中达到空前的繁荣。诗人们通过“磬”反映佛教禅理、宫廷礼仪、山水田园[1]。

磬历来主要作为礼乐乐器的特性,决定了其使用的局限性。《全宋词》中磬仅仅出现了14次,这与宋词的文体特征有关,宋词是中国文学发展史上第一个抒写艳思恋情的专门文体。宋词的题材集中在描写离愁别绪、风花雪月等方面。因此石类乐器在宋词中出现的频次较低。

1.2 “八音”诗词中动词、形容词的抽取

将从《全唐诗》与《全宋词》中抽取的含有“八音”的诗句、词句以字为单位进行切分,然后使用哈尔滨工业大学的语言技术平台(LTP)工具包进行词性标注,最后抽取出其中的动词与形容词,并取频次较高的前十个字。由于篇幅限制,仅挑选《全唐诗》“八音”所在句的形容词表与《全宋词》“八音”所在句的动词表进行分析研究,如表2和表3所示。

表 2　《全唐诗》“八音”所在句的形容词表

金	远 74	寒 74	长 67	残 61	乐 58	清 56	高 55	暮 44	微 37	深 34
石	清 43	寒 43	古 23	孤 22	高 20	远 19	香 19	静 16	幽 15	深 15
丝	清 139	高 73	长 64	素 62	闲 58	古 45	孤 42	悲 40	多 40	寒 40
竹	清 73	悲 61	寒 61	长 47	远 38	哀 33	高 30	满 29	青 27	暮 27
匏	满 20	清 16	寒 16	红 16	乐 15	远 15	白 15	长 13	香 13	闲 13
革	长 38	寒 31	乐 23	新 22	严 21	白 21	悲 21	远 20	多 20	微 20
土	乐 4	高 1	谐 1	新 1	旱 1	杂 1	紫 1			
木	残 2	乐 1	大 1							

注：“土”类与“木”类乐器抽取的形容词不足十个，原因在于包含这两类乐器的唐诗数量稀少。

从表 2 可以看出，与“竹”类乐器同现的频次较高的三个形容词分别是“清”、“悲”、“寒”，可见“竹”类乐器在唐诗中表达的情感是凄清、悲伤的。例如，李白的诗《春夜洛城闻笛》中写道：“谁家玉笛暗飞声，散入春风满洛城。此夜曲中闻折柳，何人不起故园情。”笛曲《折柳曲》表达了送别时的哀怨和对故乡的眷恋之情[5]。

表 3　《全宋词》“八音”所在句的动词表

金	晓 220	听 21	断 20	疏 18	无 14	动 14	闻 13	有 12	昏 11	是 9
石	浮 2	欣 2	落 2	击 2	疏 2	赋 2	无 1	消 1	游 1	闻 1
丝	弹 54	调 47	有 46	无 46	抱 42	听 37	鸣 36	断 35	思 34	怨 33
竹	吹 251	听 68	断 66	弄 66	横 44	倚 41	如 40	奏 39	去 39	有 39
匏	吹 55	奏 29	醉 26	沸 24	有 23	拥 23	听 18	归 17	落 17	放 16
革	催 57	吹 57	听 35	叠 28	喧 24	鸣 22	无 22	是 22	闻 21	有 19
土	传 2	奏 2	认 1	应 1	贯 1	翻 1	笑 1	醉 1	上 1	圭 1

从表 3 可以发现在宋词中与“八音”同现频次较高的动词多与乐器演奏动作相关，如丝类乐器都为弦乐，与该类乐器同现频次最多的动词就是“弹”；竹类乐器都为管状乐器，与该类乐器同现频次最多的动词就是“吹”。

2　基于 LDA 和 NMF 的“八音”诗词主题挖掘

2.1　基于 LDA 和 NMF 的主题挖掘模型

主题挖掘的目的是从大规模无标记文本中自动挖掘出文本隐含的主题信息，钱鹏[6]等利用 LDA 模型对唐诗文本进行主题建模，并取得了不错的效果。本文在利用 LDA 模型的基础上，利用了另外一种主题建模的方法 NMF 进行对比研究，从而选择出更适合诗词文本的主题建模方法。

LDA(latent dirichlet allocation)全称隐狄利克雷分配，是一种含有隐变量的概率图模型，其中的隐变量就是主题(topic)。LDA 模型的输入为一系列由词项(word)组成的文档(document)，输出为给定主题的词项分布 $P(word|topic)$和给定文档的主题分布 $P((topic|doc)$。LDA 模型假设给定主题的词项分布与给定文档的主题分布都服从多项分布，其参数分别记为 θ、Φ。由于 LDA 模型的参数估计方法是最大后验似然估计，多项分布的参数 θ、Φ 都服从其共轭先验分布，狄利克雷分布，两个狄利克雷分布的参数分别记为 α、β。

NMF(Non-negative matrix factorization)全称非负矩阵分解，是将一个非负矩阵 V 分解为另外两个非负矩阵 W、H 的技术，被广泛应用于图像处理、语音处理、文本挖掘等领域。在文本挖掘中 H 表示文档词频矩阵，W 表示文档主题矩阵，V 表示主题词项矩阵。V、W、H 矩阵的维度分别为 $n\times m$、$n\times k$、$k\times m$，其中 n 表示文档数、m 表示词项数、k 表示主题数。NMF 模型的训练目标为找到合适的 W 矩阵和 H 矩阵，使得 WH 与矩阵 V 最相近，本文采用的损失函数为基于 KL 距离的损失函数，KL 距离是信息论中的概念，用来刻画两个概率分布的相似程度，如式(1)所示。

$$J(V,W,H)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{ij}\log\frac{V_{ij}}{(WH_{ij})}\right) \quad (1)$$

2.2 基于词向量的主题一致性度量

主题一致性(topic coherence)是用来评价主题模型的一种指标,主题一致性越大,就表明主题模型的效果越好。主题一致性计算的是所有主题下前K个词语的语义相似度的平均值。Aletras[7]等利用标准化互信息的方法来计算每个主题下每两个词语的语义相似度,随着词向量技术的兴起,采用词向量计算词语语义相似度的方法取得了更好的效果[8],因此,本文采用的是基于词向量的主题一致性TC-W2V,计算如式(2)所示。

$$TC-W2V=\frac{1}{NC_K^2}\sum_{j=2}^{K}\sum_{i=1}^{j-1}\cos(v_j,v_i) \quad (2)$$

其中cos(·)表示余弦相似度,v_i、v_j为词向量,N为主题个数,C表示词向量组合数。

2.3 实验设计与结果分析

2.3.1 实验设计

本文利用LDA和NMF两种主题模型,分别对唐诗、宋词两类文档集合进行主题建模。实验采用的数据为从《全唐诗》和《全宋词》中抽取出的包含有"八音"乐器关键字的句子。我们规定每个句子为一个文档,抽取后,得到的唐诗数据集的规模为6 961,宋词数据集的规模为3 445。

(1) 语料预处理:由于现有的分词工具对唐诗、宋词的分词效果较差,因此,我们把每个字作为主题模型中的一个词项(word),每个文档由句子中的所有字组成。最后,去除数据集中的停用字,以避免其对模型最终结果的影响。

(2) 文本特征表示:本文采用了词袋模型(Bag-of-words)和词频逆文档频(TF-IDF)两种文本特征表示的方法。词袋模型将文档表示为维度为字典长度的向量,向量的值为字典中的每一个词在该文档中的词频。词频逆文档频的文本特征表示方法考虑到了文档信息对词频的修正,其计算如式(3)~式(5)所示。

$$TF-IDF((w)=TF(w)\cdot IDF(w) \quad (3)$$

$$TF((w)=\log\left(\frac{count(w)}{N_w}\right) \quad (4)$$

$$IDF(w)=\log\left(\frac{D}{D_w+1}\right) \quad (5)$$

其中N_w表示词w所在文档的词的个数,D表示文档总数,D_w表示含有词w的文档数。

(3) 实验设置:实验采用scikit-learn工具包求解LDA、NMF,采用gensim工具包训练词向量。LDA模型的训练算法为在线变分EM算法,共轭先验分布的参数α、β的值设置为0.001,算法的最大迭代次数为200。NMF模型的初始化方法为NNDSVD,该方法能有效地处理稀疏矩阵[9],损失函数设置为基于KL距离的损失函数,算法的最大迭代次数为200。词向量的维度设置为500,训练算法为Skip-gram。LDA和NMF的主题数的取值范围都设置为[5-30]。

2.3.2 实验结果

本文对两种特征(BOW、TF-IDF)和两种模型(LDA、NMF)一一组合,形成四个实验组。对每个实验组都计算给定主题数下的主题一致性,最后记录每个实验组主题一致性的最高值,结果如表4、表5所示。

表4 唐诗中"八音"主题模型结果评价

模型+特征	主题数	主题一致性
LDA+BOW	11	0.219 0
NMF+BOW	22	0.229 5
LDA+TF-IDF	28	0.232 0
NMF+TF-IDF	**18**	**0.257 2**

表5 宋词中"八音"主题模型结果评价

模型+特征	主题数	主题一致性
LDA+BOW	13	0.502 2
NMF+BOW	19	0.533 1
LDA+TF-IDF	23	0.552 3
NMF+TF-IDF	**21**	**0.559 0**

由上表可以看出,无论在唐诗还是宋词数据集上,NMF+TF-IDF组合下的主题一致性都是最高的,所以,我们选择NMF+TF-IDF组合下的主题模型进行主题建模。由此可知,NMF模型比LDA模型更适合对唐诗宋词这种短文本进行主题建模。

在NMF+TF-IDF组合下,唐诗数据集的最优主题个数为18,最优主题一致性为0.257 2,宋词数据集的最优主题个数为21,最优主题一致性为0.559 0,主题个数与主题一致性的关系的折线图如图3、图4所示。

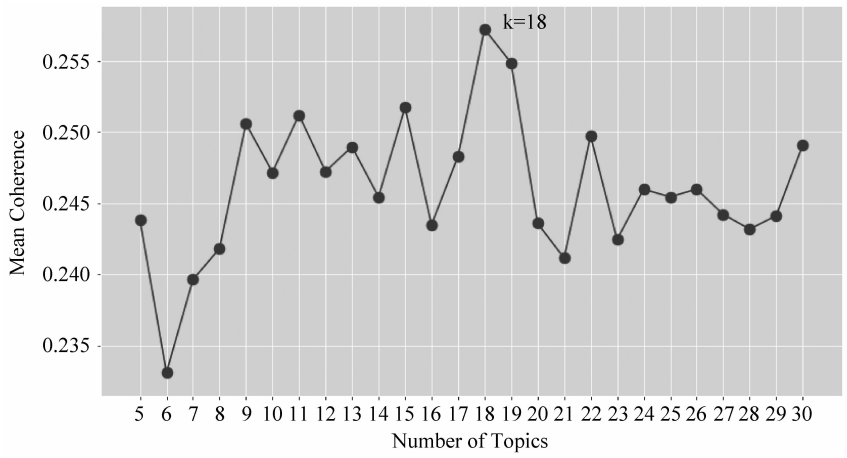

图 3　唐诗“八音”主题模型最优主题个数

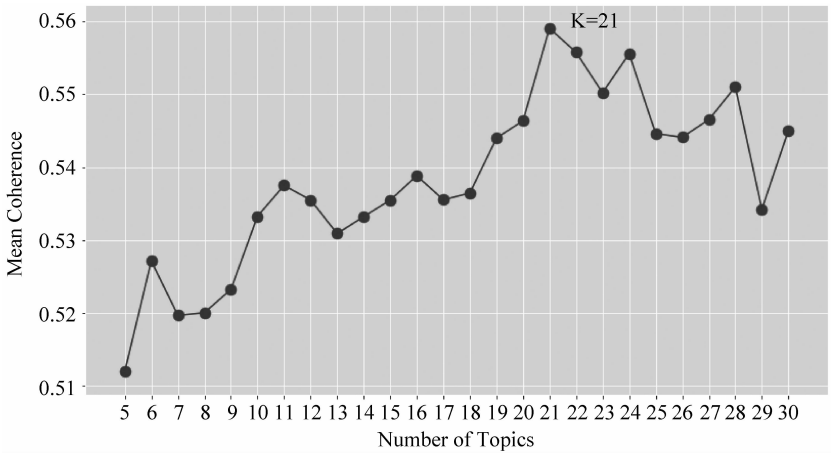

图 4　宋词“八音”主题模型最优主题个数

2.4　“八音”诗词主题聚类

2.4.1　《全唐诗》“八音”诗句主题聚类结果及分析

根据 2.3 中的实验结果，《全唐诗》中有关“八音”的诗句可以聚成十八类，由于篇幅限制将其中效果较好的五类列于表 6。

表 6　《全唐诗》“八音”诗句五类主题聚类

1	寺	磬	题	僧	林	宿	院	禅	寒	房	闻	石	师	清	云	香	疏	灯	松	竹
2	鼓	城	行	鼙	角	军	旗	风	鸣	喧	动	旌	天	马	晓	长	战	海	严	街
3	山	水	关	远	溪	居	道	路	庐	猿	隔	登	松	华	月	泰	隐	石	流	宿
4	中	寄	书	郎	州	李	酬	侍	事	外	韵	丞	怀	居	早	员	御	崔	舍	公
5	送	江	南	州	赴	西	东	李	游	陵	城	船	舟	客	君	水	府	友	阳	京

根据聚类结果给上表五类总结主题分别是：佛禅、边塞、山水、酬赠、送别。第一类佛禅主题聚出的字与寺庙、僧侣、青灯等有关，在这类主题下出现的乐器是“磬”，磬在佛教中的历史悠久，古印度佛教中就有了磬，古书中记载：“乐，石有磬。今浮屠持铜钵，亦名磬。”第二类边塞主题聚出的字都与边塞、战争有关，在这类主题下出现的乐器是“鼓”和“鼙”。

2.4.2　《全宋词》“八音”词句主题聚类结果及分析

根据 2.3 的实验结果，《全宋词》中有关“八音”的词句可以聚成 21 类，由于篇幅限制将其中效果较好的五类列于表 7。

表7 《全宋词》“八音”诗句五类主题聚类

1	歌	笙	院	里	醉	奏	散	拥	放	片	沸	酒	丛	罗	庭	舞	笑	翠	灯	绮
2	钟	疏	晓	昏	闻	残	暮	年	黄	阳	寺	林	景	时	晚	烟	动	清	漏	角
3	云	天	空	外	山	水	远	间	秋	彩	孤	流	高	鹤	飞	片	惊	梦	晚	蓬
4	笳	鸣	胡	悲	发	奏	静	动	入	怨	叠	鼓	送	哀	万	凝	举	戍	塞	恨
5	断	弦	肠	绮	梦	绿	紫	魂	船	索	寒	远	窗	春	朱	愁	信	知	未	回

根据聚类结果给上表五类总结主题分别是：宴饮、归隐、山水、边塞、相思。第一类宴饮主题出现的乐器是“笙”，可见“笙”在宋时宴饮场合的重要性。在唐诗和宋词中都出现了有关边塞主题的聚类，纵向比较唐宋诗词中有关“八音”的聚类发现，在宋词中边塞主题聚类情感是“怨”、是“哀”，是“恨”，而唐诗中的聚类多是对于战争的客观描述，这与唐宋的时代背景不无关系，唐诗中洋溢着盛唐的气象与昂扬的风骨，而宋代历经动荡与分裂，词中对于战争的描写多为流血漂橹、家破人亡的惨状。

3 八音诗词相关作者统计分析

3.1 著名诗人、词人作品中的“八音”的分布

根据王兆鹏[10-11]的研究成果，本文分别找出了唐、宋时期最著名的十大诗人、词人的作品，并就其作品中“八音”使用情况，进行统计研究。为了克服作品数量的不同，无法就“八音”的频次进行比较的问题，这部分将统计出的“八音”的频次分别除以诗人、词人作品数量，得到“八音”在诗人词人作品中的频率，并将频率结果扩大一百倍(图5)。

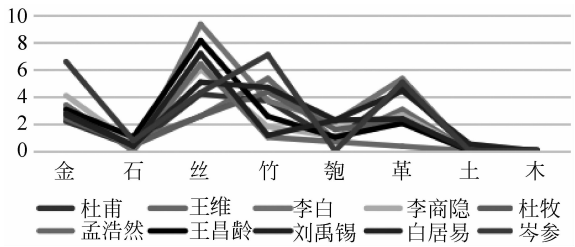

图5 唐代代表诗人作品中“八音”的分布

从图5中可以看出在唐代十大诗人作品中丝类乐器出现的均频率较高，其中又以李白作品中的丝类乐器出现频率最高。在李白的诗歌中，丝类乐器中的各种乐器几乎都出现过，例如，“赵瑟初停凤凰柱，蜀琴欲奏鸳鸯弦”、“横笛弄秋月，琵琶弹陌桑”、“佳人当窗弄白日，弦将手语弹鸣筝”。

李白不仅擅长写琴、瑟等丝类乐器，还能够精湛的演奏。例如，《示金陵子》中写道“ 金陵城东谁家子，窃听琴声碧窗城。”李白的琴声美妙，竟能吸引人来偷听，可见李白琴艺之高，绝非一般人可比。因此李白作品中的丝类乐器出现频率最高。[12]

从图6中可以看出姜夔在其词作中最擅长“八音”的使用，“八音”中的每一类乐器在其作品中的频率基本都是最高，这与姜夔精湛的音乐才能不无关系。姜夔出生于书宦门第，受父辈和市井歌舞的熏陶，擅长诗词音乐[13]。姜夔擅长、精通多种乐器，也是宋代十大代表词人中在其作品中描写乐器最多的词人。

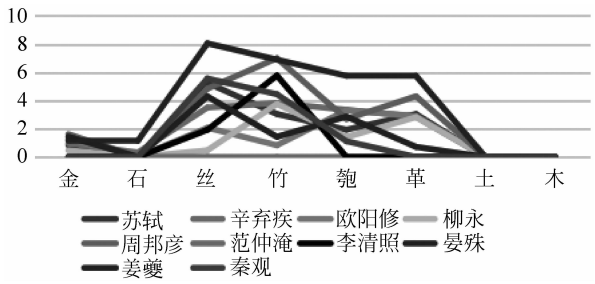

图6 宋代代表词人作品中“八音”的分布

3.2 基于 Author-Topic-Model 的作者相似度计算

3.2.1 Author-Topic-Model

Author-Topic-Model 是融入作者信息的一种主题模型，是对传统的 LDA 模型的修改[14]。Author-Topic-Model 的输入为加入作者信息的一系列文档，Author-Topic-Model 的输出为给定主题的词项分布 $P(\text{word}|topic)$ 和给定作者的主题分布 $P(topic|author)$，假设这两个分布服从多项分布，参数分别记为 θ、Φ，其参数的共轭先验分布服从狄利克雷分布，两个狄利克雷分布的参数分别记为 α、β。另外，Author-Topic-Model 中还需要假定给定文档的作者分布 $P(author|doc)$ 为均匀分布。Author-Topic-Model 较为复杂，其生成文档的过程用图7所示算法描述。

for each topic t **do**:
generate a word distribution for topic t, $\theta_t \sim Dirichlet(\alpha)$
for each author a **do**:
generate a topic distribution for author a, $\Phi_a \sim Dirichlet(\beta)$
for each document d **do**:
for each word w **in** d **do**:
random choice a author for document, according to uniform distribution
random choice a topic for author, according to $P(t|a) \sim Multinomial(\Phi_a)$
Generate a word for topic t, $P(w|t) \sim Mutinomial(\theta_t)$

图 7　生成文档的过程算法图

3.2.2　实验设计

基于 Author-Topic-Model 的作者相似度计算需要 Author-Topic-Model 输出给定作者的主题分布 $P(topic|author)$，每一个作者对应一个维度为主题数的主题偏好向量。因此，我们可以利用不同作者的主题偏好向量的相似度来衡量作者之间的相似度。本文采用的相似度计算公式为余弦距离公式，如式(6)所示。

$$\cos(v_1, v_2) = \frac{\sum_{k=1}^{n} v_{1_k} v_{2_k}}{\sqrt{\sum_{k=1}^{n} v_{1_k}^2} \sqrt{\sum_{k=1}^{n} v_{2_k}^2}} \tag{6}$$

本文利用 gensim 工具包求解 Author-Topic-Model，在融入作者信息的唐诗、宋词的主题模型上，分别计算唐代诗人之间的相似度、宋代词人之间的相似度，由于篇幅限制，这里仅挑选部分具有代表性的结果列于表 8、表 9。

表 8　十大诗人“八音”作者相似度计算举例

作者	相似度	作品数	作者	相似度	作品数
李商隐	1.000 000	91	岑参	1.000 000	94
元稹	0.851 893	144	卢纶	0.728 260	65
戴叔伦	0.824 636	32	王维	0.725 242	61
卢仝	0.810 569	22	吕温	0.721 682	15
韦庄	0.790 286	77	陈羽	0.720 881	12

表 9　十大词人“八音”作者相似度计算举例

作者	相似度	作品数	作者	相似度	作品数
辛弃疾	1.000 000	93	秦观	1.000 000	10
周邦彦	0.999 660	38	刘辰翁	0.903 175	77
刘克庄	0.889 072	82	刘埙	0.854 297	13
贺铸	0.873 276	66	高观国	0.828 850	24
晁端礼	0.838 477	18	朱敦儒	0.816 597	41

3.2.3　基于“八音”唐诗的唐代诗人相似度计算及分析

李商隐与元稹留存的诗作中脍炙人口的以爱情诗为主，风花雪月的爱恨离愁离不开乐器对气氛的烘托与渲染，李商隐与元稹有着一段相似的婚姻经历，二人同样是在意气风发时以寒门学子的身份迎娶了大家闺秀，并且与妻子琴瑟相和度过了一段美好的婚姻生活，但不幸的是妻子都早早的离他们而去，因此李、元二人在爱情诗的题材的选取、情感的表达甚至于创作的结构都存在相似的地方。

岑参是盛唐的边塞诗人，卢纶虽为中唐诗人，其边塞诗却依旧是盛唐的气象，雄壮豪迈。在他们的诗中，常见的景物是大漠、长云、旌旗等，常见的地名是楼兰、关山等，常见的乐器是笛、琵琶、鼓等，意象选择以及用字用语的相似使得二人在“八音”诗句中的相似度计算较高。

3.2.4　基于“八音”宋词的宋代词人相似度计算及分析

词坛“飞将军”辛弃疾代表了宋词豪放派的高峰，周邦彦的词则可认为是婉约词派的集大成者。

从上表可以看出似乎风马牛不相及的两人，其实在写到“八音”的词句创作上有颇多相似之处。

秦观是苏门四学士之一，词的创作上或多或少受到了苏轼的影响，刘辰翁词作风格取法苏辛而自成一体。秦观、刘辰翁均向苏词学习，因此二人在创作上的相似度较高也就不足为奇。

4 结论

本文通过对《全唐诗》和《全宋词》中有关“八音”的诗句与词句的抽取，展现了“八音”在唐诗、宋词中的宏观分布，发现唐诗中出现频次最高的是丝类乐器，宋词中出现频次最高的是竹类乐器；使用基于NMF＋TF-IDF 的主题模型将《全唐诗》、《全宋词》中有关“八音”的诗句分别聚成了十八类和二十一类，得到“笙”在宋代宴饮中是具有重要地位的乐器等结论；通过对“八音”所在诗词句中动词、形容词的抽取，找到了唐诗中与“竹”类乐器同现频次较高的三个情感词分别是“清”、“悲”、“寒”，发现了宋词中与“八音”同现频次较高的动词多与乐器演奏动作相关；本文分别找出了唐、宋时期最著名的十大诗人、词人的作品，并就其作品中“八音”使用情况，进行统计研究；还基于 Author-Topic-Model 进行了作者相似度的计算，得出了李商隐、元稹有关“八音”的诗作最相似等结论。总之，本文通过计算机挖掘的方式为中国古典文学的研究提供了新的切入视角与分析方式。

在后续的工作中，希望进一步完善模型、改进算法，使其能够具有更好的泛化性能，不仅能够应用于诗词“八音”的研究之中，还能在诗人关系挖掘和计算机自动理解诗词等更加复杂的任务中发挥作用。

参考文献

[1] 马玉婷. 唐代涉“磬”诗研究[D]. 长春：吉林大学硕士学位论文，2017.

[2] 李萌，王少杰. 试析宋词中的笙意象[J]. 丝绸之路，2013(6)：116-118.

[3] 蒲雨潇. 醉在宋词里的箫——宋词中的箫意象之浅析[J]. 鸡西大学学报(综合版)，2014(2)：120-122.

[4] 王华.《诗经》乐器“八音”类述[D]. 上海：上海师范大学硕士学位论文，2011.

[5] 贾黎明. 唐诗里的笛子[J]. 文化创新比较研究，2017(22).

[6] 钱鹏，黄萱菁. 中国古诗统计建模与宏观分析[J]. 江西师范大学学报(自然科学版)，2015,2：117-123.

[7] Aletras N, Stevenson M. Evaluating topic coherence using distributional semantics[C]//Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)-Long Papers. 2013：13-22.

[8] O'Callaghan D, Greene D, Carthy J, et al. An analysis of the coherence of descriptors in topic modeling[J]. Expert Systems with Applications：An International Journal, 2015, 42(13)：5645-5657.

[9] Boutsidis C, Gallopoulos E. SVD based initialization：A head start for nonnegative matrix factorization[J]. Pattern Recognition, 2008, 41(4)：1350-1362.

[10] 王兆鹏. 唐诗排行榜[M]. 北京：中华书局，2013,12(1)：10-13.

[11] 王兆鹏，郁玉英，郭红欣. 宋词排行榜[M]. 北京：中华书局，2012,1(1)：15-17.

[12] 李白诗歌与盛唐音乐[J]. 文学遗产，1995(3)：43-51.

[13] 晏红，罗琦卿. 姜夔曲词的音乐审美特征[J]. 江西社会科学，2015(6)：90-94.

[14] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004：487-494.

[15] Steyvers M, Griffiths T. Probabilistic topic models[J]. Handbook of Latent Semantic Analysis, 2007, 427(7)：424-440.

申资卓(1994—)，硕士研究生，主要研究领域为自然语言处理。

E-mail：1078926191@qq.com

邵艳秋(1970—)，通信作者，博士，教授，主要研究领域为自然语言处理。

E-mail：shaoyanqiu@blcu.edu.cn

杨莹(1993—)，硕士研究生，主要研究领域为计算语言学。

E-mail：13161731988@qq.com