文 献 与 数 据 学报
第 4 卷　第 3 期　　2022 年 9 月

基于异构信息网络的科技文献主题识别研究

席崇俊　徐珍珍　刘文斌　丁　楷
（中国科学技术信息研究所，北京　100038）

摘　要：[**目的 / 意义**] 科技文献主题识别研究对于把握科技领域的研究重点和热点，揭示领域内的发展态势和演化趋势具有重要意义。传统科技文献主题识别研究多基于科技文献同构信息网络进行研究，难以表达科技文献系统中丰富的对象类型和复杂的语义关系。[**方法 / 过程**] 本文利用异构信息网络表达科技文献系统中各类型对象之间丰富的语义关系，并将其转化为高阶张量的形式，利用非负张量分解算法对其进行主题挖掘。[**结果 / 结论**] 实验结果表明：基于异构信息网络可以对科技文献进行更深层次的主题识别，非负张量分解算法在处理异构信息网络时方便快捷，可以减少语义信息的丢失。

关键词：主题识别　异构信息网络　高阶张量　非负张量分解　人工智能
分类号：G350
DOI：10.31193/SSAP.J.ISSN.2096-6695.2022.03.06

0　引　言

随着科学技术的不断发展，科技文献数量也迅猛增长。科技文献是科学技术知识和信息的载体，也是科研人员在科学实践中总结出来的研究成果，从大量科技文献中识别出主题结构及其演化路径，有利于把握领域内的研究重点和热点，揭示科技领域的发展态势和演化趋势，以更好地应对科技发展过程中面临的挑战[1-2]。

目前，科技文献主题识别研究主要有基于全文文本、主题、引文、关键词等粒度的分析，研究方法包括引文分析法、词频分析法、共词分析法、文本挖掘法等。其中，引文分析法是基于文献之间的引用关系来揭示学科领域的主题与演化路径[3]，如 Shibata 等采用图拓扑结构聚类分析直接引文网络生成学科主题[4-5]；Yookyung 等从词汇粒度层次结合文献间的直接引用关系以主题词簇的形式表达学科主题[6]；Ma 等利用直接引用网络分析法探究经济和法律领域的新兴

[作者简介] 席崇俊 (ORCID：0000-0003-0334-7109)，男，硕士研究生，研究方向为技术竞争情报，Email：xicj7465@163.com；徐珍珍，女，硕士研究生，研究方向为科技大数据，Email：xuzz2019@istic.ac.cn；刘文斌，男，硕士研究生，研究方向为自然语言处理，Email：liuwb2019@istic.ac.cn；丁楷，男，硕士研究生，研究方向为知识图谱构建，Email：Dingk2019@istic.ac.cn。

问题[7]；梁镇涛等以眼动追踪领域为例，对文献的引文关系进行了提取与学科标注，构建出文献和学科层面的引文关系网络[3]；开滨等基于引文分析和可视化技术，构造了 80 个学科互相引用图，并对中国高校的科研优势学科进行了探测[8]；郭倩影等基于引文网络结合文献节点的时间、内容等性质，构建了包含引用强度、引文网络、引用时长和引用内容四个维度的学术传承性文献识别框架[9]。

词频分析法是通过统计关键词在文献中出现的频次高低来分析领域的研究热点与研究动向[10]，如 Walls 等通过构建概率模型统计生成词频分布，根据词频分布规律识别主题词簇[11]；Kleinberg 通过检测突发词识别研究热点[12]；储节旺等通过统计文献中关键词出现的频次高低，分析了知识管理近十年来的研究热点、应用领域[10]；奉国和等构建时间 – 关键词频次矩阵，并赋予词频按时间递减的权重，构建了时间加权关键词词频分析模型，以此揭示学科研究热点及变化趋势[13]；余丰民等基于词频统计方法构建了研究热点漂移程度计算模型[14]。

共词分析法则是利用词与词之间的关联关系得到文献主题之间的关系[9]，如 Kim 等通过关键词共现构建专利语义网络，采用 k-means 聚类分析方法生成专利地图来进行主题可视化分析及新兴主题的探测[15]；Neff 等通过共词网络聚类生成生态学领域学科主题来识别新兴概念[16]；唐果媛等采用人工判读法、文献计量法和对比分析法，从定性和定量两个角度对共词分析法在国际上和中国国内的研究现状进行分析[17]；钟伟金等对共词聚类分析的原理及特点进行了分析[18]；叶春蕾等基于 LDA 模型改进了共词分析方法[19]。

文本挖掘法则是通过文本挖掘技术对主题进行抽取，并用相关评价标准对主题进行分类[2]，如 Yoon 等采用 SAO 结构语义挖掘方法构建动态专利地图，以此为专家在制定研发战略的过程中提供相关技术竞争趋势信息[20]；田鹏伟等利用文本挖掘技术提取专利技术主题构建共现网络，并采用 OVL 算法及加权运算对异构信息网络进行融合，基于融合后的网络开展主题识别[21]；何伟林等基于改进的主题模型对国内情报学领域的研究主题结构及演化过程进行了分析[2]。

上述研究主要是基于科技文献某一个特征项进行主题挖掘，如引文分析法主要基于引文特征项，词频分析法、共词分析法主要基于词特征项，文本挖掘法主要也是基于文本中的词特征项进行主题挖掘。经典的主题识别研究方法主要基于科技文献之间的同构信息网络进行主题挖掘，同构信息网络中只包含了一种类型的对象，且对象之间只存在一种类型的边，最常见的有关键词共现网络、引文分析网络等[23]。但是，一篇科技文献中包含了标题、关键词、作者、机构、参考文献等众多字段，各种字段之间组成一个庞大而复杂的信息网络[22]，在实际情况中绝大部分的信息网络都是异构的，即信息网络中包含了多种类型的对象，且不同类型的对象之间存在着丰富的语义关系，这种网络称之为异构信息网络[22]。例如，在科技文献信息网络中，对象有作者、论文、刊物等，除了存在作者与作者之间的合作关系，还存在着作者与论文之间的写作关系、论文与论文之间的引用关系、论文与刊物之间的发表关系等。如果只选择科技文献信息网络中的一种对象进行主题挖掘，势必会丢失大量语义信息，主题识别结果也不精确，因此，利用同构信息网络表达科技文献系统存在局限性，而利用异构信息网络则可以完整地展示科技文献系统中各种类型对象之间的关系，获得更丰富和复杂的语义信息。

此外，经典的数据挖掘方法在对异构信息网络进行挖掘时，都倾向于将异构信息网络转化为

同构信息网络进行处理，如果将所有对象都视为同一类型来处理，即不再区分作者、关键词、标题等原本的含义，这样做不仅会丢失部分语义信息，也会损坏信息网络的内部结构，分析结果也毫无意义[21]。异构信息网络中存在着多种类型的对象，传统基于距离或相似度函数的聚类算法（如系统聚类法、K–means 聚类法）也不再适用，在异构信息网络中度量两个不同类型对象之间的距离是没有意义的，且传统聚类算法每次只能对异构信息网络中的一种类型对象进行聚类，而基于矩阵分解的聚类算法（如奇异值分解、非负矩阵分解）最多只能同时处理两种类型的对象。因此，如何将科技文献异构信息网络转化为数学形式并选择合适的聚类算法对其进行主题挖掘是本文所关注的问题。

本文尝试利用高阶张量来组织和描述科技文献系统之间的异构信息网络，通过将文献单元（关键词、标题、作者）表示成高阶张量的形式，既可以体现出文献单元之间的知识结构与内在关联性，又能最大限度地保留科技文献信息网络内部复杂的结构和语义关系，基于非负张量分解的聚类算法可以在不破坏数据的潜在含义与内在信息的前提下，将多种类型的文献单元根据其复杂的语义关系一次性同时划分到不同的类团中，从而可以较为快捷、深入地识别主题。

1　研究方法

1.1　科技文献异构信息网络的张量表示

科技文献系统中包含了众多的文献单元，各文献单元之间存在着丰富而复杂的语义关系，这种语义关系可抽象用直接或间接的矩阵表示，但矩阵是一个二维数组形式，最多只能同时表示两种文献单元之间的语义关系，这就需要一种新的数据形式来表示这种复杂的关系网络。张量是一个多维数组，一阶张量是向量，二阶张量是矩阵，三阶以上称为高阶张量，将科技文献异构信息网络表示成高阶张量形式更能反映文献单元之间的内在知识结构。

为了便于对数据的处理和结果的解读，本文以三阶张量为例进行研究方法的说明与实验结果的分析。由于作者是科技知识的创造者，文献标题是科技文献内容的直接反映，关键词是深入文献内部的单元，因此本文选取关键词、标题、作者三种文献单元进行张量的构建，其中张量的每一维度分别代表一种文献单元类型。如图 1 所示，以圆形、正方形、三角形分别代表关键词、标题、作者三种类型的文献单元，图形中的数字分别代表该类型下的文献单元序号，若某三个文献单元形成一个闭环（即它们三者共同出现），则在以它们的序号组成的三维坐标处赋值为 1，否则赋值为 0。

1.2　非负张量分解算法

在张量分解中，最常用的分解方法有 CP 分解和 Tucker 分解[24]。CP 分解是将一个 n 阶张量分解成多个秩一的张量的和的形式[25]。Tucker 分解则是将其分解成一个核心张量与若干个因子矩阵乘积的形式，其中核心张量可以看成原张量的浓缩形式[26]，当核心张量是一个对角方量时，Tucker 分解则退化成了 CP 分解（见图 2）。

但是，经典张量分解算法的分解结果中元素有正有负，而负元素在实际问题研究中是没有意义的，因此有学者提出了非负张量分解算法。非负张量分解算法是非负矩阵分解算法在高维空间

中的拓展，非负矩阵分解算法可以有效解决经典聚类算法中一个对象只属于一个类团的不足以及对象在类团中的权重值为负的缺陷，但最多只能处理两种类型的对象，而非负张量分解算法既保留了张量的优点又避免了负元素的出现，被广泛应用于图像处理、音频分类文本挖掘等领域[27–29]。

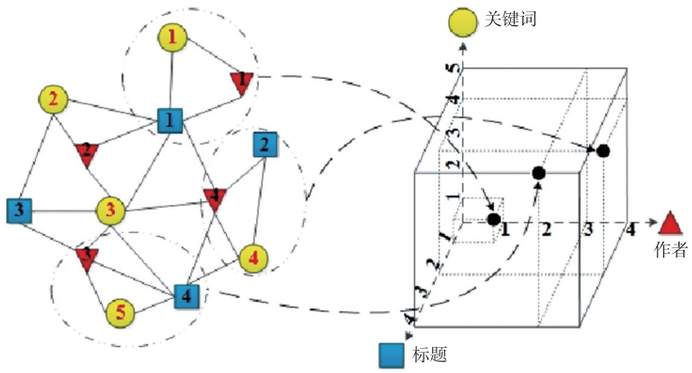

图 1　科技文献异构网络的张量建模示意图[22]

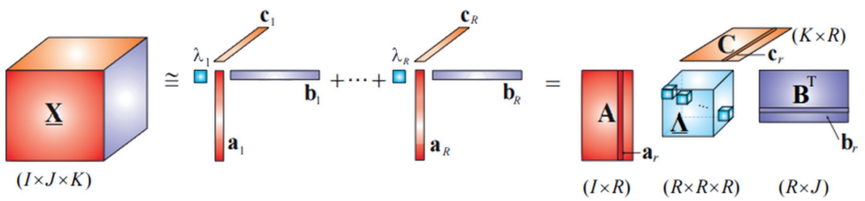

图 2　三阶张量的 CP 分解与 Tucker 分解

由于本文是基于科技文献异构信息网络进行主题识别，各类型对象聚成的类团数一致，因而基于非负 Tucker 分解后的核心张量为一个方量，阶数即为类团数，且与非负 CP 分解的结果一致。本文以关键词 – 标题 – 作者三阶张量为例进行非负 CP 分解，分解得到的秩一张量的个数即为主题个数，每个维度下的分解结果即为该维度所代表的类型对象在各主题中的因子值（即权重值），因而基于非负张量分解算法只需进行一次聚类，便可得到各主题下包含的关键词、标题、作者等信息，其中关键词和标题字段可以进行主题命名，在得到主题研究内容的同时也可以看出该主题下代表性的文献及作者团体。

2　研究路线

本文的研究路线如图 3 所示，为了对比基于异构信息网络的主题识别结果相较于基于同构信息网络的主题识别结果的优势，以及基于张量表示的异构信息网络相较于基于矩阵表示的异构信息网络的优势，分别选取关键词、标题、作者字段构建了关键词共现网络（同构信息网络）、关键词 – 标题网络（基于矩阵表示的异构信息网络）、关键词 – 标题 – 作者网络（基于张量表示的

异构信息网络）进行两组对照实验。其中，关键词共现网络和关键词 – 标题网络可以用矩阵形式表达，因此考虑用非负矩阵分解算法进行主题识别，而关键词 – 标题 – 作者网络是用张量形式表达，考虑用非负张量分解算法进行主题识别。最后，分别从主题识别结果的精度（关键词个数、关键词权重、关键词内容）和广度（主题特征项）对三种信息网络下的主题识别结果进行评价。

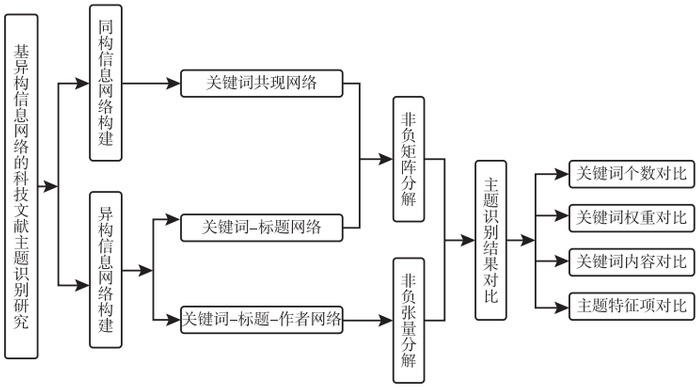

图 3　研究路线图

3　实证分析

3.1　数据集构建

人工智能是社会发展和技术创新的产物，是促进人类进步的重要技术形态，当下关于人工智能的研究十分火热，本文在 Web of Science 数据库中以“Artificial Intelligence”为主题词检索 2017~2021 年期间发表的高被引论文，共检索到 581 篇文献，包含 1827 个关键词字段、510 个作者字段、385 个机构字段等，对上述文献导出题录数据后去除本位词“Artificial Intelligence”，并利用 TDA 软件进行数据清洗。

3.2　结果分析

鉴于非负矩阵分解算法在处理同构信息网络及只包含两种类型对象的异构信息网络时具有良好的效果[21]，本文以非负矩阵分解算法的主题识别结果为对照实验组，探讨基于异构信息网络与同构信息网络在科技文献主题识别时的差异，以及利用非负张量分解算法和非负矩阵分解算法在处理异构信息网络时的优劣。关键词是三种网络中共同出现的字段，也是深入文献内部的单元，因此本文对实验结果进行横向和纵向对比分析。横向对比，即三种网络下的聚类结果中各类团中的关键词信息，以反映聚类结果的精度。纵向对比，即三种网络下的聚类结果中各类团的对象类型信息，以反映聚类结果的广度。

3.2.1　聚类结果精度分析（横向对比）

通过不断实验比较发现，当聚类数目大于 6 类时会出现部分类团内容重叠的现象，因此本文

将聚类数目初步定为 6 类。基于非负矩阵分解算法处理的关键词共现网络（同构信息网络）、关键词 – 标题网络（异构信息网络），以及基于非负张量分解算法处理的关键词 – 标题 – 作者网络（异构信息网络）的聚类结果中各类团所含关键词个数以及关键词因子值的标准差，分别如图 4 和图 5 所示。

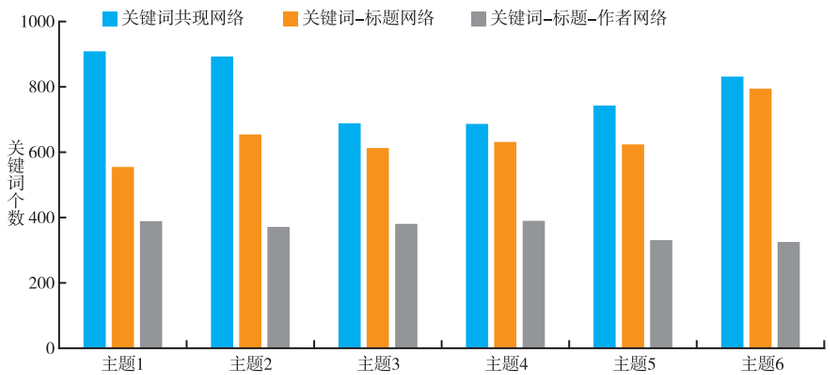

图 4　三种信息网络下各主题所含关键词个数

由图 4 可以看出，随着科技文献信息网络复杂度的提升，各主题下所含的关键词个数在不断减少。由于关键词共现网络中只包含一种类型的信息对象，且各信息对象之间只存在一种关系（共现关系），因而聚类结果中每个主题下所含的关键词个数较多，主题的综合度也较大；而在关键词 – 标题网络和关键词 – 标题 – 作者网络中，随着信息对象类型和关系的增多，在聚类时每种语义关系都可能将关键词划分到不同的类团中，因而每个主题下所含关键词个数变少，主题的深度也在提升，初步说明了基于科技文献异构信息网络可以对主题进行深层次的挖掘。

通过图 5 可以看出，基于非负矩阵分解算法在处理关键词共现网络和关键词 – 标题网络时，各主题下所含关键词的因子值标准差较大，这是由于高频关键词的因子值与低频关键词的因子值差距悬殊导致的，而基于非负张量分解算法处理关键词 – 标题 – 作者网络的聚类结果中关键词的因子值标准差较小，这是由于丰富的语义关系削弱了高频关键词对主题结果的影响，说明利用非负张量分解算法在对科技文献异构信息网络进行主题识别时可以减少语义信息的丢失，降低对内部知识结构的损坏。

此外，通过绘制三种网络下主题识别结果各类团所含关键词的词云图可以看出（如图 6~图 8 所示），三种聚类结果下人工智能领域近五年的高被引论文研究热点主要集中在机器学习（Machine learning）、深度学习（Deep learning）、物联网（Internet of Things）、新冠肺炎（COVID–19）、优化（Optimization）、地理信息系统（GIS）等方向。其中，基于关键词共现网络的聚类结果下各主题的综合度更高，如主题 3 中因子值最高的关键词为物联网（Internet of Things），而云计算（Cloud computing）、边缘计算（Edge computing）、区块链（Blockchain）等均是物联网研究方向下重要组成部分；主题 5 中因子值最高关键词优化（Optimization）均是围绕着

神经网络（Neural networks）、人工神经网络（Artificial neural network）、算法（Algorithms）展开研究的。

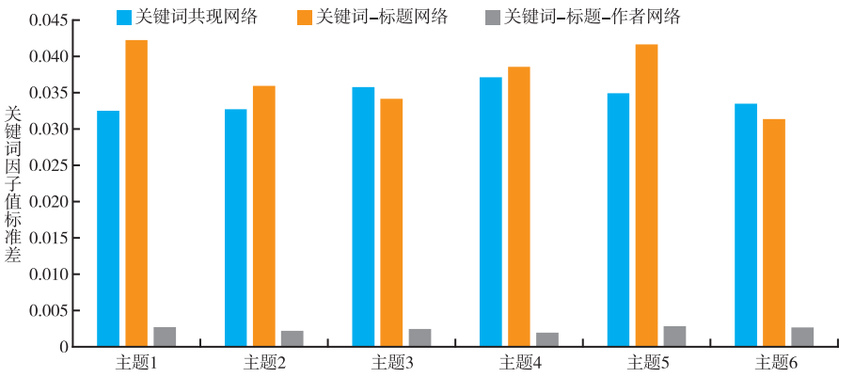

图 5　三种信息网络下各主题下关键词因子值标准差

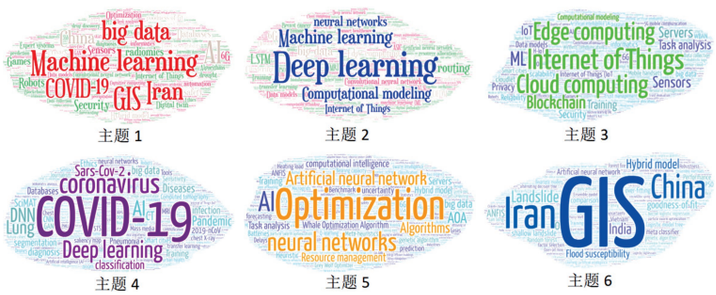

图 6　基于非负矩阵分解的同构信息网络主题识别结果词云图（以关键词共现网络为例）

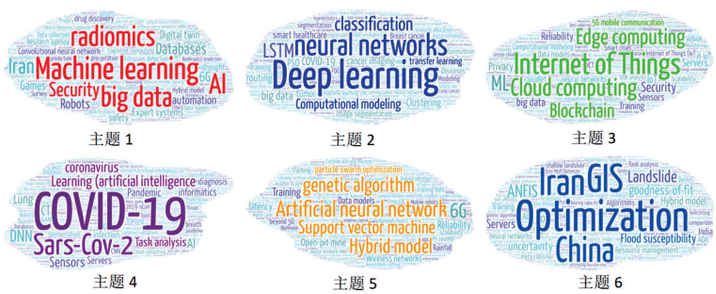

图 7　基于非负矩阵分解的异构信息网络主题识别结果词云图（以关键词 – 标题网络为例）

在基于异构信息网络的主题识别中，由于对象类型和语义关系的复杂性导致了各主题的丰富度增加，研究内容进一步细化。如图 8 所示，各主题下的因子值最高的两个关键词均为机器学习（Machine learning）、深度学习（Deep learning），其次因子较高的关键词还有物联网（Internet of Things）、新冠肺炎（COVID–19）、优化（Optimization）、地理信息系统（GIS）等，说明了人工

智能领域近五年高被引论文的核心研究内容是机器学习和深度学习，各主题纷纷围绕上述核心内容进行更深层次的研究。

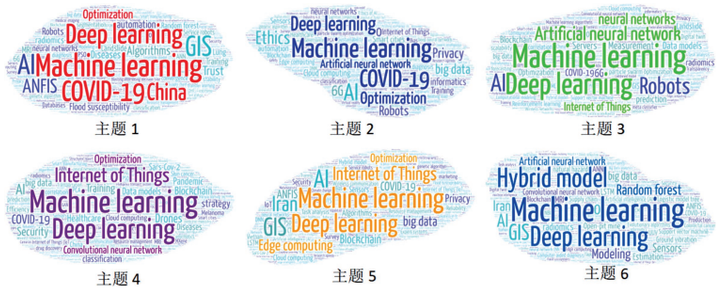

图 8　基于非负张量分解的异构信息网络主题识别结果词云图（以关键词 – 标题 – 作者网络为例）

3.2.2　聚类结果广度分析（纵向对比）

以上主要分析了三种聚类结果中的关键词信息，通过表 1~3 的聚类结果信息可以进一步看出，基于异构信息网络的聚类结果中主题所含对象类型更为丰富，表 2 和表 3 中各主题不仅包含了关键词，还包含了标题序号、作者类型对象，而标题可以辅助关键词进行类团命名，在挖掘主题内容的同时还可以得到该主题下包含的文献及作者团体。

表 1　基于非负矩阵分解的同构信息网络主题识别结果（以关键词共现网络为例）

主题 1		主题 2		主题 3	
关键词	因子值	关键词	因子值	关键词	因子值
Machine learning	0.958	Deep learning	0.881	Internet of Things	0.525
GIS	0.064	Machine learning	0.347	Cloud computing	0.386
Big data	0.060	Computational modeling	0.068	Edge computing	0.359
COVID–19	0.049	Neural networks	0.057	Blockchain	0.247
Iran	0.044	Internet of Things	0.054	Computational modeling	0.172
主题 4		**主题 5**		**主题 6**	
关键词	因子值	关键词	因子值	关键词	因子值
COVID–19	0.846	Optimization	0.783	GIS	0.659
Deep learning	0.205	Neural networks	0.241	China	0.387
Coronavirus	0.132	Artificial neural network	0.149	Iran	0.248
Sars–Cov–2	0.126	Resource management	0.141	Flood susceptibility	0.198
Classification	0.117	Algorithms	0.124	Hybrid model	0.195

表 2　基于非负矩阵分解的异构信息网络主题识别结果（以关键词 – 标题网络为例）

主题	关键词	因子值	标题序号	因子值
主题 1	Machine learning	0.976	459	0.117
	Big data	0.051	471	0.116
	Radiomics	0.042	260	0.114
	AI	0.037	130	0.110
	Security	0.036	97	0.110
主题 2	Deep learning	0.771	37	0.159
	Neural networks	0.056	379	0.159
	Computational modeling	0.042	259	0.154
	Classification	0.041	373	0.154
	Transfer learning	0.039	363	0.142
主题 3	Internet of Things	0.394	194	0.313
	Cloud computing	0.211	310	0.298
	Edge computing	0.204	176	0.283
	Blockchain	0.120	50	0.273
	5G mobile communication	0.092	397	0.257
主题 4	COVID–19	0.460	202	0.254
	Sars–Cov–2	0.081	171	0.210
	Learning artificial intelligence	0.075	499	0.203
	Coronavirus	0.074	156	0.194
	Task analysis	0.065	389	0.184
主题 5	Artificial neural network	0.416	3	0.320
	Support vector machine	0.088	524	0.267
	Genetic algorithm	0.085	340	0.256
	Hybrid model	0.080	137	0.241
	Particle swarm optimization	0.076	122	0.241
主题 6	Optimization	0.283	160	0.237
	GIS	0.259	267	0.228
	China	0.154	573	0.228
	Iran	0.097	80	0.215
	Flood susceptibility	0.086	164	0.215

表 3　基于非负张量分解的异构信息网络主题识别结果（以关键词 – 标题 – 作者网络为例）

主题	关键词	因子值	标题序号	因子值	作者	因子值
主题 1	Machine learning	0.038	259	0.039	Chen, W	0.083
	Deep learning	0.029	202	0.035	Tang, F X	0.039
	COVID–19	0.021	379	0.027	Zhao Z	0.035
	China	0.014	56	0.025	Oh, Y	0.027
	Optimization	0.012	216	0.023	Zhou, J	0.025
主题 2	Machine learning	0.030	97	0.028	Moayedi, H	0.045
	COVID–19	0.017	459	0.028	Allam, Z	0.028
	Deep learning	0.015	50	0.026	Huang, M H	0.028
	Optimization	0.011	176	0.026	Wahab, O A	0.028
	Artificial neural network	0.011	91	0.023	Wang, T	0.026
主题 3	Machine learning	0.041	158	0.034	Pham, B T	0.049
	Deep learning	0.019	338	0.034	Sands, T	0.034
	Artificial neural network	0.015	130	0.032	Wu, M H	0.034
	Neural networks	0.013	295	0.030	Dehghani, M	0.032
	Internet of Things	0.009	328	0.030	Huang, M H	0.030
主题 4	Machine learning	0.027	37	0.031	Wang, J J	0.031
	Deep learning	0.024	260	0.031	Lee, H	0.031
	Internet of Things	0.013	410	0.026	Costache, R	0.026
	Convolutional neural network	0.011	354	0.024	Hussain, T	0.026
	Optimization	0.009	17	0.022	Duan, J J	0.024
主题 5	Machine learning	0.040	26	0.027	Bui, D T	0.051
	Deep learning	0.018	85	0.025	Lv, Z H	0.034
	Internet of Things	0.016	102	0.025	Wang, X F	0.034
	Edge computing	0.016	373	0.025	Nhu, V H	0.027
	Optimization	0.011	384	0.025	Muhammad, K	0.025
主题 6	Machine learning	0.036	470	0.033	Nguyen, H	0.044
	Deep learning	0.028	182	0.028	Farivar, F	0.033
	Hybrid model	0.018	302	0.028	Zhan, J M	0.028
	Artificial neural network	0.013	313	0.028	Wang, G G	0.028
	Random forest	0.010	524	0.028	Sze, V	0.028

此外，基于表 1~3 还可以看出，随着异构信息网络对象类型和语义关系的增加，聚类结果中各类团研究内容的深度也在增加。例如，基于关键词共现网络的主题识别结果（表 1）中，表明人工智能领域近五年高被引论文研究热点主要集中在机器学习（Machine learning）、深度学习（Deep learning）、物联网（Internet of Things）、新冠肺炎（COVID–19）、优化（Optimization）、地理信息系统（GIS）等方向。基于关键词 – 标题网络的主题识别结果不仅能得到这几类研究热点，还能得到每类研究热点下的代表性文献，如表 2 中主题 3 关于物联网（Internet of Things）方向的研究，代表性文献有“Machine Learning Meets Computation and Communication Control in Evolving Edge and Cloud: Challenges and Future Perspective”（文献序号 194）、“Edge Intelligence and Blockchain Empowered 5G Beyond for the Industrial Internet of Things”（文献序号 310）、“MTES: An Intelligent Trust Evaluation Scheme in Sensor–Cloud–Enabled Industrial Internet of Things”（文献序号 176）等。基于关键词 – 标题 – 作者网络的主题识别结果中，主题内容挖掘进一步深入到作者层面，如表 3 中主题 4 和主题 5 都是聚焦于机器学习（Machine learning）、深度学习（Deep learning）在物联网（Internet of Things）上的研究，其中主题 4 是 Wang, J J、Lee, H、Costache, R 等作者关于物联网（Internet of Things）中卷积神经网络（Convolutional neural network）在区块链（Blockchain）

上的优化（Optimization）问题展开研究，代表性文章有“Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks”（文献序号 37）、“Fully Automated Deep Learning System for Bone Age Assessment”（文献序号 260）、“Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM”（文献序号 410）等，而主题 5 是 Bui, D T、Lv, Z H、Wang, X F 等作者关于物联网（Internet of Things）中大数据（Big data）、边缘计算（Edge computing）、优化（Optimization）问题展开研究，代表性文章有“Trustworthiness in Industrial IoT Systems Based on Artificial Intelligence”（文献序号 26）、“Security and Privacy in Smart Farming: Challenges and Opportunities”（文献序号 85）、“Reinforcement Learning-Based Multislot Double-Threshold Spectrum Sensing With Bayesian Fusion for Industrial Big Spectrum Data”（文献序号 102）等。

综上，本文借助于张量的数据结构形式，利用非负张量分解算法基于科技文献系统中的异构信息网络对领域主题进行挖掘，并设置了两组对照实验，实验组一对比了基于异构信息网络（关键词 – 标题网络、关键词 – 标题 – 作者网络）和基于同构信息网络（关键词共现网络）的主题识别效果的优势；实验组二对比了基于张量表示的异构信息网络（关键词 – 标题 – 作者网络）和基于矩阵表示的异构信息网络（关键词 – 标题网络）的主题识别效果优势，并对主题识别结果中的关键词个数、关键词权重、关键词内容以及主题特征项进行了评价。实验结果表明：在对科技文献进行主题识别时，基于异构信息网络的主题识别效果要比基于同构信息网络的主题识别效果好，主要体现在基于异构信息网络下的主题识别结果所包含的关键词个数和关键词权重值标准差相较于基于同构信息网络下有所降低，说明基于异构信息网络的主题识别效果更加精准深入，同时根据关键词内容和主题所包含的特征项个数也可以看出，高阶张量可以最大程度地保留科技文献异构信息网络的内部结构，基于非负张量分解算法的主题识别结果可以减少语义信息的丢失，相较于同构信息网络和基于矩阵表示的异构信息网络，基于张量表示的异构信息网络可以对主题进行更深层次的挖掘。

4　结束语

经典科技文献主题识别研究多基于科技文献同构信息网络进行分析，难以表达科技文献系统中丰富的对象类型和复杂的语义关系。本文利用异构信息网络完整地展示科技文献系统中各种类型对象之间的关系，获得更丰富和复杂的语义信息。通过将异构信息网络表示成一个高阶张量的形式，可以减少对网络内部知识结构的破坏，基于非负张量分解算法可以将不同类型的文献单元一次性同时划分到不同的类团中，减少潜在信息的损失，同时聚类结果的非负性以及类团对象的丰富性可以更有效地进行主题挖掘。本文尚存在一些不足之处：只选取了关键词、标题、作者三种文献单元，以人工智能领域近五年的高被引论文为例进行了小样本实验，实验结果是否适用于其他领域有待商榷。如何根据文献的全部信息单元构建高阶张量进行大规模样本实验，对领域内的知识结构进行更深层次的挖掘，则是本文后续研究进一步需要探讨的问题。

【参考文献】

［1］隗玲，许海云，胡正银，等. 学科主题演化路径的多模式识别与预测——一个情报学学科主题演化案

例［J］. 图书情报工作，2016，60（13）: 71–81.

［2］何伟林，奉国和，谢红玲. 基于CSToT模型的科技文献主题发现与演化研究［J］. 数据分析与知识发现，2018，2（11）: 64–72.

［3］梁镇涛，巴志超，徐健. 基于引文的跨学科领域发展路径分析——以眼动追踪领域为例［J］. 图书情报工作，2019，63（23）: 65–78.

［4］Shibata N, Kajikawa Y, Takeda Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications［J］. Technovation, 2008, 28(11): 758–775.

［5］Shibata N, Kajikawa Y, Takeda Y, et al. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications［J］. Technological Forecasting and Social Change, 2011, 78(2): 274–282.

［6］Yookyung Jo, Lagoze C, Giles C L. Detecting research topics via the correlation between graphs and texts［C］//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, San Jose, California, USA.

［7］Ma V C, Liu J S. Exploring the research fronts and main paths of literature: A case study of shareholder activism research［J］. Scientometrics, 2016, 109(1): 33–52.

［8］开滨，姚艳玲. 基于引文分析的我国高校科研优势学科探测研究［J］. 情报理论与实践，2018，41（5）: 44–49.

［9］郭倩影，杜建，李沛鑫，等. 基于引文网络的学术传承性文献识别方法研究——以2017年诺贝尔生理学或医学奖为例［J］. 情报杂志，2019，38（4）: 90–95，89.

［10］储节旺，钱倩. 基于词频分析的近10年知识管理的研究热点及研究方法［J］. 情报科学，2014，32（10）: 156–160.

［11］Walls F, Jin H, Sista S, et al. Probabilistic models for topic detection and tracking［C］// Acoustics, Speech, & Signal Processing, on IEEE International Conference. IEEE Computer Society, 1999: 521–524.

［12］Kleinberg J. Bursty and hierarchical structure in streams［J］. Data mining and knowledge discovery, 2003, 7(4): 373–397.

［13］奉国和，孔泳欣. 基于时间加权关键词词频分析的学科热点研究［J］. 情报学报，2020，39（1）: 100–110.

［14］余丰民，林彦汝. 基于关键词词频统计的学科研究热点漂移程度模型构建及实证分析［J］. 情报理论与实践，2020，43（2）: 100–105.

［15］Kim Y G, Suh J H, Park S C.Visualization of patent analysis for emerging technology［J］. Expert Systems with Applications, 2008, 34(3): 1804–1812.

［16］Neff M W, Corley E A. 35 years and 160, 000 articles: A bibliometric exploration of the evolution of ecology［J］. Scientmetrics, 2009, 80(3): 657–682.

［17］唐果媛，张薇. 国内外共词分析法研究的发展与分析［J］. 图书情报工作，2014，58（22）: 138–145.

［18］钟伟金，李佳，杨兴菊. 共词分析法研究（三）——共词聚类分析法的原理与特点［J］. 情报杂志，2008（7）: 118–120.

［19］叶春蕾，冷伏海. 基于共词分析的学科主题演化方法改进研究［J］. 情报理论与实践，2012，35（3）: 79–82.

［20］Yoon J, Park H, Kim K. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO–based content analysis［J］. Scientometrics, 2013, 94(1): 313–331.

［21］田鹏伟，张娴. 基于异构信息网络融合的专利技术主题识别研究［J］. 情报杂志，2021，40（8）: 45–52.

［22］吴继冰. 基于张量分解的异构信息网络聚类分析方法［D］. 国防科技大学，2017.

［23］吴继冰，黄宏斌，邓苏. 网络异构信息的张量分解聚类方法［J］. 国防科技大学学报，2018，40（5）: 146–152，170.

［24］Luo J, Gwun O. A Comparison of SIFT, PCA-SIFT and SURF［J］. International Journal of Image Processing, 2009, 3(4): 143-152.

［25］Cichocki A, Zdunek R, Phan A H, et al. Nonnegative Matrix and Tensor Factorizations:Applications to Exploratory Multi-way Data Analysis and Blind Source Separation［M］. Wiley Publishing, 2009.

［26］熊李艳，何雄，黄晓辉，等. 张量分解算法研究与应用综述［J］. 华东交通大学学报，2018，35（2）: 120-128.

［27］李亚芳，贾彩燕，于剑. 应用非负矩阵分解模型的社区发现方法综述［J］. 计算机科学与探索，2016，10（1）: 1-13.

［28］梁秋霞，何光辉，陈如丽，等. 基于非负张量分解的人脸识别算法研究［J］. 计算机科学，2016，43（10）: 312-316.

［29］王园园，赵亚娟. 基于非负矩阵分解的技术主题演化分析［J］. 图书情报工作，2018，62（10）: 94-105.

Research on Topic Recognition of Scientific and Technological Literature Based on Heterogeneous Information Network

Xi Chongjun　Xu Zhenzhen　Liu Wenbin　Ding Kai

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: ［**Purpose/significance**］The research on topic recognition of scientific and technological literature is of great significance for grasping the research focuses and hotspots in the field of science and technology, and revealing the development trend and evolution trend in the field. Traditional topic recognition research is mostly based on the homogeneous information network of scientific and technological documents, which is difficult to express the rich object types and complex semantic relations in the scientific and technological document system. ［**Method/process**］This paper uses heterogeneous information networks to express the rich semantic relations among various types of objects in the scientific and technological literature system, and converts them into the form of high-order tensors, and uses non-negative tensor decomposition algorithms to perform topic mining on them. ［**Result/conclusion**］The experimental results show that, based on the heterogeneous information network, it is possible to carry out deeper topic recognition of scientific and technological literature. The non-negative tensor decomposition algorithm is convenient and fast when processing heterogeneous information networks, and it can reduce the loss of semantic information.

Keywords: Topic recognition; Heterogeneous information network; High-order tensor; Non-negative tensor decomposition; Artificial intelligence

（本文责编：魏　进）