基于改进LDA模型的铁路领域主题发现研究*

龙艺璇[1] 安源[1] 王东晋[1] 翟夏普[1] 伊惠芳[2]
(1. 中国铁道科学研究院科学技术信息研究所, 北京 100081; 2. 中国科学院文献情报中心, 北京 100190)

摘要: 高效挖掘海量铁路领域科研成果数据中蕴含的主要内容是铁路领域科研人员在大数据时代亟待解决的重要问题。LDA模型是用于主题发现的主流方法, 但在面向多单词短语居多的铁路领域研究文献时存在使用受限的问题, 因此本文提出一种LDA模型的改进算法: 一方面在构建主题模型前, 对文本作预处理时抽取语料中的名词短语、动词短语、名词和动词; 另一方面在主题模型构建完成后, 融合TextRank算法与PMI算法得出关键词组块, 并以此替换LDA主题识别结果中的主题词, 进一步丰富主题的语义。最后, 以铁路领域"牵引供电系统"为例开展实证研究。结果表明, 本文提出的改进LDA模型有助于提升铁路领域主题发现结果的可解释性与可识别性, 可以为后续铁路领域科研管理中知识服务的实现提供有效的方法支持。

关键词: 主题发现; 铁路领域; 语义增强; LDA主题模型
中图分类号: G350.7 DOI: 10.3772/j.issn.1673-2286.2022.02.004
引文格式: 龙艺璇, 安源, 王东晋, 等. 基于改进LDA模型的铁路领域主题发现研究[J]. 数字图书馆论坛, 2022 (2) : 26-32.

"千古百业兴, 先行在交通。"中国铁路营业里程从1949年时仅有2.18万公里到2021年底突破15万公里[1], 中国铁路已然成为中国走向世界的亮丽名片。中国铁路取得举世瞩目成就的背后, 离不开科研人员的努力和科研成果的支撑。如今, 铁路领域科研成果数量与日俱增, 科研成果类型百花齐放。面对海量的多源异构铁路领域科研成果资源, 如何在短时间内对科研成果内容开展有效遴选, 成为铁路科研人员亟待解决的重要问题。

在科学学视角下, 主题的内涵包含两个层面: 一是单篇科技文献中的主题思想, 二是某学科领域下多篇科技文献的总体研究方向。从主题的表现形式来看, 主题由可以表征主题语义的主题词或主题短语构成[2]。本文探讨的主题为第二种, 即某学科领域下的主要研究方向, 并聚焦于铁路领域。主题发现可以实现从海量的文本中抽取主要语义内容, 有助于科研知识的梳理与归纳。当前已有众多学者尝试利用主题发现算法开展特定科研领域的主题发现研究[3-5]。本文通过调研得知, 相较于传统的文献计量方法, 以LDA (Latent Dirichlet Allocation) 主题模型为代表的主题发现算法因其能深入文本内容、适应大规模文本分析、支持计算主题之间的相关关系等优势, 成为主题发现的主流算法[6]。不过传统的LDA主题模型基于词袋 (bag of words) 模型实现, 即假设所有的单词都是相互独立的, 由于铁路行业技术化程度高、专业交叉属性强、科研成果类型丰富, 使得铁路领域科研成果呈现多源异构及多单词短语居多等特点[7], 这就极大地限制了传统LDA主题模型在铁路领域科研成果主题发现上的应用。

基于此, 本文从铁路领域科研成果特点出发, 在传统LDA主题模型基础上, 提出一种适用于英文文献的LDA模型改进算法。该算法与传统LDA算法相比, 能够

* 本研究得到中国铁道科学研究院集团有限公司科研开发项目"铁路科研知识图谱及智能知识服务体系研究" (2020YJ147) 资助。

提升多单词短语的识别效果，增加主题发现结果的可解释性和可识别性，最终实现铁路领域海量科研成果数据的语义内容分析，为科研人员快速遴选科研内容以及提高科研工作效率提供支持。

1 主题发现相关理论及方法

1.1 主题发现方法

主题发现（topic discovery）又被称为主题识别，指利用一系列语义理解方法，从复杂的大规模信息源中抽取关键词或术语，并在此基础上加以聚类，从而发现文献主题的技术方法，旨在处理和分析大规模信息并且使用户以快速有效的方式了解信息内容，发现信息中的主题[8]。

目前，主题发现主要有基于文献计量的方法和基于文本挖掘的方法两大类。文献计量分析是科学学和情报学领域的重要研究方法，基于文献计量的主题发现方法依靠文献引证、关键词等文献特征，采用引文网络聚类、词频分析、共词分析等方法开展主题挖掘研究，以期为研究前沿发现、技术机会识别、学科领域演化、研究趋势归纳等研究提供方法支持。这类主题发现方法计算成本普遍较低，且方法相对成熟，但普遍存在语义表示匮乏、主题之间关系无法揭示等缺点，不利于主题内容的解读与领域知识内容的分析。随着计算机技术逐渐渗透各个学科，加之文献数量急剧增加，近年来以文本聚类、主题模型为代表的文本挖掘方法开始应用于主题发现，为主题发现研究提供了新思路与新方法。基于文本挖掘的主题发现方法的优势在于能够深入文本内容揭示主题，但基于文本聚类的方法仍处于探索阶段，文本特征提取和聚类方法仍需进一步完善。目前，主题模型已成为基于文本挖掘主题发现领域的主流方法，近年来已有众多学者尝试利用主题模型开展主题发现研究，例如：Fang等[9]采用LDA模型从图书馆领域文献摘要中提取潜在主题，然后对文档-主题进行回归分析，并区分出冷门研究主题和热门研究主题；王曰芬等[10]采用LDA模型以国内知识流领域为例，多维度对比该领域全局主题和学科主题的差异性。

1.2 主题模型

主题模型是一种语言模型，可以发现一系列文档中隐藏的主题信息，最终实现文本语义层面的挖掘。应用主题模型可以快速扫描大数据量文本，协助研究人员理解文本内容，迅速掌握文本内容重点。目前主题模型已经广泛应用到文本聚类[11]、主题演化[12]等众多研究中。从主题模型的原理来看，可以将其视为一种生成概率模型，其基本思想可以总结为：每一篇文档都可以看作多个主题构成的概率分布，而每一个主题都可以看作多个主题特征词构成的概率分布[13]。1999年，Hofmann[14]提出PLSA（Probabilistic Latent Semantic Analysis）主题模型，开启了文本分析领域主题挖掘新篇章。目前最具有代表性的主题模型是Blei等[15]提出的LDA模型。

2 LDA主题模型的改进

随着LDA主题模型在主题发现领域的广泛应用，有学者发现LDA主题模型的识别结果在语义可解释性上仍存在很大提升空间[16]。这是因为LDA主题模型依赖词袋模型实现，即假设所有的单词都是相互独立的，而在许多文本挖掘任务开展过程中，语序和短语往往是捕获文本语义的关键，不但对于句法分析很重要，而且对于单词语义同样重要，一个短语的语义远远超越了单一单词[17]。本文聚焦的铁路领域正是传统LDA主题模型使用受限的典型领域，多单词短语形式表现的科研术语占比多。以机电系统（electromechanical system）为例，主要研究方向包含信号通信系统（signal communication system）、控制系统设备（control system equipment）、牵引供电系统设备（traction power supply system equipment）、车站设备（station equipment）、监控设备（monitoring equipment）等，因此语序与短语对于铁路领域科研成果的内容表示至关重要。

经调研，目前国内外学者在LDA主题模型基础上开展语义优化方面的改进模型主要有LDA Collocation模型[18]、PhraseLDA模型[19-20]和Chunk-LDAvis模型[17]，现有模型的改进体现出两种思路：一是在预处理阶段开展短语抽取研究，之后采用LDA主题模型对词组进行建模；二是主题模型构建完成后再对主题词进一步丰富语义。笔者认为，两种改进思路均有可取之处，且可以互为补充。第一种思路在预处理阶段抽取短语，可以尽可能将原有语序信息保留；第二种思路可以解决主题词中短语占比较低问题，进一步扩充语义。因此，笔者尝试从这两个角度同时对LDA主题模型进行改进，

充分扩充其主题发现结果的语义内涵，具体算法流程如图1所示。

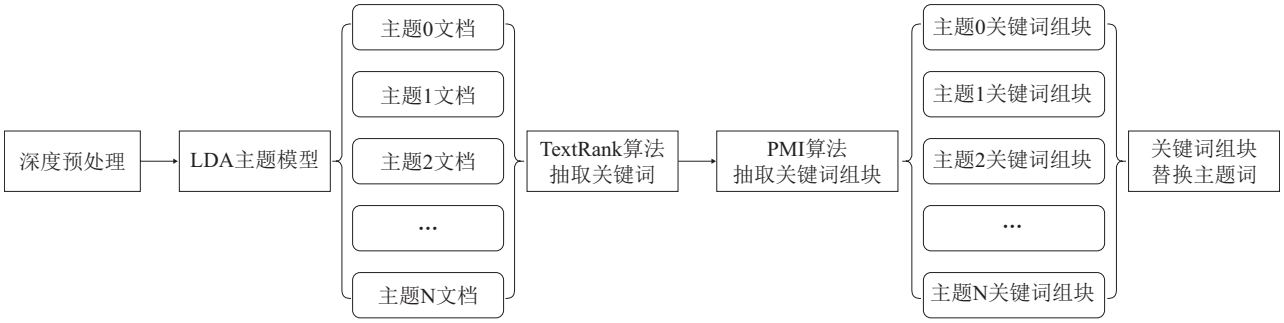

图1 改进LDA模型算法流程

词性标注、词形还原、去停用词等，还基于Python语言nltk自然语言处理库中的pos_tag对文献标题和摘要进行词性标注，标注的词性包含名词（NN）、动词（VB）、形容词（JJ）、副词（RB）、介词或从属连词（IN）、连词（CC）和感叹词（UH）。再根据词性标注结果在语料中抽取其中的名词短语、动词短语、名词和动词。

（2）LDA主题模型。对文档进行LDA主题识别后，计算每一个文档所属概率最大的主题，将其标识为该文档所属主题。按照所属主题对文档进行分类，同一主题的文档划分为一组，作为该主题下主题词扩充的语料来源。

（3）TextRank算法抽取关键词。目前文本关键词抽取算法主要有TF-IDF算法和TextRank算法。两者的区别在于以下3点。①从时间复杂度来看，TF-IDF算法在开始运算之前需要构建一个语料库，在之后的统计过程中还需要在已构建的语料库中进行读取操作，这些步骤的时间代价比较高；TextRank算法的时间主要花费在迭代算法的计算上，因此当文本数据量较小时，TextRank算法的时间复杂度应更低一些，而当文本数据量特别大时，两者差距不大。②从空间复杂度来看，TF-IDF算法主要集中在语料库对空间的需求；对于TextRank算法来说，主要集中在图链接矩阵的存储，但相比较而言，TF-IDF算法对内存的要求更高一些。③从语义复杂度来看，TF-IDF在计算的过程中没有考虑文档中词间的关联性，仅基于单词出现的次数计算，而TextRank算法在构建有向加权图时考虑了词间的关联性。基于以上分析，使用TextRank算法对每一主题下的扩充语料抽取其关键词，并按照关键词权重降序排列。

（1）深度预处理。在构建LDA主题模型前需要对文本进行预处理。除了进行常规的特殊字符剔除、分词、

（4）PMI算法抽取关键词组块。使用TextRank算法抽取某主题下的关键词之后，为了扩展主题词语义信息，需要考虑关键词之间的相关性，因此提出关键词组块的概念，即关联程度密切的关键词。在信息论中，互信息（NMI）是对信号之间关联程度的描述[21]。PMI正是从互信息中衍生而来。PMI从统计的角度出发，通过计算词语之间的共现次数，得出词语之间的相关性，即统计两个关键词在文本中同时出现的概率，如果概率越大，其相关性就越高，关联度也越高。通过PMI算法计算词语之间的相关性，可以找到与关键词最相关的组块，相比于单词而言，关键词组块包含的语义信息会更加丰富。因此，用PMI算法计算结果中最大概率的关键词组块替换相应主题下的主题词，可以提高主题发现结果的可解释性与可识别性。

3 改进LDA模型在铁路领域主题发现中的应用

3.1 数据获取与实验条件

本文从Web of Science核心合集（以下简称WoS核心合集）中获取相关实验数据，并选择“牵引供电系统”（traction power supply system）研究方向开展实证研究。牵引供电系统可以将地方电网中的电能源源不断地输送至动车组，为动车组的高速运行持续提供强大电能，其安全性与稳定性是动车组运行的重要保障。铁路牵引供电系统主要包含三部分，分别是牵引变电所、自耦所和牵引网，其设计与实现离不开材料科学与工程、计算机科学与技术、电气工程、交通运输工程

等多学科的交叉融合，技术性强、精度要求高、结构复杂、专业性明显，是铁路领域重要的研究方向。因此，通过主题发现研究协助科研人员及时梳理牵引供电系统的研究进展，把握牵引供电系统的研究主题具有重要意义。

依托中国铁道科学研究院，在铁路领域相关研究方向专家的指导下，构建英文检索式。由于WoS核心合集类别中未将“铁路”相关研究单独列为一个研究领域，因此以“traction power supply”为主题词进行英文文献检索，检索时间限定为2017—2021年，文献类型选择“paper”，再请相关研究方向专家人工对检索结果进行筛选，剔除与铁路领域不相关的文献，最终得到774条检索结果。

实验基于Window10系统，内存为16GB，处理器为X64。经典LDA算法选择的Python3版本下第三方模块LDA工具包，具体运行时LDA主题模型的参数设定为alpha（document—topic associations）=50/k，beta（topic—term associations）=0.01，5 000次迭代，其中k代表设定语料库中的最优主题个数。关于最优主题个数的选取，选择Perplexity-Var指标，该指标将主题相似度与困惑度相结合，使用JS散度表征主题相似度，并将相似度视为随机变量，引入随机变量方法作为主题抽取结果差异性与稳定性的评判依据[22]，改善了困惑度指标数目偏大、辨识度不高的缺陷[23]。

3.2 实验过程

第一，将WoS核心合集中获取的数据导入MySQL数据库，利用中英文标题、摘要和关键词开展深度预处理，抽取其中的名词短语、动词短语、名词和动词。

第二，利用传统LDA主题模型对2017—2021年时间窗内深度预处理后的数据进行建模，根据Perplexity-Var指标，最优主题个数确定为10。

第三，对建模后的数据开展语义扩充。使用PMI算法和TextRank算法融合处理，成功抽取出共同出现频率高且相对重要的词语组块，按照PMI值降序排列，与原有主题词对比后替换，实现关键词组块对主题词扩充。由于篇幅有限，下文实验结果部分仅列出语义扩充后主题强度排名前三的主题，对比主题语义扩充前后主题词变化，展示改进LDA模型主题发现结果的可解释性与可识别性的优越性。主题强度主要描述主题在某时间窗口中的活跃程度。在同一时间窗口中主题强度值越大，说明主题热度越高，受到的关注程度越高。主题强度计算参考任智军等[24]在2015年提出的算法，该算法已被业内学者认可并广泛应用[25]。

3.3 实验结果与分析

为更清晰地展示改进LDA主题模型与传统LDA主题模型建模结果的不同，本文以表格形式展示建模结果（见表1）。

表1 两种主题模型对比结果（主题强度排名前三）

主 题	传统LDA主题模型主题词	改进LDA主题模型主题词	主 题
T0：牵引供电系统环保相关研究	traction power system，substation，energy conservation，renewable energy，efficiency，regenerative，management，energy storage，battery，modeling	traction power system，traction substation，energy conservation renewable energy，energy efficiency，regenerative braking energy，energy management，energy storage system，battery，load modeling	T0：提升牵引供电系统的能源效率，集中于再生制动能量的利用以及新型材料的应用
T4：牵引供电系统电能质量研究	traction power supply system，power quality，electric railway，high-speed railway，resonance，power system，frequency，traction load，compensation，filter	traction power supply system，power quality，electric railway，high-speed railway，harmonic resonance，power system harmonics，low-frequency oscillation，traction load，reactive power compensation，power harmonic filter	T4：提升牵引供电系统的电能质量，集中于谐波谐振的特征识别与抑制技术研究
T7：牵引供电系统故障识别研究	rail transportation，traction power supply system，model，stray current，fault，motor，monitoring，transformer，analytical model，catenary	rail transportation，traction power supply system，mathematical model，stray current，fault diagnosis，traction motor，monitoring，traction transformer，analytical mode，catenary fault	T7：牵引供电系统故障诊断与定位研究，特别是接触网相关故障

表1中，右侧展示的为采用改进LDA主题模型进行语义扩充后的关键词组块，可见经语义扩充后，语义信息明显更加丰富。T0中，左侧表格根据“traction power system”“energy conservation”“renewable energy”等主题词确定该研究方向主要内容为铁路领域牵引供电系统环保相关研究，但难以聚焦到具体研究方向，在右侧表格中“energy efficiency”“regenerative braking energy”“energy management”“load modeling”等新增关键词组块，可以进一步将研究方向集中在通过再生制动能量的充分利用以及新兴复合材料尝试等方式开展牵引供电系统的能源效率研究，具体主要包括制动储能控制、能量评估方法研究、再生制动仿真建模研究等研究内容；T4中，左侧表格根据“traction power supply system”“power quality”“electric railway”等主题词确定该研究方向主要内容为铁路领域牵引供电系统电能质量研究，但主题词描述相对模糊，在右侧表格中“harmonic resonance”“power system harmonics”“power harmonic filter”“reactive power compensation”等新增关键词组块，将研究方向进一步集中于无功、谐波、负序和末端电压偏低等4类电能质量问题，由于电气化铁路的单相独立不对称性、负载电流随机波动性、谐波的频谱分布广等特点，负序、谐波问题影响也日益严重，从而影响到电力系统公共电网。无源和有源电力滤波器的应用研究用以谐波治理，其中SVG控制无功补偿、APF控制谐波补偿，在交流牵引网中逐步使用SVG来兼顾低次谐波的治理；T7中，左侧表格根据“rail transportation”“traction power supply system”“fault”等主题词确定该研究方向总结为铁路领域牵引供电系统故障处理研究，而在右侧表格中，根据扩充的“mathematical model”“fault diagnosis”“traction transformer”“catenary fault”等关键词组块，可以将研究内容进一步锁定在牵引变压器、接触网故障、牵引电机故障等，此外，采用数学模型或仿真模型针对牵引供电系统故障分析属于故障定位分析的主要途径。通过以上分析，可以证明本文提出的主题模型改进方法对铁路领域文本数据的主题发现有较为明显的提升作用。

3.4 实验结果验证

本文主要采取了主观与客观两种实验验证方式。首先，本文依托中国铁道科学研究院，采用专家验证方式，咨询机车车辆研究所牵引供电系统相关研究专家对于本文主题发现结果的指导意见。专家认可本文提出的改进LDA主题模型在提升主题发现语义时发挥的积极作用，认为该算法相比较于传统的LDA主题模型发现结果，更为准确地总结了国际上牵引供电系统的主要研究方向，并相对明确地指出了每个方向的主要研究突破点，同时提出增加可视化效果等未来工作指导。此外，本文通过将2017—2021年铁路领域牵引供电系统研究方向下WoS核心合集所有科技文献数据导入VOSviewer软件，依据共现关系，构建关键词共词网络，自动生成“Network Visualization”，通过调整“Attraction”参数为1，“Repulsion”参数为0，“Resolution”参数为0.7，最终得出界限较为清晰的3个关键词聚类簇。通过与本文提出的改进LDA主题模型的主题发现结果对比，关键词聚类结果与本文改进LDA主题模型建模后得出的主题强度排名前三的主题基本吻合。其中，关键词类1根据“traction power supplies”“energy management”“energy storge system”等关键词可以得出与T0内容基本一致，均为铁路领域牵引供电系统能源节约环保相关研究；关键词类2根据“power quality”“harmonics”“electrified railway”等关键词可以得出与T4内容相符，均为铁路领域牵引供电系统电能质量提升相关研究；关键词类3根据“traction power supply system”“fault diagnosis”“analytical model”等关键词可以得出与T7内容更为吻合，均为铁路领域牵引供电系统故障发现与处理相关研究。此外，通过对比可以得出，本文主题发现结果的主题词中，多单词短语出现频率要比关键词共现聚类结果更高，因此对于研究内容的概括更全面，主题发现结果语义可解释性更强。综上，证明本文提出的改进LDA主题模型可以极大提升铁路领域英文文本数据主题发现能力。

4 结语

伴随着大科学时代的到来，科研管理信息服务除满足日常科研管理需求之外，面对科研人员科研信息服务的新需求，亟需调整自身服务业态，实现从信息服务到知识服务的结构性变革。快速遴选重要研究成果，是铁路领域科研人员面对海量国际铁路科研成果资源时亟待解决的重要问题。以LDA主题模型为代表的主题发现方法可以实现从海量的文本数据中快速抽取主要语义内容，为海量铁路科技信息资源的有效梳

理与组织提供支持。本文针对传统LDA主题模型在面对多单词短语居多的铁路领域研究文本时使用受限的问题，创新性地提出改进LDA主题模型，并以铁路领域“牵引供电系统”为例，验证了该算法有助于提升铁路领域主题发现结果的可解释性与可识别性，可以为后续铁路领域科研管理中知识服务的实现提供技术支撑，也可以为科研人员在针对其他多单词短语居多的领域开展主题发现研究时提供有效的方法支持。

参考文献

[1] 2021年全国铁路营业里程突破15万公里 高铁超4万公里[EB/OL]. [2022-01-16]. http://finance.people.com.cn/n1/2022/0105/c1004-32324386.html.

[2] 逯万辉，谭宗颖. 基于知识基因游离与重组的领域主题演化研究[J]. 情报理论与实践，2019，42(2)：101-107.

[3] 游毅，索传军. 国内信息生命周期研究主题与趋势分析——基于关键词共词分析与知识图谱[J]. 情报理论与实践，2011，34(10)：17-21.

[4] MENG Z Q，SHEN S M，CHEN Q L. A network decomposition-based text clustering algorithm for topic detection[J]. Applied Mechanics & Materials，2013(239/240)：1318-1323.

[5] 黄先蓉，张窈. 我国新闻出版领域有关意识形态研究的主题结构及演化趋势[J]. 出版科学，2019，27(6)：19-25.

[6] CHOI S Y，SEO J Y. An exploratory study of the research on caregiver depression：using bibliometrics and LDA topic modeling[J]. Issues in Mental Health Nursing，2020，41(7)：592-601.

[7] 赵占芳，刘鹏鹏，李雪山. 基于改进TextRank的铁路文献关键词抽取算法[J]. 北京交通大学学报，2021，45(2)：80-86.

[8] 图书馆·情报与文献学名词审定委员会. 图书馆·情报与文献学名词[M]. 北京：科学出版社，2019.

[9] FANG D，YANG H，GAO B，et al. Discovering research topics from library electronic references using latent Dirichlet allocation[J]. Library Hi Tech，2018，36(3)：400-410.

[10] 王曰芬，傅柱，陈必坤. 基于LDA主题模型的科学文献主题识别：全局和学科两个视角的对比分析[J]. 情报理论与实践，2016，39(7)：121-126.

[11] 曲靖野，陈震，郑彦宁. 基于主题模型的科技报告文档聚类方法研究[J]. 图书情报工作，2018，62(4)：113-120.

[12] 王丽，沈湘. 文本预处理后的LDA模型主题发现与技术演进研究[J]. 农业图书情报，2019，31(4)：19-28.

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

[14] HOFMANN T. Probabilistic latent semantic analysis[C]// Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco. USA：Morgan Kaufmann Publishers Inc，1999.

[15] BLEI D M，ANDREW Y N，MICHEAL I J. Latent dirichlet allocation[J]. The Journal of Machine Learning Research，2003(3)：993-1022.

[16] LONG Y X，CHEN L Y，YI H F，et al. Semantic Provenance of Innovation Research[EB/OL]. [2022-01-15]. https://www.asist.org/sig/sigmet/events/.

[17] 刘自强，许海云，岳丽欣，等. 基于Chunk-LDAvis的核心技术主题识别方法研究[J]. 图书情报工作，2019，63(9)：73-84.

[18] GRIFFITHS T L，STEYVERS M，TENENBAUM J B. Topics in semantic representation[J]. Psychological Review，2007，114(2)：211-244.

[19] E-LKISHKY A，SONG Y，WANG C R，et al. Scalable topical phrase mining from text corpora[J]. Proceedings of the VLDB Endowment，2014，8(3)：305-316.

[20] LI B，WANG B，ZHOU R，et al. CITPM：A cluster-based iterative topical phrase mining framework[C]//International Conference on Database Systems for Advanced Applications. Dallas：Springer International Publishing，2016：197-213.

[21] ROBERTO B. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks，1994，5(4)：537-550.

[22] 关鹏，王曰芬. 科技情报分析中LDA主题模型最优主题数确定方法研究[J]. 现代图书情报技术，2016，32(9)：42-50.

[23] NIKITA M. Select number of topics for LDA model[EB/OL]. [2020-11-15]. https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html.

[24] 任智军，詹淑琳，范婷婷. 基于论文与专利整合数据的研究方向发现模型研究[J]. 情报工程，2015，1(4)：88-94.

[25] 李昌，伊惠芳，吴红，等. 无人驾驶汽车专利技术主题分析——基于WI-LDA主题模型[J]. 情报杂志，2018，37(12)：42，50-55.

作者简介

龙艺璇，女，1994年生，博士，研究方向：科学计量、科技情报分析、科技信息组织与管理。
安源，男，1963年生，硕士，正高级工程师，通信作者，研究方向：科研信息管理与服务，E-mail：13661388282@139.com。
王东晋，男，1992年生，硕士，工程师，研究方向：信息技术。
翟夏普，女，1994年生，硕士，助理工程师，研究方向：信息组织与信息服务。
伊惠芳，女，1993年生，博士研究生，研究方向：科学计量、专利分析、文本挖掘。

Research on Railway Field Topic Discovery Based on Improved LDA Model

LONG YiXuan[1] AN Yuan[1] WANG DongJin[1] ZHAI XiaPu[1] YI HuiFang[2]
(1. Scientific & Technical Information Research Institute, Chinese Academy of Railway Sciences, Beijing 100081, P. R. China; 2. National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract: The era of big data has brought difficulties for researchers in the railway field to quickly select the main research directions, obtain international research trends, and understand international research hotspots. Efficiently excavating the main content contained in the massive scientific and technological literature in the railway field has become an important problem to be solved urgently by researchers in the railway field. In view of the fact that the topic model represented by LDA is used as the mainstream method for topic discovery, there is a problem of limited use in the face of scientific and technological literature in the railway field with many multi-word phrases. In this study, we innovatively propose a semantic enhanced LDA topic model. On the basis of in-depth preprocessing of extracting nouns phrases, verb phrases, nouns and verbs, we combine TextRank algorithm and PMI algorithm to obtain keyword chunks. We use the sorted keyword chunks to replace the topic words in the LDA topic recognition results. In this study, we conduct an empirical study on the "traction power supply system" as an example. The results show that the semantic enhanced LDA topic model proposed in this paper can help to improve the interpretability and recognizability of topic discovery results in the railway field. In addition, it can also provide effective method support for the realization of knowledge services in scientific research management in the railway field.

Keywords: Topic Discovery; Railway Field; Semantic Enhancement; LDA Topic Model

（收稿日期：2022-01-20）