基于LDA的信息资源挖掘与可视化研究

丁玲[1] 叶佳鑫[2] 曾婷[1]
(1. 武汉市国土资源和规划信息中心，武汉 430014；2. 华中师范大学信息管理学院，武汉 430079)

摘要：LDA (Latent Dirichet Allocation) 是一种从文档资源中抽取主题的概率模型，将其用于文档的主题提取通常具有不错的效果。档案信息资源是一种具有较高利用价值的文档资源，但其目前存在碎片化、建设不足等问题。基于此，本文将LDA与聚类、层次空间构建技术相结合应用于档案信息资源建设，并进行实证研究。从实验结果来看，将LDA应用于档案信息资源建设可以挖掘资源间的隐含联系，明确资源间的等级层次，并有助于信息资源的可视化展示。

关键词：信息资源建设；主题提取；聚类；层次空间
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2019.02.005

随着信息技术的不断发展与大数据时代的来临，信息资源的数字化转型已成为目前的重要任务之一，而数字档案资源的建设无疑是数字化转型中重要的一环[1]。针对档案信息资源建设，全国档案事业发展“十三五”规划提出了相应措施，规划要求提升档案资源利用的便捷性以及加快档案管理信息化进程[2]。目前档案信息资源存在碎片化、用户对档案信息价值认识较低、档案资源间相关性难以发现、档案信息资源整合不足等问题[3]。因此，很有必要对档案信息资源进行挖掘与建设。LDA主题模型是一种针对文档资源的主题抽取模型，本文尝试将其与聚类、层次空间构建等数据挖掘技术结合并应用于档案信息资源建设，以提高档案资源利用的便利性，帮助用户及档案工作者更好地使用档案信息资源；同时，也为LDA主题模型在信息资源建设中的应用方向提供参考。

1 信息资源建设模型框架

本文主要通过主题提取、聚类、层次空间构建3种技术方法对文档类档案信息资源进行建设，信息资源建设模型框架见图1。

档案信息资源的碎片化使得用户对相关资源的利用变得困难，而为解决资源的碎片化就需要对零散的信息资源间关系进行挖掘。为了解决档案资源建设中存在

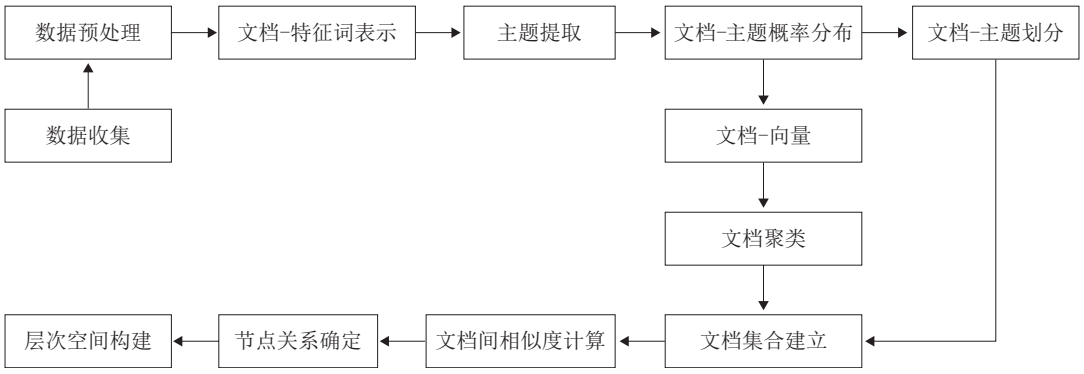

图1 信息资源建设模型框架

的相关问题，本文首先利用主题模型对资源进行主题提取，挖掘文档类资源蕴含的主题信息，以便从主题的角度进行资源的整合，将碎片化的信息资源转为主题表示；主题挖掘其局限在于只有资源具有明确的主题划分时，才能将其归入特定主题，而针对某些不属于任意主题的资源，其与其他资源间可能存在一定的相关关系，而主题提取无法对这种关系进行挖掘。为此，在主题提取后，采用聚类方法对档案资源进行进一步的相关性挖掘，以便更好地发现资源间联系，对主题表示进行补充与完善；实现相关资源的整合之后，如何从资源集合中选取重要的资源进行优先利用是资源建设的重要问题，从同类文档中优先发现重要文档可有效地提高资源利用效率。为进一步挖掘资源间关系，在主题提取与文档聚类的基础上，进一步计算文档间的相似度，并以此为基础来进行资源间的层次空间构建。

2 信息资源建设方法

2.1 基于LDA主体模型的信息资源主题提取

本文选用LDA来进行信息资源的主题提取，LDA是一种生成主题概率模型，常被用来处理大规模文档[4]。其思想源于一个基本假设，即文档由多个隐含主题构成，隐含主题由若干特征词构成。文档中每个词通过“以一定概率选择某个主题，并从这个主题中以一定概率选择某个语词”来得到[5]。生成一篇文档，其中每个词出现的概率都可以通过公式（1）来计算。

$$P(词|文档)=\sum_{主题}P(词|主题)\times P(主题|文档) \quad (1)$$

LDA是一个完备的主题模型，其文本生成方式可由图2的贝叶斯网络图来表示。LDA采用Dirichlet分布作为概率先验分布，模型中，K为文档的主题总数，M为文档集合，N是每篇文档中总词数，隐变量Z表示某一个主题，W为文本的单词，参数α和β分别是文档-主题概率θ以及主题-语词概率分布φ的先验分布超参数，W是唯一可观测的变量[6]。

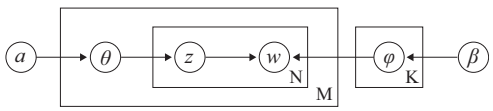

图2 LDA模型的贝叶斯网络图

LDA使用的是词袋思想，以某一概率选取主题，再以某一概率选出主题中每个单词，不断重复该步骤产生文档中所有语词。对词汇进行模糊聚类，聚集到一类的词可间接表示一个隐含主题。LDA挖掘了文本信息，衡量了不同文档间的潜在关系，也能用某一类词来表达隐藏主题。

2.2 基于主题挖掘的信息资源聚类

在主题挖掘的基础上，本文选择DBSCAN算法来实现文档资源的聚类。DBSCAN是一种经典的密度聚类算法，其优势在于可自动确定簇的数目，且能发现所有任意形状簇。其部分概念如下[7]。

（1）点的邻域。空间中任意一点P的邻域是以该点为圆心、以Eps为半径的圆区域内包含的点集合，记作$NEps(P)=\{q\in D|dist(p,q)\leqslant Eps\}$，集合中点P的最小个数由密度阈值minPts控制。

（2）噪声。数据库D中不属于任何类的点为噪声。基于定义，DBSCAN算法可以描述为[8]：①检测数据集中尚未检测的样本P，若P未被处理，则检查其Eps邻域，如果其中包含的样本数≥minPts，则构成新簇C，将其邻域中所有其他样本加入C；②对C中没有被处理的样本P，检查Eps邻域，若其中样本数≥minPts，则将尚未归入任何一个簇的样本加入C；③重复步骤②，直到没有新对象加入簇C；④重复步骤①~③，直到处理完所有样本。

2.3 基于资源关系的信息资源层次空间构建

层次空间是概念空间的一种表现形式，而概念空间是指概念间相关关系的集合（如同义关系、近义关系、上下位关系等），层次空间主要用于体现概念间上下关系及近义关系。构建资源的层次空间，有利于用户清楚地发现资源间等级层次关系。文档资源间的四层结构层次空间构建步骤如下[9]。

（1）选择与所有资源相关度最大的资源作为根节点。

（2）选择根节点为起点即第一层级；设定某一阈值A，将与根节点间相关性大于阈值的资源按相关性大小依次作为根节点的子节点加入层次空间建立第二层级；随后按同一方法设定阈值B，建立第三层级，设定阈值C建立第四层级，层次空间建立完毕。

（3）从剩余的未加入层次空间的资源中选择与剩余资源平均相关度最大的资源作为新的根节点。

（4）重复步骤（2）、步骤（3），直到所有资源加入层次空间中。

3 实验数据预处理

本文的实验数据来源于武汉市国土资源和规划信息中心，从中搜集了100条文档型业务公文数据，对数据进行编号（0~99），数据名称见图3。

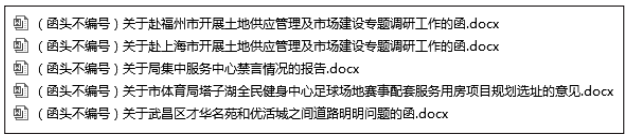

图3 数据名称

在进行资源建设之前须对文档数据进行预处理，首先利用中国科学院NLPIR汉语分词系统来对文档进行分词，之后利用哈尔滨工业大学停用词表过滤掉文档中如“啊”“咦”“哦”等无实际意义的词以及“《”“∈”等特殊符号。考虑到所使用的档案文档资源的特殊性，为使其能表示出较为准确的主题信息，剔除文档中出现频次在100次以上且对主题区分意义较小的词，如“工作”“项目”“建设”“位于”等。对于词典中没有的词，通过NLPIR的自定义词典功能加入词典。

4 信息资源建设实例

4.1 基于主题的信息可视化展示

在数据预处理的基础上，利用LDA主体模型对文档集合进行建模，得出文档数据-主题概率分布。利用Python中的LDA工具包对图3中的文档数据进行模型训练，选择主题数K分别为3、4、5、6、7进行研究与比较，对比不同主题数下各主题中文档间相关性，发现选取的文档经预处理后有较大可能分布为4个主题，因此选择主题数K=4，LDA参数结合国内外相关文献知识取值$\alpha=\frac{50}{k}$，$\beta=0.01$[10]。经过500词迭代训练后，得出文档-主题概率分布，即文档分属于每个主题的概率，表1显示了前20个文档数据的文档-主题概率分布。

对所得的文档-主题概率矩阵进行分析，发现对于部分文档如文档6、文档11、文档12等来说，其分属于某个主题的概率明显高于其分属于其他主题的概率，说明其与该主题高度匹配。而对文档2、文档3、文档4等来说，其分属于每个主题的概率大小相近，难以直接将其分入某一个主题。因此，须设置相应的阈值来进行主题分布控制，在对文档-主题概率进行分析的基础上，选取阈值为0.45，设置文档-主题概率大于该阈值的文档可归入相应主题。例如，对文档6来说，文档6-主题2概率为0.703，因此将其划分入主题2；对文档2来说，其分别与主题0、1、2、3的相关概率都低于0.45，因此不将其归入任何主题。在进行相关文档展示时，为更直观地观测文档的内容，使用TF-IDF方法来对文档进行处理，对于每个文档选择其TF-IDF值最高的5个词来替代文档名。对100个文档分别进行主题划分，其结果如图4~图7所示。

表1 文档-主题概率矩阵

主题 / 文档	0	1	2	3
0	0.206	0.245	0.304	0.248
1	0.199	0.241	0.326	0.233
2	0.200	0.277	0.270	0.253
3	0.263	0.224	0.230	0.282
4	0.267	0.264	0.229	0.241
5	0.216	0.278	0.242	0.264
6	0.071	0.703	0.124	0.101
7	0.247	0.246	0.253	0.253
8	0.177	0.219	0.331	0.273
9	0.299	0.228	0.259	0.217
10	0.260	0.229	0.237	0.275
11	0.520	0.386	0.043	0.051
12	0.395	0.474	0.058	0.074
13	0.286	0.662	0.026	0.026
14	0.472	0.406	0.041	0.081
15	0.450	0.314	0.123	0.113
16	0.369	0.531	0.044	0.055
17	0.615	0.244	0.073	0.068
18	0.727	0.189	0.039	0.044
19	0.420	0.206	0.143	0.230

从主题-文档划分结果可以看出，100个文档中仅有23个文档被划分到相应主题，而在剩下的77个文档中，有部分文档与其他文档间存在较强的相关性，仅依靠主题提取技术难以发现其间的相关性。因此，须进一步的进行文档间聚类分析，以期对文档间关系进行更深入的挖掘。

4.2 基于聚类分析的信息可视化展示

在运用DBSCAN算法进行聚类分析时需确定Eps

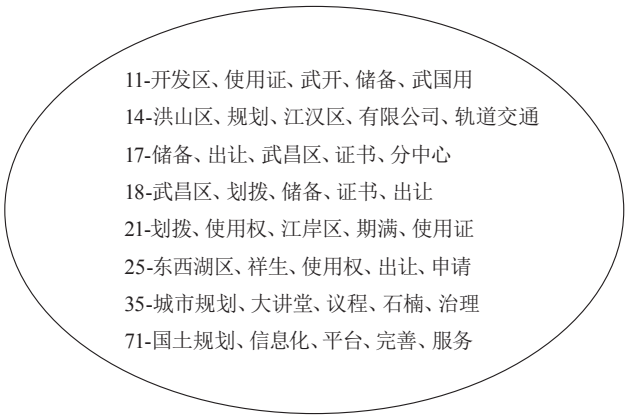

图4 主题0相关文档

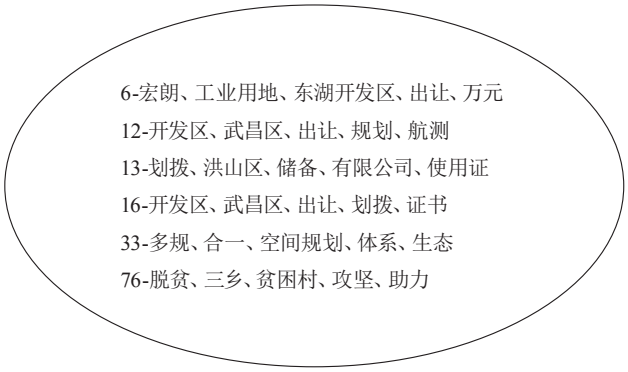

图5 主题1相关文档

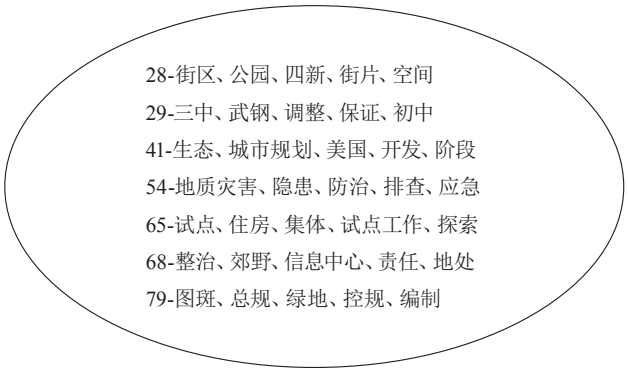

图6 主题2相关文档

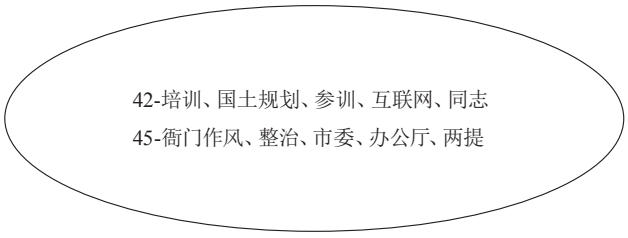

图7 主题3相关文档

与minPts值，Eps的取值可通过聚类对象间的欧式距离值来设置，欧式距离计算见公式（2）。

$$d_{ab}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^{2}} \quad (2)$$

其中，d_{ab}为对象a与对象b的欧式距离，X_{1k}、X_{2k}分别为对象a、b在第k维的向量值。将表1中的4个主题视为4个维度，各文档在维度的向量值等同于其文档-主题概率。如对文档0来说，其向量表示为（0.206、0.245、0.304、0.248）。利用公式（2）计算表1中文档的欧式距离，结果见表2。

表2 部分文档间欧氏距离矩阵

	1	2	3	4	5	6	7	8
0	0.000	0.028	0.047	0.102	0.099	0.073	0.531	0.066
1	0.028	0.000	0.070	0.126	0.121	0.098	0.537	0.090
2	0.047	0.070	0.000	0.096	0.081	0.034	0.492	0.059
3	0.102	0.126	0.096	0.000	0.057	0.075	0.557	0.046
4	0.099	0.121	0.081	0.057	0.000	0.059	0.512	0.038
5	0.073	0.098	0.034	0.075	0.059	0.000	0.492	0.047
6	0.531	0.537	0.492	0.557	0.512	0.492	0.000	0.529
7	0.066	0.090	0.059	0.046	0.038	0.047	0.529	0.000
8	0.054	0.051	0.090	0.133	0.147	0.114	0.564	0.110
9	0.109	0.122	0.117	0.080	0.062	0.109	0.556	0.066
10	0.092	0.116	0.086	0.011	0.050	0.067	0.551	0.035
11	0.475	0.487	0.455	0.425	0.387	0.435	0.558	0.423
12	0.423	0.436	0.392	0.391	0.343	0.375	0.403	0.379
13	0.554	0.564	0.516	0.547	0.496	0.506	0.251	0.527
14	0.440	0.454	0.416	0.391	0.351	0.395	0.506	0.388
15	0.340	0.352	0.324	0.288	0.252	0.305	0.543	0.287
16	0.462	0.474	0.428	0.438	0.388	0.413	0.356	0.423
17	0.503	0.514	0.496	0.441	0.419	0.477	0.714	0.450

对表2中的欧氏距离进行比较分析，发现当欧式距离为0.094时，可保证大部分对象被划分到对应簇，且各簇内的对象间相关性较大。因此，将Eps值设置为0.094，DBSCAN中另一参数minPts的取值通常为2。选定Eps和minPts值后，利用Python的DBSCAN工具包来实现对象的聚类。文档在主题1-主题2维度上的聚类结果如图8所示，经过整合的最终聚类结果见图9。

如图8所示，各文档按其在主题1及主题2上的向量值分布于图中，其中十字型符号表示一个聚类簇，颜色深浅不同则表示属于不同聚类簇，图中文档间的相关性可通过距离来判断，即距离越近的文档间相关

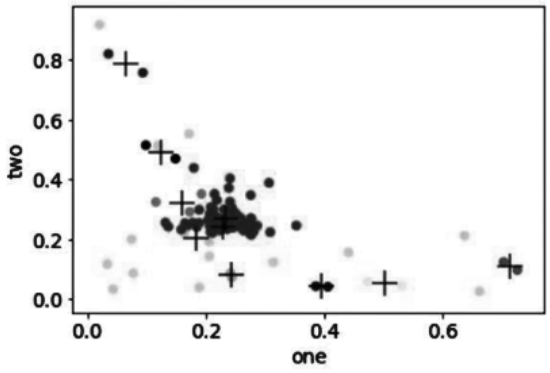

图8 文档在主题1-主题2维度上的聚类结果

性越大。

文档在不同维度上的分布存在差异，共有1-2、1-3、1-4、2-3、2-4、3-4六个维度。将文档在各维度上的分布结果进行整合，得到图9所示的最终聚类结果。

图9显示的是文档最终聚类结果，可看出，有13个文档被划分为噪声，其他的87个文档被分别分入到9个聚类簇中（第零簇至第八簇），其中第零簇中的文档数量最大，该类文档在采集的文档资源中所占数量最多，可判断其为武汉市国土资源和规划信息中心日常工作

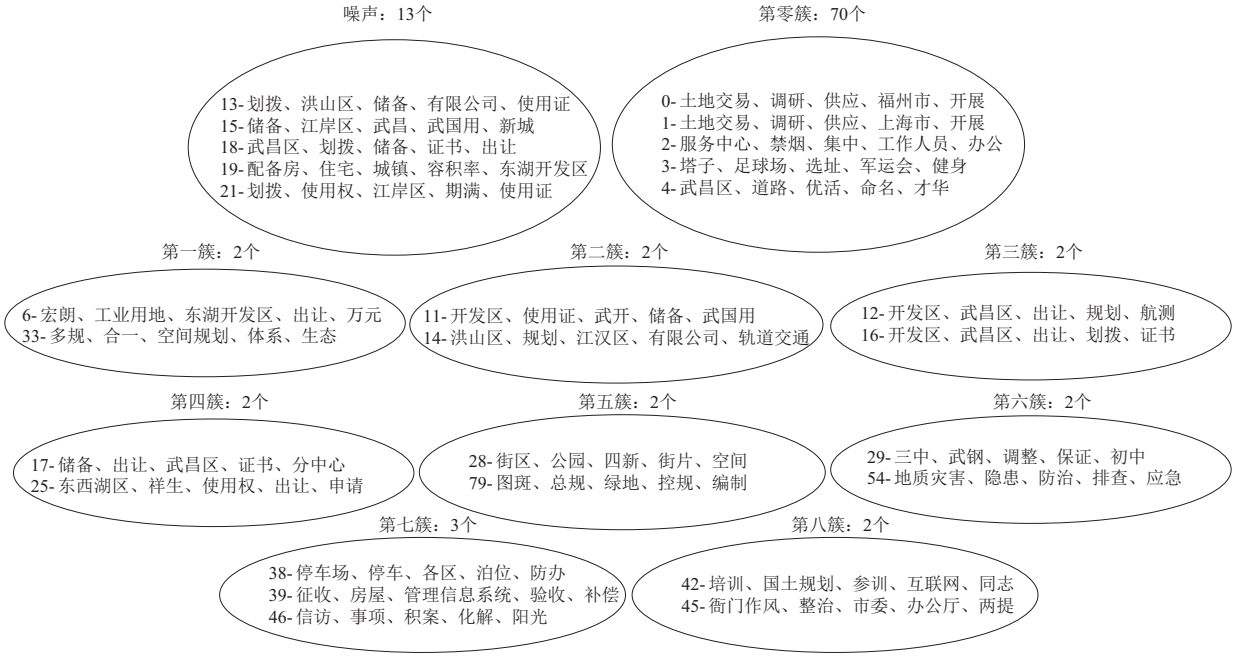

图9 文档最终聚类结果

中面对类型最多的文档；而其他8个类中的文档数量则相对较少，可判断武汉市国土资源和规划信息中心在日常工作中面对的这些类型文档较少。

比较主题划分结果与聚类结果可以发现，两种分析结果存在一定的相似性。如对文档42与文档45来说，其在主题划分中被划分为同一主题（主题3），在聚类时也被划分到同一簇（第八簇）。但是，两种分析结果也存在一定的差异，如文档38、39、46在聚类时被划分到同一簇（第七簇），说明其间存在一定的相关性，而在主题分析中文档38、38、46未被划分到任意主题，3个文档间未显示出相关性；对文档6、12、16、33来说，其在主题划分时被划分到同一主题（主题1），而在聚类时，文档6、33被分入第一簇，文档12、16被分入第三簇，这说明在同一主题中，文档也可能属于不同类别。聚类可以对主题划分的结果进行补充与改进，可以发现更全面的文档间关系。

4.3 基于等级结构的信息可视化展示

在主题提取与聚类的基础上，为得到更全面的文档间关系，可将主题提取与聚类的结果进行综合，并在此基础上尝试构建层次空间，以发现文档间更深入的相关关系。基于聚类到一簇的文档间具有较强的相似性这一概念，利用聚类结果来对主题划分结果进行改进。分析现有主题（主题0、1、2、3）中文档的聚类结果，为聚类时聚到一簇的文档建立更为深入的相关关系。在聚类过程中，文档7与文档25、文档11与文档14等聚集到一簇，为这些文档建立更为深入的相关关系以

此来改进主题划分的结果，对主题0、1、2进行了改进，改进后的主题0、1、2如图10~图12所示。

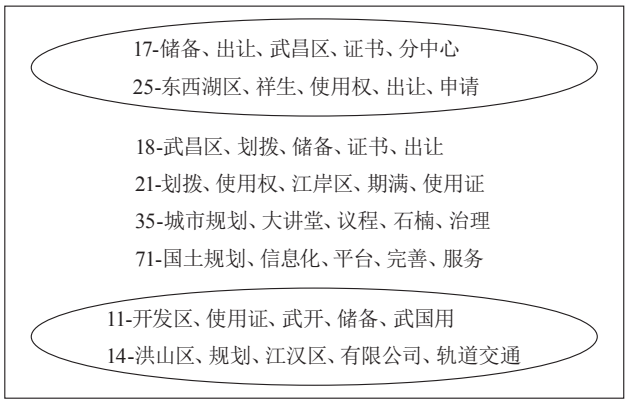

图10 改进后主题0相关文档

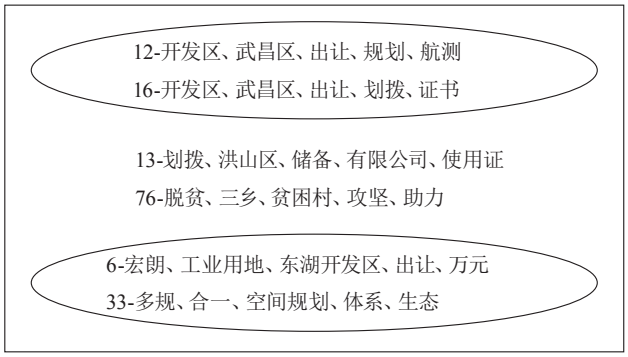

图11 改进后主题1相关文档

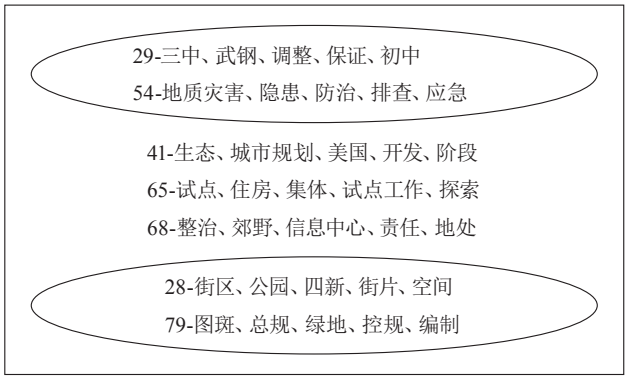

图12 改进后主题2相关文档

在利用聚类结果对主题划分结果进行改进后，同一主题中的文档可以被划分到不同簇，主题中文档间相关性被近一步定义，即聚到一簇的文档间其相关性相较于其他文档间更强。

对主题划分结果改进后，考虑到主题中文档个数及文档所属簇个数，最后选择主题0及主题2中的文档进行文档间层次结构构建。利用余弦相似度算法（式3）来计算主题中文档间相似度，结果见表3和表4。

$$Sim(T_i, T_j) = \cos(\theta_{ij}) = \frac{\sum_{k=1}^{N} T_{ik} T_{jk}}{\sqrt{\sum_{k=1}^{N} T_{ik}^2 \sum_{k=1}^{N} T_{jk}^2}} \quad (3)$$

在余弦相似度算法中，T_i、T_j分别为文档i的向量和文档j的向量。

表3 主题0中文档相似度

文档	11	14	17	18	21	25	35	71
11	1.000	0.996	0.966	0.927	0.831	0.966	0.861	0.800
14	0.996	1.000	0.945	0.897	0.787	0.953	0.823	0.776
17	0.966	0.945	1.000	0.990	0.942	0.995	0.963	0.914
18	0.927	0.897	0.990	1.000	0.979	0.975	0.986	0.924
21	0.831	0.787	0.942	0.979	1.000	0.919	0.993	0.931
25	0.966	0.953	0.995	0.975	0.919	1.000	0.950	0.927
35	0.861	0.823	0.963	0.986	0.993	0.950	1.000	0.959
71	0.800	0.776	0.914	0.924	0.931	0.927	0.959	1.000
均值	0.918	0.897	0.964	0.960	0.923	0.961	0.942	0.904

表4 主题2中文档相似度

文档	28	29	41	54	65	68	79
28	1.000	0.931	0.924	0.892	0.991	0.955	0.996
29	0.931	1.000	0.942	0.988	0.889	0.982	0.912
41	0.924	0.942	1.000	0.895	0.876	0.922	0.913
54	0.892	0.988	0.895	1.000	0.840	0.981	0.861
65	0.991	0.889	0.876	0.840	1.000	0.915	0.997
68	0.955	0.982	0.922	0.981	0.915	1.000	0.929
79	0.996	0.912	0.913	0.861	0.997	0.929	1.000
均值	0.956	0.949	0.925	0.922	0.930	0.955	0.944

得到文档间相似度之后，为主题0及主题2中的文档建立层次空间。按2.3节的方法，首先选择与其他文档平均相关度最大的文档作为等级结构的根节点。对主题0来说，选择文档17为根节点，即第一层级；随后设置第二层级阈值为0.99，将与文档17相似度大于阈值的其他文档按相似度大小依次加入第二层级，即依次加入文档25与文档18；设置第三层级阈值为0.96，将与文档25、文档18相似度大于阈值的其他文档按相似度大小依次加入第三层级，即依次在文档25下加入文档11，在文档18下加入文档35与文档21；设置第四层级阈值为0.95，将与文档11、文档35、文档21相似度大于阈值的其他文档按相似度大小依次加入第四层级，即依次在文档11下加入文档14，在文档35下加入文档71。主题0文档层次空间见图13。

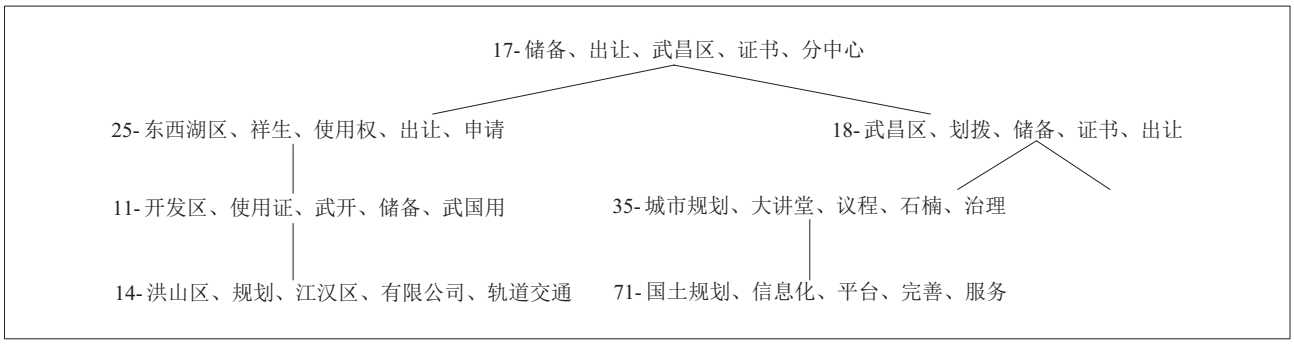

图13 主题0文档层次空间

为主题2中的文档建立层次空间。对主题2来说，选择文档28为根节点，即第一层级；随后设置第二层级阈值为0.99，将与文档12相似度大于阈值的其他文档按相似度大小依次加入第二层级，即依次加入文档79与文档65；设置第三层级阈值为0.91，将与文档79、文档65相似度大于阈值的其他文档按相似度大小依次加入第三层级，即依次在文档79下加入文档68、文档29与文档41；设置第四层级阈值为0.90，将与文档68、文档29、文档41相似度大于阈值的其他文档按相似度大小依次加入第四层级，即在文档29下加入文档54。主题2文档层次空间见图14。

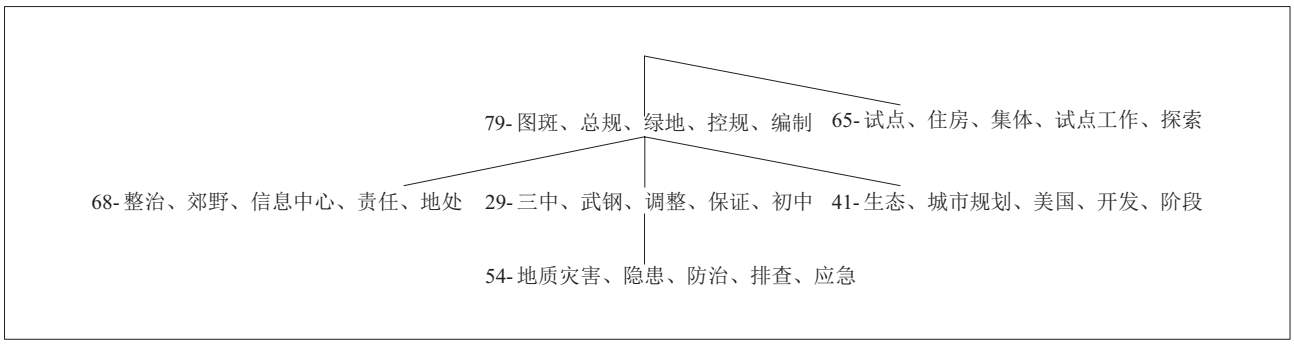

图14 主题2文档层次空间

在层次空间中，根节点与子节点可能会存在一定的连续性关系。如对图13中的文档17、文档18与文档21来说，文档18与文档21可能是在文档17基础上产生的新文档。层次空间中，距离越近的节点其关系越为紧密，如对文档14来说，其与文档11距离最近，即在层次空间中其关系最为紧密；而文档14与文档71距离最远，即在层次空间中其关系最为疏远。

5 结语

将数据挖掘技术应用于信息资源建设，可以更深入地挖掘信息资源间的相互关系，更全面地展示信息资源间的联系。通过数据挖掘技术来建设信息资源，可以建立挖掘更深入、展示更全面、使用更便利的信息资源。

本文以档案信息资源为例，综合应用主题提取、聚类、层次空间构建3种技术方法来进行信息资源建设，其具体意义如下。

第一，在分析档案信息资源特征的基础上对档案信息资源进行处理，利用LDA主题模型对档案信息资源进行主题提取，并对提取的主题进行了主题展示。从主题划分的结果来看，近50%的档案信息资源存在较为明确的主题偏向，是围绕某些主题而进行的工作；而其他档案资源主题偏向不太明确，涉及多个主题。

通过将主题提取技术应用到档案信息资源建设，使得档案信息中蕴含的主题信息得以被挖掘；通过主题来进行档案信息资源的展示，可以从特定主题的角度来进行档案资源的浏览与查找，并能同时发现与某一主题相关的所有档案信息资源，大幅提高了档案信息资源查找与利用的效率。

第二，在主题分析的基础上，用DBSCAN聚类算

法对档案信息资源进行聚类分析，将每一个主题视为一个维度，分析了档案信息资源在多维度上的相关性，使得在多个维度上都具有较强联系的资源聚集到一类。从聚类结果来看，有大多数的档案信息资源被聚集到同一聚类簇，而其他的档案信息资源则分布在其他的几个小聚类簇中。

通过将聚类技术应用到档案信息资源建设中，使得档案信息资源间的关系能够被更清晰地展示，从多个维度对档案信息资源间的相关性分析，保证了聚类结果的准确性。以本文档案资源聚类结果来说，大多数资源都被聚集到同一类簇，而这类档案信息资源也恰好是资源来源规划中心在实际工作中处理最多的一类档案信息资源。通过聚类，可以帮助相似资源的查找，实现对档案信息资源的合理调节与配置，并能展示出档案信息资源间的相互关系。

第三，在主题提取与聚类分析的基础上，进一步进行了档案信息资源间层次空间的构建。选取适宜的、能进行层次构建的信息资源集合，利用余弦相似度算法进行了档案信息资源间的相似度计算，通过相似度分析，构建档案信息资源间的层次关系。

通过为档案信息资源构建层次空间，可以更好地发现档案信息资源间的关系强度，并且可以在一定程度上发现档案信息资源间的上下位关系及近义关系。即在层次空间中，根节点与子节点间通常存在一定的上下位关系，同一根节点的不同子节点间通常存在近义关系。而两个节点在层次空间中的距离可以表示节点间的相关性大小，即距离越近的节点间联系越紧密，距离越远的节点间联系越疏远。层次空间的构建，使得档案信息资源间的关系得到了更充分、全面的展示，为档案信息资源的利用带来了极大的便利。

参考文献

[1] 冯湘君. 国家综合档案馆数字档案资源建设理念探析 [J]. 档案学通讯，2017（3）：56-60.

[2] 国家档案局. 全国档案事业发展“十三五”规划纲要 [EB/OL]. [2018-06-04]. http://www.saac.gov.cn/news/2018-12/04/content_136280.htm.

[3] 赵彦昌，毛丽敏. “互联网+”环境下档案信息资源建设若干问题研究 [J]. 档案学研究，2017（4）：31-35.

[4] KRESTEL R，FANKHAUSET P，NEJDL W. Latent dirichlet allocation for tag recommendation [C]//ACM Conference on Recommender Systems，Recsys 2009，New York，Ny，Usa，October. DBLP，2009：61-68.

[5] 邓丹君，姚莉. 基于微博标签和LDA的微博主题提取算法 [J]. 计算机与数字工程，2017，45（5）：954-957.

[6] 李慧宗，胡学钢，杨恒宇，等. 基于LDA的社会化标签综合聚类方法 [J]. 情报学报，2015，34（2）：146-155.

[7] 石陆魁，何丕廉. 一种基于密度的高效聚类算法 [J]. 计算机应用，2005，25（8）：1824-1826.

[8] 李双庆，慕升弟. 一种改进的DBSCAN算法及其应用 [J]. 计算机工程与应用，2014，50（8）：72-76.

[9] 熊回香，叶佳鑫. 基于同义词词林的社会化标签等级结构构建研究 [J]. 情报杂志，2018（1）：126-131.

[10] 王茜，王均波. 一种改进的协同过滤推荐算法 [J]. 计算机科学，2010，37（6）：226-228.

作者简介

丁玲，女，1974年生，硕士，高级工程师，研究方向：数字档案馆。
叶佳鑫，男，1993年生，硕士研究生，研究方向：网络信息组织、数据挖掘，通信作者，E-mail：jxye@mails.ccnu.edu.cn。
曾婷，女，1988年生，硕士，馆员，研究方向：数字档案馆。

Research on Information Resources Mining and Visualization Based on LDA

DING Ling[1] YE JiaXin[2] ZENG Ting[1]
(1. Wuhan Land Resource and Urban Planning Information Centre, Wuhan 430014, China;
2. School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: LDA is a probabilistic model for extracting topics from document resources, and it is usually effective for extracting topics from documents. Archival information resources are of high utilization value, but there are problems such as fragmentation and insufficient construction. Based on this, this paper applies LDA, clustering and hierarchical space construction technology to the construction of archival information resources, and conducts an empirical study. From the experimental results, the application of LDA to the construction of archival information resources can excavate the hidden connections between resources, clarify the hierarchy of resources, and contribute to the visual display of information resources.

Keywords: Information Resource Construction; Topic Distillation; Clustering; Levels of Space

（收稿日期：2019-01-10）