文 献 与 数 据 学报
第 4 卷　第 3 期　　2022 年 9 月

基于文本分类和主题模型的文化遗产信息资源知识发现方法

彭　博[1, 2]
（1. 华中科技大学建筑与城市规划学院，武汉　430074；
2. 华中师范大学信息管理学院，武汉　430079）

摘　要：[**目的 / 意义**] 如何详尽地为受众挖掘文化遗产信息资源文本中蕴含的有关知识，成为了中华历史文化传承与推广中的重要问题。[**方法 / 过程**] 文章提出基于文本分类和主题模型的文化遗产信息资源知识发现框架：针对文本特征将文化遗产信息资源分类，使用关键词抽取方法获取信息资源内容有关的关键词，而后与知识图谱进行知识耦合，根据信息资源内容特征融合不同关键词抽取方法进行知识发现。文章以《清明上河图》信息资源文本为例，对知识发现方法进行实验。[**结果 / 结论**] 融合后的知识发现方法较单一方法在知识实体的发现数量以及实体关系的发现数量上均有提升。实验表明，依照信息资源内容特征的不同对其进行分类，在此基础上使用有针对性的关键词抽取方法，能够显著提高文化遗产文本知识发现效率。

关键词：知识图谱　文本分类　主题识别　知识发现
分类号：G254；TP391.3
DOI：10.31193/SSAP.J.ISSN.2096-6695.2022.03.05

近年来出现的“文博热”成为了文化领域的突出现象之一，各种文博类专题节目、展览、讲座、文章等受到了社会各界的广泛关注与热议。然而这些对象所承载的信息资源却有着专业性强、分类复杂、动态累积、资源散乱等特点[1]，这些特点增加了文化遗产信息资源受众理解以及掌握相关知识的难度。因此，如何从文化遗产信息资源中准确且高效地挖掘出其中隐含的文化遗产知识并直接传递给用户，成为了发扬和传承中华优秀传统文化的关键问题之一。

1　相关研究

1.1　国内外文化遗产信息资源的有关研究

目前国内外有关文化遗产信息资源的研究主要集中在数字化与数字化后信息资源的利用两方

[作者简介] 彭博（ORCID：0000-0003-0262-3095），男，实验员，博士研究生，研究方向为信息资源管理，Email：pbresearch@sohu.com。

面。前者多集中在如何将现有的文化遗产资源通过数字化形式保存。Filip 等[2]认为，在复杂信息环境和丰富通信产品的影响下，文化遗产有关信息通过虚拟展示的方式能够在竞争日益激烈的社会和经济环境中，取得文化遗产机构、文化创意产业、政府和公众间的双赢，发展更容易获得和自我维持。在国内，文化遗产数字化的有关研究与应用起始于 2000 年前后。刘刚等[3]将计算机三维扫描技术应用于敦煌莫高窟石窟的三维数字化工作中，通过数字摄影测量和三维激光扫描进行石窟的三维信息获取、建模、纹理绑定等，为使用数字化手段进行有关研究提供了基础。王婷[4]利用三维激光扫描技术建立三维模型，以秦兵马俑一号坑陶俑为例进行了数字化模型展示、分析与研究，为数字博物馆的建立、藏品的鉴赏和保存提供了新的方式。国内外学者在文化遗产数字化方面的研究已经进行的十分深入，三维建模、虚拟现实、模拟仿真、地理信息系统等技术已经被广泛应用于数字化工作中。

而文化遗产信息资源利用的有关研究主要集中在数据标准制定与发布以及从语义角度对信息资源中的知识进行组织两方面。Van 等[5]认为，文化遗产保护机构的两个重要使命是保存历史文化遗迹资源和将已发现的遗迹、遗产尽可能展现给受众，然而事实上这两个使命存在极大冲突，因此信息资源元数据标准的提出为化解两种使命间的冲突提供了新方法，为文化遗产管理机构提供了一个完整展现文化遗产信息的标准。Tsai 等[6]将非结构化信息资源通过语义标注、焦点上下文等方法转化为文物元数据的著录格式。Hu 等[7]从用户角度构建了针对壁画和石窟的不可移动文化遗产的元数据模型，尤其强调了文化遗产信息的编目标准和元数据模式之间的互操作性问题是提高文化遗产信息数字化平台用户体验的关键。Matusiak 等[8]通过翻译后的词汇映射构建了多语言的元数据记录，进行了文化遗产资料多语种索引的研究和实践，拓宽了同一元数据标准下文化遗产信息资源的来源途径。龚花萍等[9]融合多种文物元数据标准，提出了针对文物信息资源元数据的著录标准，构建了文物信息资源的元数据模型。艾雪松等[10]通过元数据标准的复用构建了针对博物馆领域的文物信息资源元数据模型。Hyvönen[11]使用语义网整理了文化遗产有关的数据资料，为文化遗产信息资源的语义挖掘提供了数据基础。De Boer 等[12]使用关联数据有关技术采集、存储、加工与发布具有语义关联的文化遗产数据。曾子明等[13]针对文化遗产多媒体资源、视频资源中的潜在语义关联进行了文化遗产知识组织研究。

从上述研究可以发现，目前国内外学者有关文化遗产信息资源的研究已经发展到从多来源、多标准、多语言融合的视角进行，但大多从元数据角度基于结构化数据开展研究，展现文化遗产有关信息，对于非结构化数据构成的文化遗产信息资源研究较少。

1.2　文本分类与知识发现的有关研究

“实体–关系–实体”三元组是知识的基本载体，广泛分布在各种类型的数据中，通常情况下结构化与半结构化数据由于拥有明确的组织结构和统一的著录标准，比较容易挖掘其中的知识[14]。但是众多的非结构化数据中也蕴含有大量的知识，完整且精确地挖掘这些知识能够极大地扩充知识来源的范围，提高领域知识研究中知识挖掘的能力[15]。有学者发现大量的、无结构的文本信息中蕴含有大量的知识，对文本中的知识进行挖掘对于知识管理有着重要的意义[16]。在文本分类与知识发现的有关研究中，范宇中等[17]认为知识库中的领域知识是进行文本分类的重要手段，基于知识的系统能够进行快速、准确的文本分类。Hu 等[18]结合专利术语与专利文本

的主题挖掘，进行了专利技术的自动分类研究。Kim 等[19]通过对文档的总结，发现了专利技术中的共同特征，并在此基础上对专利中的知识进行挖掘。夏立新等[20]对网络文本进行挖掘，构建知识与就业岗位间的关联关系，提高了网络招聘信息的利用效率。王燕鹏等[21]对含有相同知识的文献进行聚类，使用标记的关键词作为知识热点并构建知识网络，采用结构洞理论分析网络和节点特性，以此识别潜在的新兴技术。以上研究表明，利用知识间关联关系结合文本挖掘技术，能够将相似文本进行聚类，提高文本挖掘的效率；同样地，利用文本的内容特征进行分类，能够增强文本中知识的标记特征，提高知识发现效率。

综合上述文献可以发现，现有方法多从信息资源的结构化数据中的知识角度进行研究，其研究对象是经过标注和整理后的文化遗产信息资源。然而受众在了解文化遗产知识的过程中多以各类型描述文本为主，从用户角度来说其更需要的是对文本中知识的提炼与直观的展示。因此，如何从非结构化文化遗产信息资源中发现和挖掘相关知识并呈现给受众，成为了智慧数据时代所需要解决的新问题。而文化遗产信息资源中大量的结构化著录数据为知识发现提供了良好的数据基础与知识来源，因此，将知识库知识与文本分类、关键词抽取等技术结合，成为了进行非结构化数据知识发现有关研究的一条有效途径。

2　基于文本分类的文化遗产信息资源知识发现框架

有研究显示，同一主题文本中与内容有关的关键词分布各不相同，关键词抽取的效率与文本类型有着密切关系[22]。不同文本间质量各异，按照一定方法对文本进行聚类能够有效提高文本主题的挖掘效率[23]。在现有的面向文化遗产信息资源文本进行的文化遗产知识发现研究中，单一的关键词抽取方法仅能对部分类别的文化遗产信息资源进行有效的知识发现[24]。因此文章提出从文本中知识内容分布特点的不同出发，根据文本内容特征进行更为深入的主题识别与关键词抽取，力图从有限数量的关键词集合中尽可能挖掘在文本主题中与之有关的知识，将识别后的主题关键词与知识图谱进行知识耦合，实现非结构化数据信息资源的知识发现，其过程如图 1 所示。

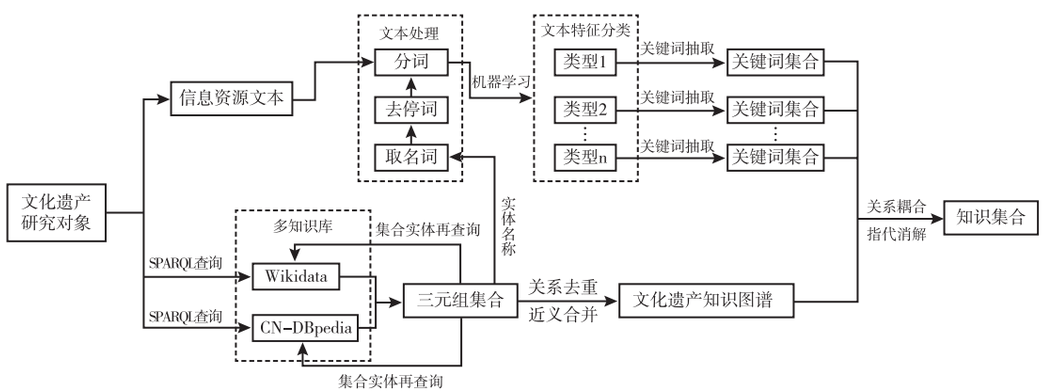

图 1　基于文本分类的文化遗产信息资源知识发现框架

文章进行文化遗产信息资源知识发现的过程主要可以分为三部分：一是文化遗产知识的获取，以信息资源的研究对象为切入点，从外部知识库中抽取文化遗产知识并构建知识图谱，作为知识发现工作的知识数据来源；二是文化遗产信息资源文本的处理，将非结构化数据组成的信息资源转化为与内容有关的关键词集合，作为发现知识三元组的基础对象；三是文化遗产信息资源的知识发现，该部分将信息资源中抽取到的关键词与知识图谱的知识三元组进行耦合，获得与信息资源主要内容有关的知识集合，完成文化遗产信息资源的知识发现。

2.1 文化遗产知识图谱的构建

知识图谱是实体与其有关概念的集合，由描述实体、属性间关系的三元组构成。知识的本质是人类在实践中认识客观世界的成果，“实体–关系–实体”三元组可以被看作是知识的载体。文化遗产信息资源知识具有唯一性、标准化特征，知识图谱的构建是进行知识发现的前提，也是目前学者们研究的重点。文章在多知识库中使用 SPARQL 查询与研究对象有关联的实体并根据实体间关联关系进行多次查询，参考元数据构建方法合并与去重后得到基于研究对象实体的三元组集合[25]。如图 2 所示，查询的过程是以文化遗产实体在知识库中的唯一命名为入口，获取所有有关实体及关联关系。在得到集合后，通过图模型将三元组中的实体及属性映射为节点和边，其映射过程可以表示为（S，P，O）$\rightarrow G_i=(V_n, E_m)$，其中 $V=\{S\}\cup\{O\}$，$E=\{(S\rightarrow O)\}$，S 和 O 代表获取知识三元组中的两个关联实体，P 代表实体间关联关系或属性，V 代表知识图谱中的节点，与三元组中的实体对应，边 E 的标签表示为 P，构建面向研究对象的知识图谱。

```
SELECT ?p ?plabel ?o ?olabel
WHERE
{
  wd:Q714802 ?p ?o .
  SERVICE wikibase:label { bd:serviceParam wikibase:language "zh". }
}
```

图 2　Wikidata 知识库查询语言示例

2.2 文化遗产信息资源的文本处理及分类

文本的关键词是信息资源主要内容的概括，文本中与主题有关的知识也大多包含在其中，有针对性地高效抽取文化遗产信息资源文本的关键词对于文化遗产知识的挖掘有着至关重要的作用，也是进行知识发现的首要工作。

不同来源与类型的信息资源包含的文化遗产知识不尽相同，即便是描述同一对象的信息资源文本也各有侧重。在文本信息挖掘中，针对不同文本进行分类能够显著提高词义消歧、智能检索、主题识别的计算效果[26]。文章依据文化遗产信息资源来源的不同，将其划分为两类。一类是以普及文化遗产知识为主的百科类文本，如对文化遗产的年代、作者、传承、主要描述内容、主要特征等知识点的介绍，知识分布宽泛。同时，由于互联网的开放性特征，部分百科类文本内容存在内容重复、质量较低等特征，采用基于主题的关键词抽取方法能够较好地获取该类信息资源文本中与主题有关的关键词。另一类是深入研究文化遗产有关发现的论文、说明、公告等研究类信

息资源文本。相较于百科类文本，发表在专业期刊中的研究型文化遗产信息资源文本会针对实体的某一特征或属性进行深入的分析，提出新发现，创造新知识，该类型文本一般不会就研究对象的基本信息进行过多描述。同时，由于期刊文献刊载要求，研究型文化遗产信息资源文本重复率较低，采用基于统计的关键词抽取方法能够捕捉不同文本间的差异性，挖掘信息资源主题中独立出现的关键词。该类关键词往往是深入研究后获取新知识的代表。由于不同类型文本差异性较大，文章提出通过已知分类的信息资源作为训练集，利用机器学习对未知来源数据进行分类，在进行知识发现前对文本进行分类，有针对性地选择关键词抽取方法，提高知识发现效率。

2.3　文化遗产信息资源的知识发现

文本的主题中包含着创作者所讨论的实体或概念[27]，这些实体或概念之间如果进行关系标注就构成了知识，知识标注得越详尽就越利于用户理解。在面对数量众多、描述对象各异、承载内容复杂的文化遗产信息资源时，文章依据信息资源中知识分布的不同特征进行分类，采用不同的关键词抽取方法进行主题关键词的挖掘。在得到文本主题关键词后，文章使用 $Topic_m$ 表示主题关键词集合，即 $Topic_m$={$topic_1$、$topic_2$……$topic_n$}，知识发现的过程是将 3.1 小节中知识图谱 G_i=（V_n，E_m）中的节点标签与主题挖掘得到的主题关键词集合求交集，其结果为 $K=V_i \cap Topic_m$，K 表示与知识图谱中节点标签一致的关键词，耦合后的知识发现结果集合为 R={（K_1，E_1，K_2），（K_2，E_2，K_3）……（K_{n-1}，E_m，K_n）}，挖掘过程如图 3 所示。

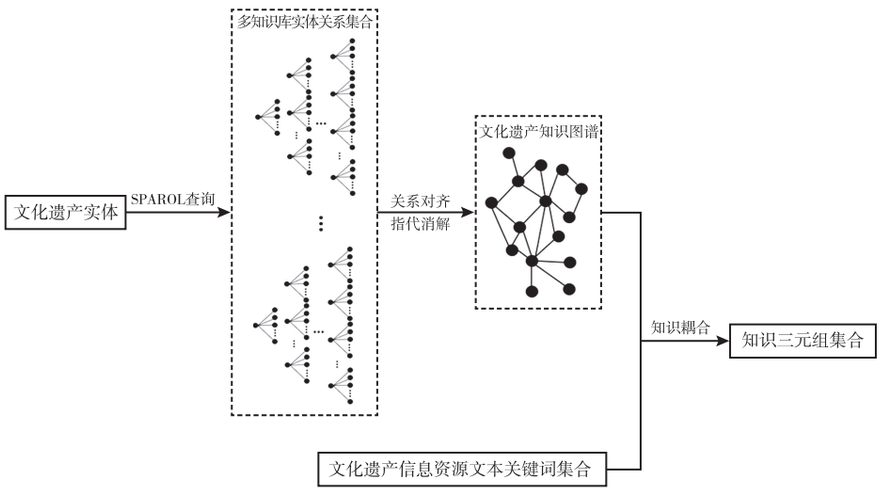

图 3　基于多知识库的文化遗产信息资源知识发现过程

3　文化遗产信息资源知识发现实验

为了验证文本分类与主题识别在文化遗产信息资源知识发现中的效果，文章选择中国十大传世名画之一的北宋风俗画《清明上河图》有关信息资源文本全文为例进行研究。选择全文为研究对象的原因主要有两点：一是全文对于研究对象的描述与标题和摘要相比更为详尽，能够最大限

度地挖掘信息资源中与文化遗产有关的知识；二是本文研究的文本分类与关键词抽取方法能够更好地面向全文发挥作用，捕捉信息资源文本中的特征，进行知识发现。

文章收集的文化遗产信息资源文本可以分为两种来源。一种为百科类文化遗产信息资源文本。文章选取百度百科、维基百科、搜狗百科等关于“清明上河图”词条文本以及在中文网络知识社区“知乎”中与“清明上河图”话题有关的文章，共得到百科类文化遗产信息资源文本 50 篇。第二种为研究类文化遗产信息资源文本。文章在网络学术平台“中国知网”中以“清明上河图”为主题检索期刊文献，经过去除无关文献后获得研究类文化遗产信息资源文本 50 篇。以这两种来源（共计 100 篇文献）为研究对象进行知识发现实验，该样本基本代表了目前互联网中与《清明上河图》有关信息资源的全貌。

3.1 文化遗产知识图谱构建

在知识图谱构建中，文章首先在网络知识库 Wikidata 中以“清明上河图”（编号 Q714802）为检索入口，得到与“清明上河图”节点距离小于等于 3 的节点所构成的三元组 380 对。随后在中文知识库 CN–DBpedia 中以“Named–Entity Disambiguation: 清明上河图（北宋张择端风俗画）”进行 3 次检索得到三元组 108 对。以 3 次检索为阈值是由于检索获取的三元组数量呈指数型增加趋势，多于 3 次的检索后会获取大量与文化遗产实体无关的三元组，影响知识图谱构建质量。最后，经去重后构建拥有 401 个节点、409 条边的文化遗产知识图谱，结果如图 4 所示。

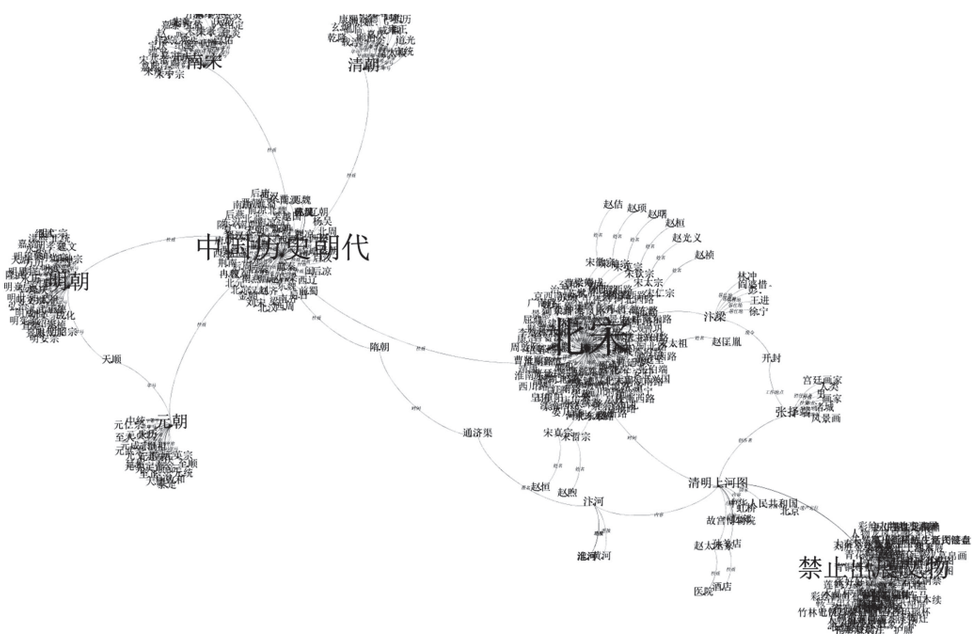

图 4　文化遗产信息资源知识图谱

3.2 文化遗产信息资源的特征及分类

对文化遗产信息资源文本进行分类是有针对性地获取主题关键词的首要步骤。首先对语料文本进行分词，停用词库使用了常用的哈尔滨工业大学停用词表以及百度停用词库，使用停用

词1 893个。取名词依据文化遗产知识图谱中的节点标签确定，401个节点名称列为白名单词语，采用jieba分词工具对文本进行分词。分词后选择TF-IDF[28]值作为文本中关键词的特征表示构建文档-词项矩阵，每个矩阵行代表一篇文化遗产信息资源文本。在分类模型的选择上，支持向量机[29]（Support Vector Machine，SVM）、逻辑回归[30]（Logistic Regression，LR）、朴素贝叶斯[31]（Naive Bayesian，NB）是较为常用且理想的选择[32]。使用这三种分类模型，通过Python程序语言中的scikit-learn包将实验数据集中的100篇文档按8:2的比例分为训练集和测试集，即80篇作为训练集，20篇作为测试集进行百科类和研究类文化遗产信息资源文本的分类，其结果如表1所示。

表 1　《清明上河图》文化遗产信息资源文本分类结果

	SVM	LR	NB
测试集正确率	65%	70%	80%
测试集 AUC 值	0.824 176	0.824 176	0.846 154

从表 1 可以发现，朴素贝叶斯分类模型效果最好，测试集正确率达到了 80%。由于实验中文本分类为二分类，文章使用 ROC（Receiver Operating Characteristic）即接受者操作特征曲线与曲线下面积（Area Under Curve，AUC）对分类效果进行评价[33]。如图 5 所示，平面的横坐标是 False Positive Rate（FPR），纵坐标是 True Positive Rate（TPR），横坐标代表测试集中被错误分到正样本类别中真实的负样本所占负样本总数的比例，纵坐标代表测试集中真实的正样本所占正样本总数的比例。对于分类模型可以根据其在测试样本上的表现得到 TPR 和 FPR 点对，分类模型就映射成 ROC 平面上的一个点。AUC 值为 ROC 的曲线下面积，AUC 值越接近 1 说明分类模型的效果越好，三种分类模型 AUC 值均大于 0.8，说明文章对文化遗产信息资源百科类和研究类的划分符合客观规律。

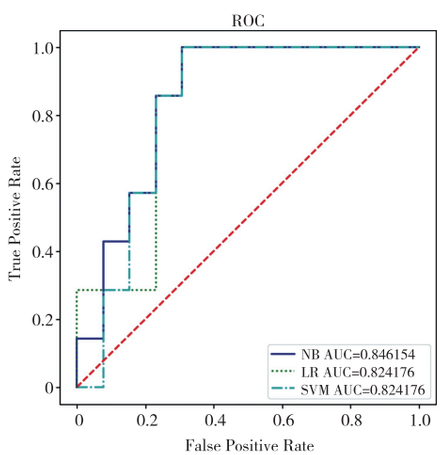

图 5《清明上河图》文化遗产信息资源文本分类效果评价图

3.3 文化遗产信息资源的关键词抽取

关键词是文本类信息资源主要内容的体现，文章对不同分类下的文本采用针对性关键词抽取方法以获取与主要内容有关的实体。根据不同关键词抽取方法的算法特征，对主题较为集中的一类文本采用基于主题的关键词抽取方法；对主题较为分散的文本采用基于统计的关键词抽取方法，从文本间的差异化内容中挖掘与各自主题有关的实体。

针对主题较为分散的一类文本，文章选取基于统计特征的 TF–IDF［33］算法，计算公式如公式 1 所示。G 代表文档总数，n_t 为包含特定词语 t 的文档数，TF 代表词频，IDF 是包含词语文档数与总文档数的对数。一个词语的重要性随其在文档中出现频率的增加而增加，但随其在文档集中出现的总频率的增加而减小。

$$TF-IDF = TF_{ti} \cdot IDF_t = \frac{n_{ti}}{N} \cdot \log \frac{|G|}{n_t} \quad (1)$$

针对主题较为集中的一类文本，文章选择 LDA［34］主题模型进行关键词抽取。该模型通过概率大小提取文档关键词，每个特征词在文档中出现的概率为：

$$p（特征词 | 文档）= \sum 主题\, p（特征词 | 主题）\times p（主题 | 文档） \quad (2)$$

同时，文章选择基于词语间关系的 TextRank［35］作为对比算法，其按文本窗口来统计文本中词之间的关系，提取片段阈值中关键词的共现关系，迭代计算各词在网络中的权重，而后以各词的权重进行排序，其迭代算法如公式 3 所示。其中，$WS（V_i）$ 表示节点 V_i 的权重，$\mathrm{In}（V_i）$ 表示 V_i 的入度，$Out（V_j）$ 表示 V_j 的出度，w_{jk} 表示节点 V_j 和 V_k 间的边权重，w_{ji} 表示节点 V_j 和 V_i 间的边权重，$WS（V_j）$ 表示节点 V_j 的权重，d 为阻尼系数，取值范围为 0 至 1。

$$WS(V_i) = (1-d) + d * \sum j \in \mathrm{In}(v_i) \frac{w_{ji}}{\Sigma v_k \in out(v_j) w_{jk}} WS(v_j) \quad (3)$$

运用三种有代表性的关键词抽取方法与主题模型对测试集中的 20 篇文化遗产信息资源文本进行关键词抽取的比较研究。在关键词数量的选取上，由于最短一篇文档关键词数量为 323，文章按 10% 的比例选择关键词数 K=30，通过不同关键词数分别进行知识发现，部分结果如表 2 所示。

3.4 文化遗产信息资源的知识发现

在得到《清明上河图》文化遗产信息资源文本的关键词后，将关键词与知识图谱中的知识节点进行耦合以实现信息资源的知识发现。耦合规则如下：当只有单一主题关键词与知识图谱中的节点匹配成功时，选择核心节点“清明上河图”与匹配节点路径中知识节点的集合作为知识发现结果，这时知识集合 R 可以表示为 R={（K_1，E_1，K_2），（K_2，E_1，K_3）...（K_{m-1}，E_m，K_m）}，其中 m 为核心节点与匹配节点间的距离；而当有词对与知识图谱中的节点匹配成功时，知识集合 R' 则表示为 R'={（K_1'，E_1，K_2'），（K_2'，E_1，K_3'），...（K_{n-1}'，E_n，K_n'）}，其中 n（$n > 1$）表示两个

节点在知识图谱中的距离，耦合后得到的知识由“知识节点–关联关系–知识节点”三元组的方式表示。

表 2　《清明上河图》文化遗产信息资源文本关键词抽取列表（部分）

		TF–IDF	TextRank	LDA
文档 1	关键词 1–10	世俗、生活、院本、风貌、展示、社会、剧版、蓝本、观众、堪称	社会、生活、风貌、研究、世俗、观众、画作、风俗画、疫情、讲究	清明上河图、北宋、生活、社会、研究、院本、清、清平乐、画、幅
	关键词 11–20	研究、再现、风俗画、画作、剧刚、主演、奇观、治沉、静物、征服	展示、意境、院本、回复、用笔、圆熟、电视剧、堪称、时期、剧版	展示、风貌、世俗、宋仁宗、里、堪称、赵祯、风俗画、时期、一种
	关键词 21–30	服化、非常少、文人墨客、剧中、庙宇、朝堂、置身于、展厅、消息、列置	古装、开创、主演、独特、西式、社会现状、作者、房舍、作画、透视	剧版、观众、中、蓝本、画作、画家、展、宋代、再现、疏影
…	…	…	…	…
文档 20	关键词 1–10	候风、华表、木杆、相风、表木、装置、立鸟、身份、组合、风仪	候风、华表、木杆、装置、表木、功能、身份、相风、书写、立鸟	鸟、中、候风、华表、河图、乌、木杆、鹤、相风、装置
	关键词 11–20	风向、功能、图像、书写、记载、形态、图式、吹动、鸟头、梦华	记载、组合、后人、图像、制作、图式、形态、风向、位置、鸟头	身份、两只、立鸟、表木、记载、功能、组合、古代、宋代、立鹤
	关键词 21–30	制作、位置、发明、典故、方向、所绘、进一步、材料、后人、诽谤	方向、历史、材料、典故、吹动、所绘、风仪、做成、谓之、描绘	倪、图、风仪、书写、木、图像、形态、梦华、风向、录

经统计发现，LDA、TextRank 关键词抽取方法在百科类文化遗产信息资源中获得的知识相对较多，TF–IDF 在研究类文化遗产信息资源中获得的知识较多。由此可以发现不同关键词抽取方法针对不同类型文本的关键词抽取效果各异。为了对比研究，文章在原有三种方法抽取而来的关键词集合基础上，使用 LDA 对 20 篇中 12 篇被分类为百科类的文本以及 TF–IDF 对剩余 8 篇分类为研究类的文本的关键词抽取结果形成新的关键词集合，以及 TextRank 与 TF–IDF 按照同样的方法形成关键词集合，最后分别将识别结果汇总，使用 5 种不同模式的关键词抽取方法与知识图谱进行知识发现工作，其结果如图 6 所示。

在图 6 中圆形节点及连线表示 LDA 主题模型的知识发现结果，三角形节点及连线表示 TextRank 算法的知识发现结果，正方形节点及连线表示 TF–IDF 算法的知识发现结果，五边形节点及连线表示 LDA+TF–IDF 模式的知识发现结果，六边形节点及连线表示 TextRank+TF–IDF 模式的知识发现结果，节点连线间的标签代表节点之间的关联关系。从图中可以发现，五边形节点数量最多，这代表 LDA+TF–IDF 模式挖掘的实体数量最多，知识发现结果相对最好。据比较得知，经过文本分类后，LDA+TF–IDF 模式对文本中的实体关系发现数量最多，较单一方法中表现最好的 LDA 主题模型的实体关系发现数量提升近 30%，知识实体发现数量提升近 50%。这说明

对不同类型文化遗产信息资源文本进行分类，而后根据文本特征选取针对性的关键词抽取方法，能够显著提高通过主题进行知识发现的数量与效率。

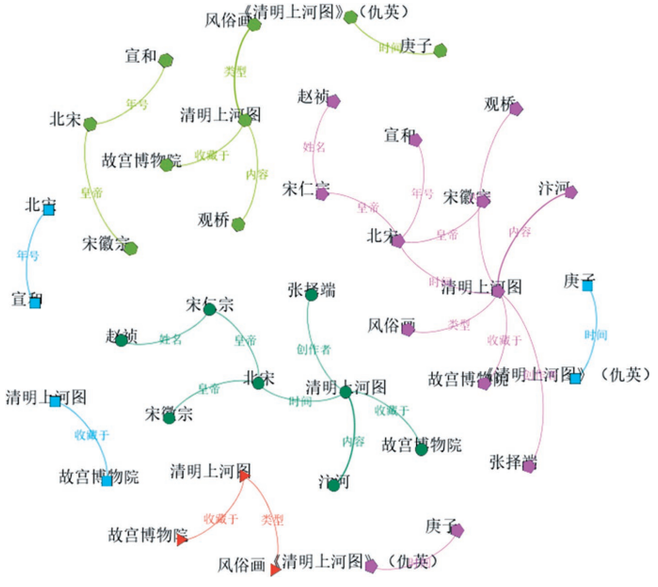

图 6 《清明上河图》文化遗产信息资源知识发现结果

为了进一步研究关键词抽取方法在文化遗产信息资源知识发现中的效果，文章采用 P@K［36］指标对关键词抽取的结果与知识关联程度进行评价，其中 K 的取值为 K=［10、20、30］。P@K 指前 K 个关键词的正确率，也就是和知识实体的匹配率，其计算公式为：

$$P@K = \frac{r(k)}{k} \tag{4}$$

其中，r (k) 表示抽取到的前 K 个关键词中有效关键词的数量，在文中计算为抽取出的关键词与实体耦合的数量。在知识发现的过程中有效关键词的数量越多、排序越靠前，则 P@K 指标越高，该关键词抽取方法也越容易发现文化遗产信息资源中存在的知识。

表 3 《清明上河图》文化遗产信息资源文本有效关键词实验结果

关键词抽取方法	P@10	P@20	P@30
LDA	0.034 96	0.029 09	0.022 11
TextRank	0.013 16	0.009 87	0.006 79
TF–IDF	0.023 12	0.017 75	0.011 95
LDA+TF–IDF	0.058 82	0.038 96	0.030 84
TextRank+TF–IDF	0.025 64	0.016 13	0.013 13

由 P@K 指标可以发现，文章对文化遗产信息资源进行分类后采取 LDA+TF–IDF 分别抽取百科类与研究类文本关键词的方法，在所有指标上均优于其他关键词抽取方法。但随着抽取关键词阈值的上升，其他方法与 LDA+TF–IDF 方法的差距逐渐缩小，这说明随着关键词抽取量的增加，与文化遗产知识有关的关键词数量减少，知识发现效率也随之降低。

3.5　结果分析

（1）从知识发现的实体间关系数量上看，采用不同的关键词抽取方法对分类后的文化遗产信息资源文本进行知识发现，所产生的效果各异。主题模型 LDA 对常见的百科类文化遗产信息资源知识发现结果较好，这是由于百科类文化遗产信息资源主要描述实体的基本知识，目的明确，语言简单，同时文本中存在大量的常用语，经分词过滤后更能突出其主要内容。基于统计特征的 TF–IDF 模型由于其逆文档特性，会忽略普遍存在于文本中有关描述对象基本情况的介绍部分，其在百科类文化遗产信息资源文本中的知识发现效果较差也验证了这一点。相对应的，研究类文化遗产信息资源文本由于其受众的专业性特征，通常不会就研究对象的基本情况作过多介绍，使用 TF–IDF 模型忽略其共同存在的关键词，反而会突出其研究内容中更加深入的部分，致使其在研究类文本分类中知识发现效果较好。两类文本混合后，由于百科类文本对其词频矩阵的干扰，使得研究类文本中深入研究的部分主题权重下降，导致其对实验集整体的知识发现效果较差。实验结果表明，使用不同的关键词抽取方法进行知识发现的效果最好，单一主题模型无法胜任混合类文化遗产信息资源文本的知识发现工作。

（2）从知识发现的内容上看，LDA 主题模型发现的主题关键词如“北宋”“故宫博物院”等多属于研究对象的基础知识，说明该模型对浅层的文化遗产知识发现效果较好。而 TF–IDF 模型对研究类文化遗产信息资源文本发现的主题关键词有“宣和”“汴河”等较为深入的知识点，这些关键词也存在于部分较为详尽的百科类文本中，但处于文本中词频统计靠后的位置。以上结果说明：百科类文化遗产信息资源文本按照一般的关键词抽取方法就能够对其描述对象进行主要知识内容的概括；而研究类文化遗产信息资源文本则需要进行一定的统计处理，去除部分概括浅层次知识的内容才能够对其所表达的主题知识进行揭示，因此相对于其他关键词抽取方法，TF–IDF 模型能够更好地挖掘研究类文化遗产信息资源中较为深入的文化遗产知识关键词。

（3）从知识发现的数量角度看，所有方法的知识发现效果均随关键词阈值的增加而降低。这一方面说明通过关键词与知识图谱耦合的知识发现方法能够有效挖掘文化遗产信息资源文本中的主要内容与主要知识点；另一方面也说明若要发现文本中蕴含的更深层次知识，需要融合更多的关键词筛选方法以控制阈值增加所带来的无关关键词干扰问题，若仅通过增加主题关键词阈值的模式反而会降低挖掘效果，增加计算难度。

（4）从知识发现在文化遗产管理中的作用看，由于现有的文化遗产展示过程多以非结构数据构成的文本为主，若不直接对其中的文化遗产知识进行揭示是无法满足受众需求的。本研究发现了一种效率较高的信息资源文本知识发现方法，围绕文本内容进行隐性知识的发现与展示，将外部知识库与知识图谱进行结合，拓展有关机构进行藏品管理的途径，拓宽文化遗产知识推广的方法，提供文化遗产知识研究的新视角，提出文化遗产智慧化管理的新手段，促进文化遗产管理工作的发展。

4 结 论

中华文化博大精深，随着互联网及各种类型传播平台的推广，数量众多的文化遗产信息资源作为媒介将原本晦涩难懂的文化遗产知识推介给广大受众。但由于对领域知识的了解是一个由浅入深、循序渐进的过程，受众的知识水平极大地影响其理解能力。通过不同类型文本的分类知识发现，快速且精准地挖掘文化遗产信息资源中的知识，能够更好地帮助受众由浅入深地全面了解文化遗产知识，提高受众对文化遗产知识的理解能力，弘扬和推广中华传统文化。

文章通过将分类算法与关键词抽取方法融合，提出了具有针对性的文化遗产信息资源知识发现方法。在以《清明上河图》为例的文化遗产信息资源知识发现实验中，多类型融合的知识发现方法比单一类型方法在知识实体发现与实体关系挖掘上均有较大提升，取得了较好的知识发现效果。

在下一步的研究中，需要对关键词阈值增加后取得更好的知识发现效果进行深入的研究，同时还要扩大与方法对应的知识发现对象的适用范围，提升融合文本分类与关键词抽取的文化遗产信息资源知识发现方法的使用效果及应用范围。

【参考文献】

［1］肖花，刘春年. 文物信息资源分类与特征分析［J］. 现代情报，2012，32（10）: 8–11，92.

［2］Filip F G, Ciurea C, Dragomirescu H, et al. Cultural heritage and modern information and communication technologies［J］. Technological and Economic Development of Economy, 2015, 21(3): 441–459.

［3］刘刚，张俊，刁常宇. 敦煌莫高窟石窟三维数字化技术研究［J］. 敦煌研究，2005（4）: 104–109.

［4］王婷. 文物真三维数字建模技术在秦始皇兵马俑博物馆中的应用——以一号坑陶俑为例［J］. 文物保护与考古科学，2012，24（4）: 103–108.

［5］Van Hooland S, Rodríguez E M, Boydens I. Between commodification and engagement: On the double-edged impact of user-generated metadata within the cultural heritage sector［J］. Library Trends, 2011, 59(4): 707–720.

［6］Tsai C M, Qamra A, Chang E Y, et al. Extent: Inferring image metadata from context and content［C］//2005 IEEE International Conference on Multimedia and Expo. IEEE, 2005: 1270–1273.

［7］Hu X, Ng J, Xia S. User-centered evaluation of metadata schema for nonmovable cultural heritage: Murals and stone cave temples［J］. Journal of the Association for Information Science and Technology, 2018, 69(12): 1476–1487.

［8］Matusiak K K, Meng L, Barczyk E, et al. Multilingual metadata for cultural heritage materials［J］. The Electronic Library, 2015, 33(1): 136–151.

［9］龚花萍，孙晓，刘春年. 文物信息资源元数据模型、实施标准与应用策略［J］. 情报科学，2015，33（2）: 80–84.

［10］艾雪松，石宪，彭超，等. 文物信息资源元数据模型构建与应用研究［J］. 情报科学，2019，37（6）: 69–74.

［11］Hyvönen E. Publishing and using cultural heritage linked data on the semantic web［J］. Synthesis Lectures on the Semantic Web: Theory and Technology, 2012, 2(1): 1–159.

［12］De Boer V, Wielemaker J, Van Gent J, et al. Supporting linked data production for cultural heritage institutes: the amsterdam museum case study［C］//Extended Semantic Web Conference. Springer, Berlin, Heidelberg, 2012: 733–747.

［13］曾子明，周知，蒋琳. 基于关联数据的数字人文视觉资源知识组织研究［J］.情报资料工作，2018, 39

（6）: 6–12.

［14］邵作运，李秀霞. 引文分析法与内容分析法结合的文献知识发现方法综述［J］. 情报理论与实践，2020，43（3）: 153–159.

［15］王颖，吴振新，谢靖. 面向科技文献的语义检索系统研究综述［J］. 现代图书情报技术，2015（5）: 1–7.

［16］韩客松，王永成. 文本挖掘、数据挖掘和知识管理——二十一世纪的智能信息处理［J］. 情报学报，2001（1）: 100–104.

［17］范宇中，张玉峰. 文本知识的自动分类方法初探［J］. 情报科学，2003（1）: 103–105.

［18］Hu Zhengyin, Shu Fang, Tian Liang. Empirical study of constructing a knowledge organization system of patent documents using topic modeling［J］. Scientometrics, 2014, 100(3): 787–799.

［19］Kim J H, Choi K S. Patent document categorization based on semantic structural information［J］. Information Processing & Management, 2007, 43(5): 1200–1215.

［20］夏立新，楚林，王忠义，等. 基于网络文本挖掘的就业知识需求关系构建［J］. 图书情报知识，2016（1）: 94–100.

［21］王燕鹏，韩涛，陈芳. 融合文献知识聚类和复杂网络的关键技术识别方法研究［J］. 图书情报工作，2020，64（16）: 105–113.

［22］王健，张俊妮. 统计模型在中文文本挖掘中的应用［J］. 数理统计与管理，2017，36（4）: 609–619.

［23］王莉军，姚长青，刘志辉. 一种文本挖掘和文献计量的科技论文评估方法［J］. 情报科学，2019，37（5）: 66–70.

［24］高劲松，彭博. 基于主题识别的文物信息资源知识发现方法研究［J］. 情报科学，2021，39（4）: 9–14.

［25］台钰莹，王乐春，杨东波. 元数据标准登记系统平台构建——以文物行业为例［J］. 图书馆建设，2019（S1）: 15–19.

［26］王春柳，杨永辉，邓霏，等. 文本相似度计算方法研究综述［J］. 情报科学，2019，37（3）: 158–168.

［27］林杰，苗润生. 专业社交媒体中的主题图谱构建方法研究——以汽车论坛为例［J］. 情报学报，2020，39（1）: 68–80.

［28］Ramos J. Using TF–IDF to determine word relevance in document queries［C］//Proceedings of the First Instructional Conference on Machine Learning. 2003, 242: 133–142.

［29］Suykens J A K, Vandewalle J. Least squares support vector machine classifiers［J］. Neural processing letters, 1999, 9(3): 293–300.

［30］Ng A Y, Jordan M I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes［C］//Advances in Neural Information Processing Systems, 2002: 841–848.

［31］Kononenko I. Semi–naive Bayesian classifier［C］//European Working Session on Learning. Springer, Berlin, Heidelberg, 1991: 206–219.

［32］丁晟春，俞沣洋，李真. 网络舆情潜在热点主题识别研究［J］. 数据分析与知识发现，2020，4（Z1）: 29–38.

［33］Aizawa A. An information–theoretic perspective of TF–IDF measures［J］. Information Processing & Management, 2003, 39(1): 45–65.

［34］Goldberg Y, Levy O. Word2vec explained: Deriving Mikolov et al.'s negative–sampling word–embedding method［J］. arXiv preprint arXiv:1402.3722, 2014.

［35］Mihalcea R, Tarau P. TextRank: Bringing order into text［C］//Proceedings of the 2004 Conference on

Empirical Methods in Natural Language Processing. 2004: 404-411.

［36］Davis J, Goadrich M. The relationship between precision-recall and ROC curves［C］//Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 233-240.

Research on Knowledge Discovery Method of Cultural Heritage Information Resources Based on Text Classification and Topic Models

Peng Bo[1,2]

(1.School of Architecture and Urban Planning, Huazhong University of Science and Technology, Wuhan 430074, China;
2.School of Information Management, Central China Normal University,Wuhan 430079, China)

Abstract:［**Purpose/significance**］How to mine the relevant knowledge contained in the texts of cultural heritage information resources for users has become an important issue in the inheritance and promotion of Chinese historical culture.［**Method/process**］This paper proposes a knowledge discovery framework of cultural heritage information resources based on text classification and topic models: Different types of cultural heritage information resources are classified according to text features, and different topic models are used to couple the knowledge, then knowledge is discovered by integrating different keyword extraction methods according to the content features of information resources. This paper takes the information resources concerned “Riverside Scene at Qingming Festival” as a case study for the proposed knowledge discovery method.［**Result/conclusion**］Compared with the single method, the integrated method can discover more knowledge entities and more entity relationships. The experiment shows that the effectiveness of knowledge mining can be improved by classifying the texts of cultural heritage information resources according to their content characteristics and then using targeted keyword extraction methods according to different types of cultural heritage information resources.

Keywords: Knowledge map; Text classification; Topic recognition; Knowledge discovery

（本文责编：周　霞）