基于知识元的学术论文内容创新性智能化评价研究*

李贺[1] 杜杏叶[1,2,3]
[1] 吉林大学管理学院 长春 130002 [2] 中国科学院文献情报中心 北京 100190
[3] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要 [目的/意义] 创新性是对学术论文质量最基本的要求,是学术论文的灵魂,是学术论文评价的核心。知识元是学术论文基本组成单元。基于知识元理论和机器学习相关理论与算法,从学术论文内容层面研究计算机如何智能化地进行创新性评价及其实现过程与方法。 [方法/过程] 首先,构建学术论文的研究问题、理论、方法、结论 4 个知识元本体,接着提出基于知识元的学术论文创新性判断模型。其次,根据学术论文研究特点,构建理论与方法机器分类模型及知识元的抽取规则与抽取方法,建立规则库和知识语料库。最后,基于语义相似度计算方法,根据判断规则和相关权重对学术论文 4 个维度的创新性进行评分。 [结果/结论] 基于知识元抽取的学术论文创新性评分系统的实证结果表明,该智能化评价方法具有一定的可行性,可为学术论文内容创新性智能化评价系统的最终实现提供方法借鉴。

关键词: 学术论文 知识元 内容创新性 智能评价
分类号: G230
DOI: 10.13266/j.issn.0252-3116.2020.01.012

创新是国家发展的重要基础。从国家知识创新战略层面来看,知识创新反映国家科研产出能力、知识传播能力和科技整体实力。以学术论文为载体的科研成果是知识创新的重要组成部分,学术论文的质量和数量是衡量一个国家创新能力与活力的重要标志。学术论文评价是知识创新能力评价的基础与重要内容,是国家综合创新能力测度体系的重要指标之一。学术论文评价的核心在于对学术论文质量、学术价值和学术影响力的评价。当前,学术论文发表前评价主要通过专家匿名评审方式进行,这种评审方式受专家自身学术水平和学科领域方向等限制,具有一定的局限性,可能使一些好的成果被遗漏或迟滞发表,一些不好的成果则发表在有影响的期刊上,从而给国家综合创新能力评价带来负向影响。知识管理、大数据及人工智能技术的发展为克服学术论文评审弊端提供了新的可能性。同时,由于学术论文中的知识元不仅可以用来表达、存储、检索和利用知识,还可以用来描述知识的发展脉络,进行知识发现。故本文尝试基于知识元理论,借助大数据及人工智能技术,研究学术论文创新性智能化评价的理论与方法。

1 学术论文创新性评价概述

创新性是对学术论文质量最基本的要求,是学术论文的灵魂,是学术论文评价的核心。学术论文的创新性评价包含多个维度,从内容来看,包括观点的创新,即在某一领域提出了他人所不曾提出的观点或研究问题;学术理论创新,即发现了新现象或揭示了新规律,或者是提出了新的理论;结构或方法的创新,即在已有研究的基础上提出崭新的视角或者研究方法,或对现有的方法进行了改进、完善,或者利用现有的方法解决应用领域中存在的新问题;结果结论创新,结果结论创新伴随着理论创新和方法创新,指在前 3 个创新基础上,获得了与原有成果不同的结果,得出了不同的结论。

从创新程度方面来看,陈建青[1]将创新程度划分为开创性、独创性和改进性 3 个层次。其中开创性的研究成果是指在国内外某个专业学科或领域所做的具有深远影响的、具有全局性、前瞻性、战略性、突破性、颠覆性的创新成果,是最具创新性的成果;独创性的研究成果是指在某个已有专业领域提出某项新的课题并做出具有原始创新或独立知识产权的研究成果;改进

* 本文系国家自然科学基金面上项目"基于图模型的多源异构在线产品评论数据融合与知识发现研究"(项目编号:71974075)研究成果之一。
作者简介: 李贺(ORCID:0000-0001-8847-3619),教授,博士生导师;杜杏叶(ORCID:0000-0001-5016-0561),副研究馆员,副编审,硕士生导师,博士,通讯作者,E-mail:duxy@mail.las.ac.cn。
收稿日期:2019-12-11 本文起止页码:93-104 本文责任编辑:王传清

性的研究成果指在已有研究课题、研究对象及其研究成果的基础上，做进一步的补充、改进或完善性的研究工作。创新是有层次性的[2]，层次最高的是开创性创新，其次是阶段性创新，第三是应用性创新。从国外权威学术期刊的界定看，*Nature* 认为创新的科研成果应具备新颖性、引人注意，而且该项研究在该领域之外还具有广泛的意义。*Science* 则认为创新是指对自然或理论提出新见解，而不是对已有研究结论的再次论证。

关于创新性的评价方法，目前国际学术界最为认可的就是同行评议[3]，引文分析也是当前使用得较多的一种方法。如文献计量领域的学者[4]认为，论文具有高创新力特点时，该论文更有可能成为高被引论文。但同行评议和基于引文的文献计量分析这两种方法在评价学术论文创新性时都存在一定的局限性。同行评议的局限性主要表现在：①严重依赖评审专家的主观判断，评审标准不一。②选择有资格的同行范围很窄，评审专家给出评价的可信度难以确定。③难以保证匿名评审的公正性。同时，同行评议作为定性评价方法还存在主观随意、低效性、评价过程隐蔽、结果难以复证和监督等缺陷[5]。而量化评价则因学科的不同其普遍适用性较为有限，应用计量学来评价学术成果还可能误导学者[6]。

当前，基于同行评议的学术论文发表前评价与基于文献计量（涵盖补充计量）的学术论文发表后评价所具有的缺陷没有得到根本的解决。对同行评议的缺陷进行改进的研究，如开放式同行评议[7]、同行评议专家遴选问题研究[8]、对同行评议表单进行量化处理[9]等，虽然在一定程度上解决了如公正性要求等方面的缺陷，但无法解决同行评议中基于专家主观判断的问题；对于以文献计量为基础的学术论文影响力评价虽然经过多次改进，如补充计量指标 Altmetrics 对引用的评价，但始终无法从学术内容层面解决影响力问题。

研究人员因此将研究方向转向研究学术论文内容本身，通过对学术论文的内容分析与挖掘将学术论文中的创新点进行识别并构建学术论文创新力测度指标来评价学术论文的创新性。目前，已经有一些相关研究。如沈阳[10]从关键词角度对论文创新度进行评价，利用统计不同时期的关键词频度的方法，对已有的关键词进行提取，该研究认为词频越高，时间越长，用户评价越低，则创新度越低；贺婉莹[11]从创新吸收、创新扩散的角度从多个维度对论文创新力进行评价，虽然取得了一定的效果，但对学术论文本身的内容因素关注仍较少。索传军等[12]利用学术论文中知识元转移的数量测度单篇学术论文的老化度和创新度；杨京等[13]基于研究主题对学术论文的创新力进行评价，认为如果某篇学术论文的研究主题和当前的科学研究前沿主题相契合，同时发表在影响因子较高的期刊上，那么这篇论文即具有较高的创新力。阮光册[14]采用 Doc2Vec 方法对文本内容进行向量计算与相似度计算以生成热点选题论文集，在此基础上再利用主题模型和聚类算法进行主题识别与挖掘，在语义特征的识别上获得了更优的效果，可以用来对学术论文主题新颖性和创新性进行识别和判断，是本文研究内容创新性智能化评价的重要基础。

2 学术论文智能化评价概念与过程

评价是指在一定的标准下，对评价对象进行比较分析，使用户更好地认识评价对象，并指导用户做出决策[15]。智能化评价是将人工智能的理论、方法、技术运用于评价对象并对评价对象进行认识的过程。学术论文智能化评价是指在学术论文评价过程中，判断论文质量好坏的若干关键指标可以由人工智能技术进行判断，或者说由计算机程序自动完成，并给出评价结果的评价过程。通过智能化评价，可以解决或部分解决以往以同行评议为主要评价方法的学者主观判断的弊端。

学术论文智能化评价的相关理论与方法，最初适应于学术论文发表后进入出版传播平台即正式学术交流领域后所产生的学术影响（如基于海量数据从计量指标角度判断学术论文影响力等）。随着大数据技术发展，尤其是知识表示、知识推理、文本识别与分析、知识发现等技术与机器深度学习等人工智能技术的发展，使得计算机智能地对未进入传播领域（或未进入正式学术交流领域）的单篇学术论文（学术手稿或稿件）的质量进行判断与评价成为可能。智能技术运用于学术论文发表前的评价是一个发展的过程，是随着技术与方法进步逐步渗透到学术论文评价的各个流程[16]或主要的内容环节的[17]。评价的因素包括评价的目的、评价的主体、评价的客体、评价的指标体系、评价的标准、评价的模型和评价的结果。

根据智能化技术实现评价的自动化程度或参与传统评审流程的程度，学术论文智能化评价可以分为初期的计算机辅助评价，主要对学术论文的外在指标进行判断和评价；中期的主要依靠计算机进行内容层面的识别与评价，主要对学术论文内容创新性等的评价；成熟期的完全由计算机智能地完成的评价（计算机自动给出主要的评审语）3 个阶段。其中，判断学术论文外在因素的计算机辅助评价的相关技术已较成熟；判断学术论文内容本身的技术正是当前研究的热点，也

是本文主要研究的内容；以计算机自动给出评审语的完全智能化评价是未来的发展方向。

智能化评价过程是使用智能化系统对评价对象进行认识的过程，通过模型的自组织、自学习、自适应、自识别、自协调等功能成为智能化综合评价模型，可以更好地为用户提供决策服务。对学术论文内容进行智能化评价过程包括3个方面：一是内容的智能化识别；二是内容的智能化抽取；三是内容的智能化比对。智能化识别是根据学术论文的内容特征，依据智能识别方法进行识别，如识别研究主题等；智能化抽取是在识别的基础上，根据描述规则进行内容抽取；智能化比对是评价的重要一步，是基于语义相似度计算算法与技术，将识别及抽取的内容进行语义相似度计算和比对，再利用机器学习自动进行特征分类，判断相关内容的新颖性、创新性等。学术论文智能化评价过程如图1所示：

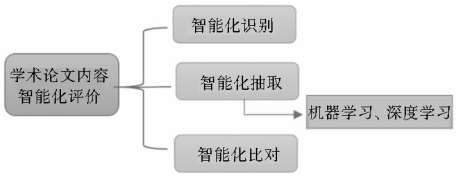

图1　学术论文内容智能化评价过程

3　知识元理论

3.1　知识元概念

知识元是表示、控制、管理和操作知识的基本单元，是为了解决以文献为单位的知识组织方式所包含的知识内容太少而无法满足用户增长的知识需求而逐渐发展起来的[18]。20世纪70年代后期，美国情报学家弗拉基米尔·斯拉麦卡指出将知识的控制单位由文献表层深入到文献内部知识元，文献中的知识元及其链接将产生极大的知识增值，从而提高知识利用和知识创造的效率[19]。英国情报学家B C. Brookes随后也提出利用“认知观点”地图的概念来连接、表征知识内容和知识创造[20]，同时将文献网演变为知识元关联的概念网，使知识体系由外部宏观结构演变为内部微观结构[21]。知识元不仅可以用来表达、存储、检索和利用知识，知识元之间的链接关系还可以用来描述知识的发展脉络，进行知识发现，并预测未来发展方向。知识元概念在不同学科领域和不同时期有不同的表现形态，如教育学领域的知识元是指知识体系的“知识点”，人工智能领域的知识元则指“语义网”，图情领域的知识元则表示文档中的基本概念[22]。温有奎认为知识元是构成知识结构的基元，是知识分解成可独立使用的最小单位，可用来表达一个完整的知识内容或概念，是一组包含了某些知识成分的信息单元集合[23]。根据知识元是基本单元的界定，利用知识元可以有效解决[24]：①知识的自由切分与存取；②知识的自由组织与检索；③知识的自由组合与检索；④知识的准确计量与评价。

3.2　知识元描述与抽取

对知识元的描述有描述模型和描述规则两种。知识元的描述模型[18]是对知识元的语义内容和结构进行揭示的一种抽象表示，是知识元表示的方法，其目的是促进知识元的管理与利用。知识元的描述规则是为了对知识元进行识别和抽取，是根据知识元的描述模型和特征分析而制定或总结的知识元的表示总和。知识元描述模型一般包括属性、内容和关系3个方面，索传军等[18]用语义三元组描述创新知识元，认为每一个创新知识元都可以分解为至少1个主语、谓语、宾语形式，这些语义三元组由于描述的是同一主题下的知识内容，因而存在一定的逻辑关系。袁名依等[25]提出一种基于本体的知识元表示方法。本体是对某一领域中的术语及术语间关系的规范说明，提供对领域知识的共同理解和描述，用于共享、交流和复用，由经过精确定义的概念及概念间的关系组成。其中知识元本体包含Creator、Knowledge Element、Knowledge Element Abstract、Knowledge Element Description和History等5类，Creator用于描述创建者，Knowledge Element用于描述不同的知识单元，Knowledge Element Abstract用于表示知识元抽象体，Knowledge Element Description用于表示知识元描述体，History用于记录知识元的演进发展过程。

从数字资源中抽取知识元是知识元应用的基础，当前学者所提出的方法大致可分为基于文本结构的抽取方法与基于规则的抽取方法两种类型。基于文本结构的抽取方法如姜永常[26]提出的基于物理结构和逻辑结构的抽取方法；周宁等根据事先给定的结构约束来抽取文本片段；方龙等[27]根据学术文本的功能结构进行识别。基于规则的抽取方法如王忠义等[28]提出基于规则的知识元抽取方法，首先建立了概念知识元、事实知识元、数值知识元、方法知识元和关系型知识元的描述规则，并对各知识元的特征词进行详细描述，然后基于描述规则对知识元进行识别和抽取。

本文基于知识元的相关研究，构建学术论文知识元本体，根据知识元本体的描述规则抽取学术论文的知识元，利用知识元包含的语义信息与学术论文的语义信息进行语义相似度计算，根据相似度计算结果，对其创新性进行评价。

4 基于知识元的学术论文创新性评价过程

如上文所述，学术论文内容创新性主要在于新论点和新论据，新论点包括新问题、新理论、新结论，新论据包括新方法和新数据。学术论文成果应在研究问题、理论、方法、结论等微观方面体现其创新性。因此，本文将学术论文内容创新性评价划分为 4 个维度——研究问题创新、理论创新、方法创新及结论创新。

研究问题创新（也可称作研究主题创新或研究选题创新）是指研究者提出一个新的研究问题，或新的研究主题，或新的观点，或新的研究视角，从研究问题可以初步判断研究的价值和创新性。理论创新是指研究者在社会实践活动中，对出现的问题，作出新的理性分析和解答，对认识对象或实践对象的本质、规律和发展变化的趋势作新的揭示和预见，对人类历史经验和现实经验作新的理性升华，是对原有理论体系或框架的新突破，对原有理论的新修正、新发展，对未知领域的新探索。方法创新是指对已有的研究对象提出了新的方法，或对现有的方法进行了改进，或者利用现有的方法解决应用领域中存在的问题。结论创新伴随着研究问题创新、理论创新和方法创新，指在以上创新基础上，获得了与原有成果不同的结果或结论。具体见图 2。

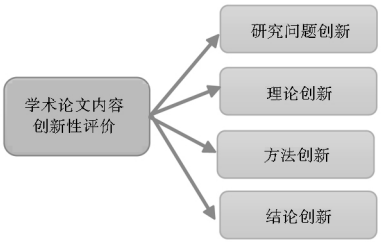

图 2 学术论文内容创新性评价的维度

要进行上述内容的创新性判断，必须首先对目标学术论文中的相关知识进行识别和特征抽取，并与现有学术论文知识库中的内容进行对比分析或相似计算，以判断是否具有创新性。具体步骤如下：

（1）建立知识元本体和知识元描述规则库和术语库：描述反映创新性 4 个维度的知识元的抽取规则，建立描述规则库；对规范术语进行描述，建立术语库。

（2）依据知识元描述规则，对一定时间窗口的已发表学术论文知识库进行知识元抽取，建立知识元本体库，包括研究问题知识元本体库，理论知识元本体库，方法知识元本体库，结论知识元本体库。

（3）建立知识元图谱库（即知识链接网络）。识别一定时间窗口的学术论文中的知识元，并建立每篇学术论文的知识元图谱，标注时间，形成知识元图谱库。

（4）抽取目标学术论文的知识元，建立目标学术论文的知识元图谱。

（5）计算目标学术论文知识元创新性，获得目标学术论文的创新指数。将目标学术论文与知识元本体库和知识元图谱库进行匹配和相似度计算，获得目标学术论文在理论、方法和应用层面的创新指数。

学术论文内容创新性评价的总体流程见图 3。首先对学术论文知识元进行描述，建立知识元抽取规则，形成学术论文知识元抽取规则库，并采用规则库依据知识元本体抽取学术论文中的知识元，构建学术论文的知识元本体，与目标论文中抽取的知识元进行相似度计算，获得创新性评价结果。

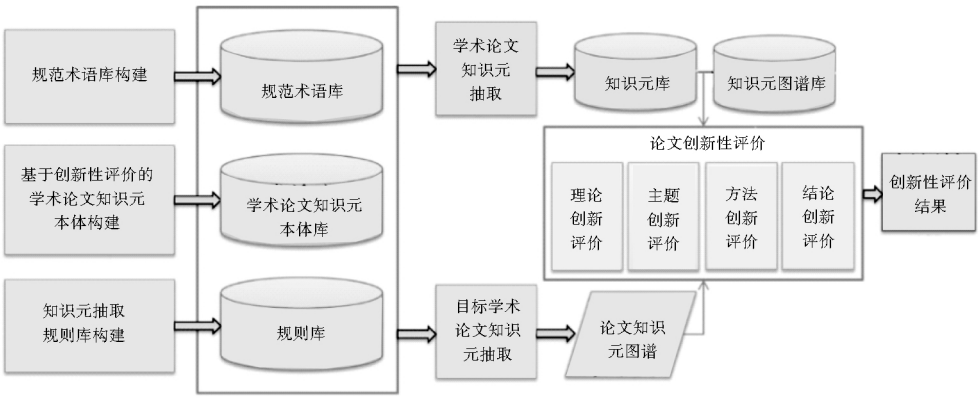

图 3 基于知识元的学术论文创新性评价模型

4.1 学术论文知识元本体

学术论文具有通用的元数据。学术论文基本元数据特征包括篇名、作者、作者单位、摘要、关键词、分类号、DOI、期刊名、发表时间、支持基金、主题、关键词、被引文献等内容。其中，题名、摘要、关键词、主题、分类号、被引文献与论文的正文内容相关。学术论文正文

内容的结构也类似，以《图书情报工作》论文为例，通常包括引言、研究现状或相关研究、理论、方法研究或模型构建、实例或实验、结论、参考文献等基本结构。这些基本结构构成了学术论文的分层体系和各自的功能。本文聚焦创新性评价，主要通过抽取学术论文的研究问题、理论知识点、方法知识点及结果知识点等代表论文核心内容的知识。研究问题属于主题领域，可从学术论文的题名、关键词、摘要、引言部分获得；理论知识点源自学术论文的题名、关键词、摘要、相关理论、理论模型构建、结论等；方法知识点源自学术论文的题名、关键词、摘要、相关理论、方法研究、实例或实验；结果知识点源自学术论文的摘要、实例或实验、结论。在上述分析基础上，构建基于创新性评价的学术论文的知识元描述规则和知识元本体。

4.1.1 学术论文知识元本体总体结构

本文在对知识元、本体模型分析的基础上，构建知识元本体的逻辑结构描述模型。知识元的描述模型目前有三元组、四元组、五元组、六元组模型，尚缺乏统一的标准和框架，本文考虑采用 RDF 格式存储知识元本体，因此选择三元组模型作为知识元本体的逻辑描述模型。RDF 模型是 W3C 提出的用来描述网络资源的标准数据模型，采用主语－谓语－宾语的语义三元组形式描述资源，包括资源的元数据。主语是指被描述的资源，谓语是属性，宾语是属性值。基于以上分析，本文构建了学术论文本体、研究问题本体、理论本体、方法本体和结论本体，各本体之间的关系如图 4 所示：

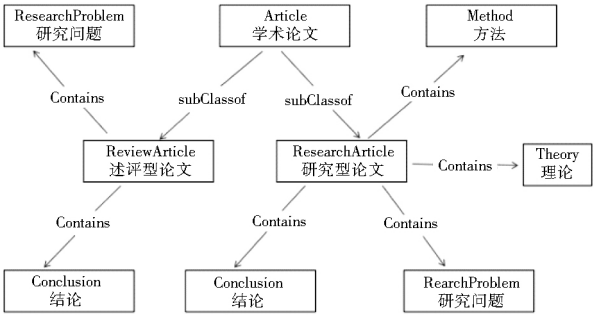

图 4 学术论文本体关系结构

4.1.2 学术论文知识元本体

学术论文属性主要由元数据构成，包括篇名、作者、摘要、关键词、分类号、DOI、期刊名、发表时间、支持基金、主题、被引文献等。学术文献的种类一般包含书籍、报告、会议文章，本文主要针对的是期刊论文，所以并未以此建立学术论文的分类，而是以述评型论文、研究型论文对论文进行分类，并作为学术论文实体的子类。为了提高本体的共享和重用，本文构建的本体继承了 doco 本体、fabio 本体、deo 本体的一些概念。具体概念层次结构如图 5 所示：

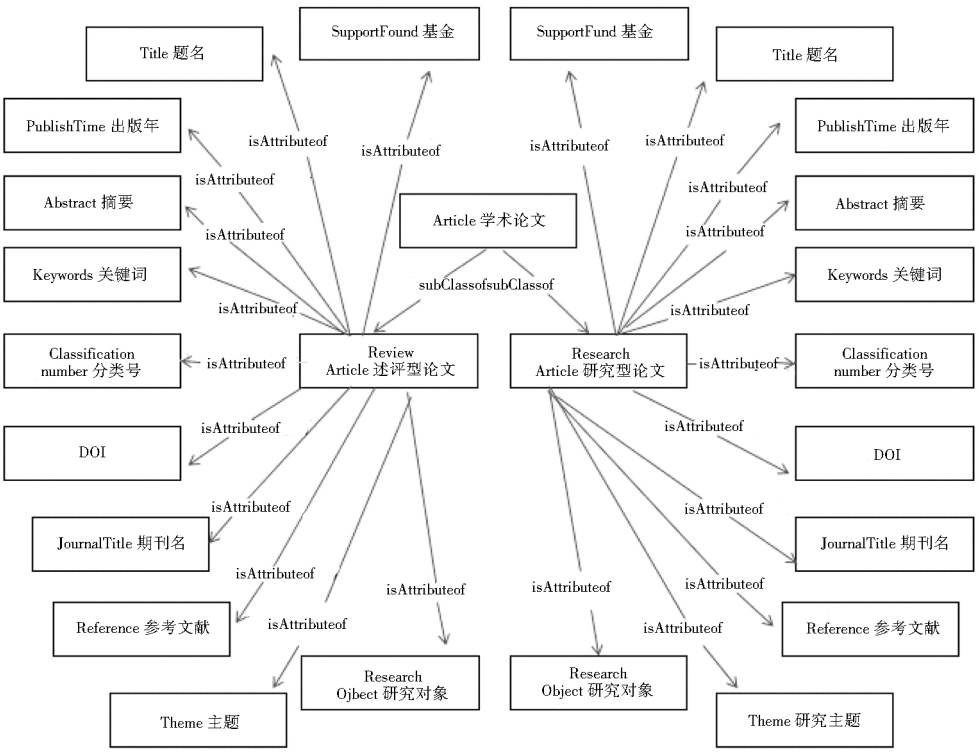

图 5 学术论文知识元本体层次结构

4.1.3 学术论文研究问题知识元本体

研究问题是学术论文研究的基石，没有研究问题，学术研究就失去了研究的意义。学术论文中研究问题是通过研究对象、研究背景、研究目的、研究意义体现的。学术论文研究问题知识元本体层次模型见图 6。

4.1.4 学术论文理论知识元本体

理论创新是研究创新的重要部分，包括思想、学说的创新。学术论文中理论创新的内容主要通过题名特征词、主题词、论文中包含的理论观点、假设模型、框架模型、结论来体现。依据论文的不同结构，将学术理论实体分为理论观点、假设模型和框架模型。根据以上分析，本文构建学术论文理论知识元本体层次模型，如图 7 所示，为理论创新评价的数据准备提供基础。

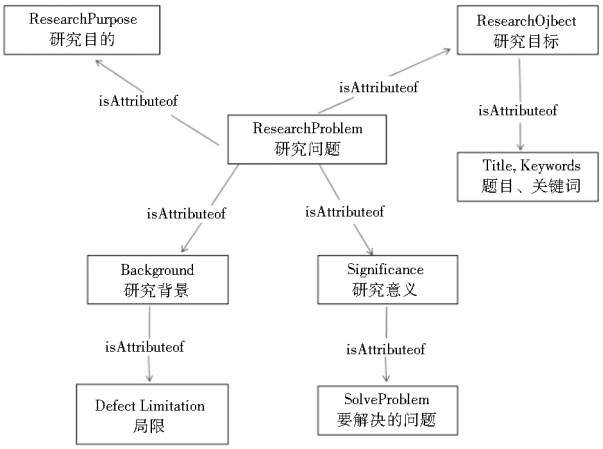

图 6 学术论文研究问题知识元本体层次结构

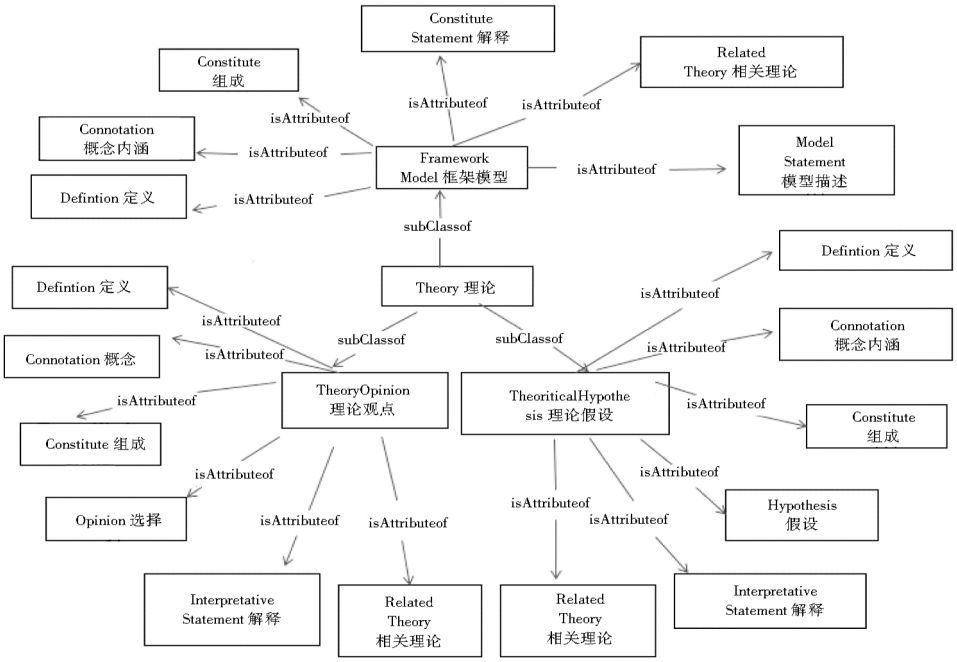

图 7 学术论文理论知识元本体层次结构

4.1.5 学术论文方法知识元本体

学术论文中方法的使用比较复杂，总体可以分为科学研究方法和问题解决方法。科学研究方法主要包括调查问卷法、专家访谈法、案例分析法、观察法、文献研究法、实验等。科学研究方法包括算法、技术方法、评价模型、数学模型等。由于不同的方法具有不同的属性，因此，在方法实体中，将调查问卷法、专家访谈法、案例分析法、观察法、文献研究法、实验、算法、技术方法、评价模型、数学模型等都作为方法的子类，构建了学术方法知识元本体层次结构，见图 8。

4.1.6 学术论文结论知识元本体

论文的结论是论文的重要构成部分，它包括了主要的结论性、观点性、创新性知识。结论的核心要素主要有对策、建议、启示、研究价值、优势、创新点等内容。以此构建学术论文结论知识元本体层次结构见图 9。

4.2 学术论文知识元抽取

4.2.1 知识元抽取规则

通过分析学术知识元抽取需求，剖析学术论文特点，设计知识元抽取方案。学术论文的题名、关键词、摘要是重点抽取对象，其次，学术论文正文具有类似的内容结构，通常由引言、相关理论、研究方法/内容、实验/案例、结论等组成，可依据知识元内容酌情进行抽取。为了抽取过程顺利，需制定一些约束条件：首先，提取过程中不区分英文大小写；其次，在正则表达式设

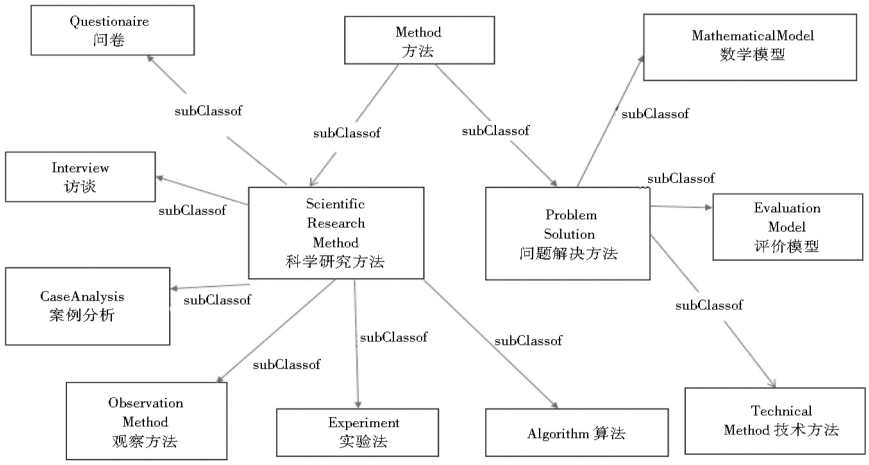

图 8　学术论文方法知识元本体层次结构

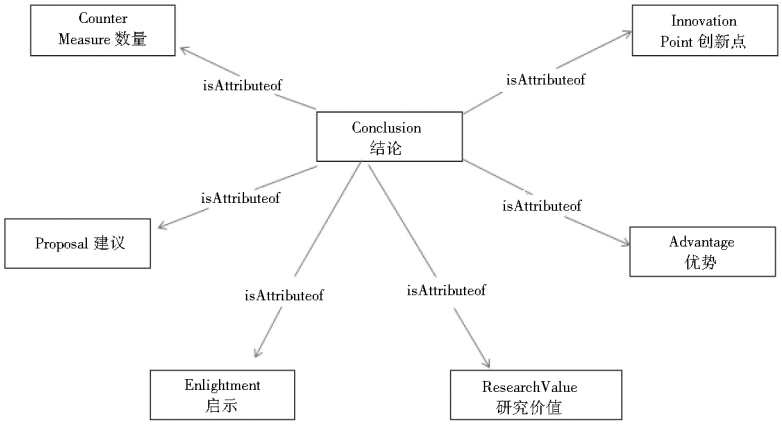

图 9　学术论文结论知识元本体层次结构

计过程中设定长字符串优先；最后，选择贪婪匹配模式，尽可能地获取更多的信息。在进行文本分析之前，首先对文本进行预处理，主要是滤除格式符、换行符等不需要的字符。学术论文文本内容具有很强的规则性，尤其摘要部分，在基于规则的知识元识别过程中结合规范术语库的数据，最后获得论文的知识元。

(1) 学术论文研究问题知识元抽取。识别与抽取目标文本：①篇名与关键词：对篇名进行主题识别，与关键词匹配，确定研究对象和研究内容。②摘要中研究目的、研究意义、应用/实践意义。③引言中本文的研究内容。

学术论文研究问题知识元抽取规则(部分)如下：

题名抽取

关键词抽取

研究/分析(.* ?)领域的(.* ?)的(.* ?)问题

面向/针对(.* ?)领域，研究(.* ?)问题；

基于(.* ?)，研究(.* ?)，解决(.* ?)问题；

提出/阐述(.* ?)解决方案/问题/方法

探讨/分析(.* ?)影响要素/因素

建立(.* ?)流程/体系/程序

对(.* ?)进行研究/评估/可行性分析

(2) 学术论文理论知识元抽取。识别与抽取目标文本：①摘要中提及的运用理论说明。②引言及文献综述中理论不足的相关论述。③相关理论基础部分中的相关理论陈述。④结论中提及的理论贡献。

学术论文理论知识元抽取规则(部分)如下：

提出(.* ?)理论/假设/框架/模型

对(.* ?)理论进行了改进/改善/完善

具有(.* ?)理论意义

对(.* ?)理论进行述评/分析/评价

运用/借鉴(.* ?)理论

(3) 学术论文方法知识元抽取。识别与抽取目标文本：①摘要中提及的运用方法说明。②引言及文献综述中方法不足相关论述。③实验部分中的方法。④结论中提及的方法。

学术论文方法知识元抽取规则(部分)如下：

提出(.* ?)方法/流程/算法/程序/过程

对(.* ?)方法进行了改进/发展/完善
对(.* ?)算法进行了改进/发展/完善
对(.* ?)流程进行了改进/发展/完善
对(.* ?)过程进行了改进/发展/完善
对(.* ?)方法进行述评/分析/评价
运用/借鉴(.* ?)方法
如调查问卷的数量(.* ?)
问卷的信度/效度(.* ?)
查准率/查全率/召回率(.* ?)

(4)学术论文结论知识元抽取。识别与抽取目标文本:①摘要中提及的结果结论。②引言及文献综述中研究目的。③实验部分中的结果。④结论。

学术论文研究结论知识元抽取规则(部分)如下:
提高了(.* ?);
得出(.* ?)结论:(1)(.* ?),(2)(.* ?),(N)(.* ?)
对(.* ?)进行了验证/改进/改善
验证了/证明(.* ?)可行性/有效性/是可行的/是有效的
正确率达(.* ?)
研究发现(.* ?)
结果表明(.* ?)

4.2.2 基于机器学习的理论与方法分类模型

(1)理论与方法分类模型构建。创新性判断是对知识元的创新性进行判断的过程。基于机器学习的理论与方法分类就是让计算机自动发现且充分理解训练集(发表论文知识元)的基本规则和语义,并以计算机可识别的方式表示,进而作为未知文本的判断依据的过程,即计算机自动分类的过程。近年来,基于机器学习的方法进行文本的分类研究很多,尤其是在情感分类研究方面。如杜慧等[29]利用蕴含上下文语义信息的词向量构建文本的特征表示,进而用机器学习的方法对语料进行情感分类;李惠富等[30]将主成分分析、潜在语义分析、Word2Vec 以及 TF-IDF 特征提取方法作为多类型分类器融合的特征提取方法,该方法在各类型语料库中都有很好的表现。

本文在对多种机器学习算法的应用进行考察和思考后最终选择了朴素贝叶斯(Naive Bayes)模型。朴素贝叶斯分类器是机器学习常用的方法之一,是一种有监督的学习算法,其分类鲁棒性好,速度快,尤其适合大数据处理,近几年常被用做文本分类领域。

基于贝叶斯的理论方法分类模型如图 10 所示:

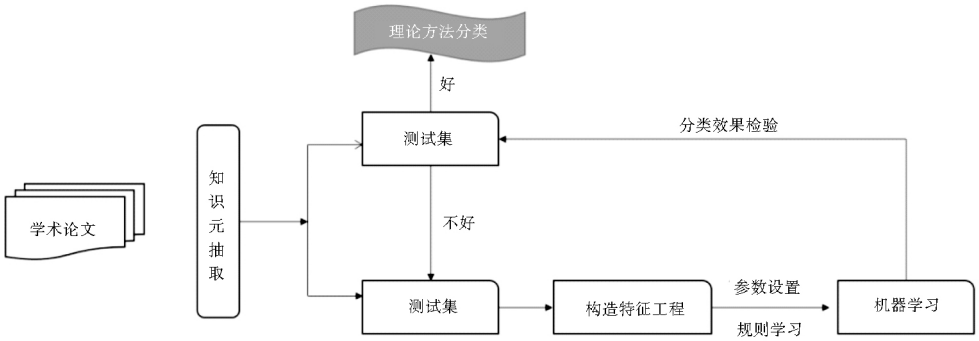

图 10 基于朴素贝叶斯的理论与方法分类模型

(2)理论与方法量化。作者在学术论文中对其理论、方法的表述有一定的规则,基本表现为对这两个方面的描述。通过对大量的学术论文分析发现,这种描述的结构可归纳为“动词+特征词+副词”或“动词+特征词”。例如在关于研究方法的论述中,常用“基于＊＊＊,提出了一种＊＊＊方法。”“对＊＊＊方法进行了改进”。因此评价指标量化就是以论述中的＊＊＊作为特征词,并以其为轴心,定位特征词前后[-u,u]的区间,结合语料库对区间内的动词或副词进行赋值,部分赋值(5 分法)情况见表 1。

4.2.3 规则库构建

基于规则的知识元抽取方法是通过规则与文本内容进行匹配,从而抽取所需要的内容。基于规则的信息抽取方法是一种确定的信息抽取方法,通常通过正则表达式实现,优点是准确率高,缺点是缺乏灵活性。本文通过对论文内容进行分析,构建抽取规则库。具体构建过程见图 11。

主要包括:①依据学术论文知识元本体的实体、分类及属性,对论文内容梳理和分析,筛选包含所需信息内容的完整句子,构成初选集;②利用 SVM 模型对初

表1　理论、方法赋值表(部分)

理论	赋值	方法	赋值
首次提出(.*?)理论	5		
提出(.*?)理论	4	提出(.*?)方法	4
对(.*?)理论进行了改进	3	提出(.*?)流程	4
对(.*?)理论进行了优化	3	提出(.*?)算法	4
对(.*?)理论进行了改善	3	提出(.*?)程序	4
对(.*?)理论进行了完善	3	提出(.*?)过程	4
具有(.*?)理论意义	2	对(.*?)方法进行了改进	3
对(.*?)理论进行述评	2	对(.*?)方法进行了发展	3
对(.*?)理论进行分析	2	对(.*?)算法进行了改进	3
对(.*?)理论进行评价	2	对(.*?)流程进行了改进	3
运用(.*?)理论	1	对(.*?)过程进行了改进	3
借鉴(.*?)理论	1	对(.*?)方法进行了完善	3
		对(.*?)方法进行述评	2
		对(.*?)方法进行分析	2
		对(.*?)方法进行评价	2
		运用(.*?)方法	1
		借鉴(.*?)方法	1

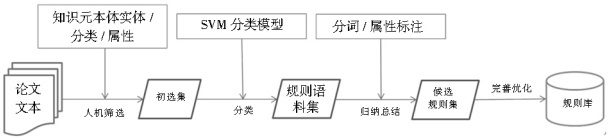

图11　规则库构建过程

选集中的句子进行分类，形成规则语料集；③对规则语料库中的句子进行分词和词性标注，分析句子的结构，例如“……构建一种……模型”，“……提出了一种……方法”，“……构建一个……框架”，判定知识元本体类型，并将这些句子结构进行归纳总结，形成候选规则集；④在候选规则集的基础上，构建规则，采用正则表达式表达规则，并通过不断的信息抽取实验对规则进行优化和完善，构建稳定的规则模板，最终得到规则库。

在知识元抽取过程中，因为方法、理论等具体实例存在一些不规范的名称，为了保证实体命名的一致性，根据领域知识术语库匹配抽取内容，构建了规范术语语料库，以保障理论和方法识别的准确性。规范语料库的构建主要依据中国知网和全国科学技术名词审定委员会合作项目《中国规范术语库》、中国知网《方法库》，在此基础上，补充学术论文中的不规范实体名称，形成统一的方法、理论命名实体语料库。

4.3　学术论文创新性智能化评价过程

在学术论文知识元库构建的基础上，通过抽取论文知识元(经过训练后)与现有的学术论文知识元比较，获取论文创新性评价的基本数据，提出论文研究问题创新性、理论创新性、方法创新性、结论创新性评价的基本方法。

具体步骤如下：

第一步：学术论文知识元抽取。首先对文本进行预处理，滤除不需要的字符；在基于规则的知识元识别过程中结合规范术语库的数据，最后获得学术论文知识元。

第二步：进行数值比较。学术论文知识元包含数值和文本。数值知识元主要包含方法知识元，如调查问卷的数量、问卷的信度和效度等，主要涉及到论文的科学性问题；结论知识元，如查准率、查全率，涉及到论文结论的创新性评价。本文主要是对论文创新性的评价，因此，选取结论数值，根据其具体定义，比较大小，确定论文结论的创新性。

第三步：文本相似度计算。知识元的类型是文本时，需要判断文本的相似度，本文主要采用词向量的方法。Word2vec是产生词向量的模型，使用该模型将使每个词语都获得一个相对应的词向量，通过计算词向量的余弦值获得两个词的相似度值。引入词向量之后，可以识别两个字型不同但是相关或疑似相近的词语，能够弥补传统文本相似度算法的不足。本文采用中文维基百科语料训练词向量。

第四步：目标学术论文创新性评价。

本文将依据文本相似度和数值比较结果计算论文创新性结果。首先，计算研究问题创新评价结果；其次，对理论创新进行评价；再次，评价方法创新；最后，对结论创新成果进行评价。

构建的学术论文创新性智能化评价过程见图12。

4.4　学术论文创新性智能化评价实证检验

基于上述研究过程，本文利用python语言，采用python中的NLP工具包，结合python中的Flask框架，对学术论文创新性4个维度智能化评价进行实证检验。

4.4.1　数据集获取

实验数据集主要以《图书情报工作》2015－2018年4年的投稿论文和部分图书馆学情报学核心期刊2015－2017年已发表的学术论文数据组成(约6千余条数据)。考虑到实验的便捷性，本次实验仅获取论文题名、中文摘要和关键词。部分摘要直接录入为结构化摘要，如将摘要分为【目的/意义】、【过程/方法】、【结果/结论】，以提升计算机识别的有效性。

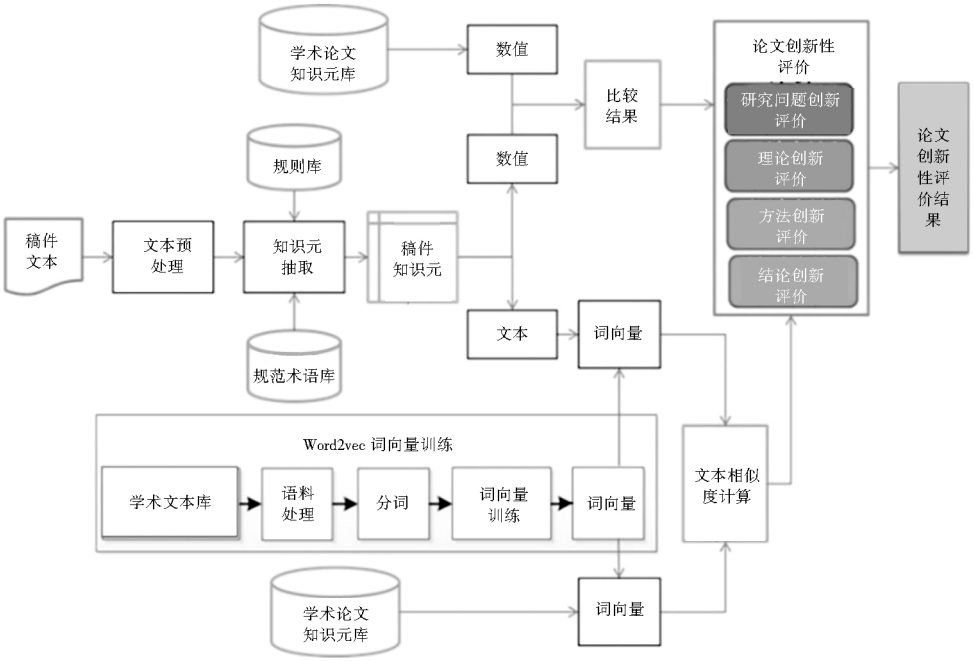

图 12 学术论文创新性智能化评价过程

4.4.2 实验模块组成及功能

实验模块由数据集管理模块、抽取规则管理模块、创新性评价模块、数据集训练模块、论文综合评分模块构成。

数据集管理模块: 该模块主要对数据库中的数据集进行维护。在这里主要指已发表的学术论文知识库。

抽取规则管理模块: 该模块主要是对理论维度和方法维度的评分规则进行设置,基于设置的规则计算这两个维度的评分分数。

创新性评价模块: 该模块主要是基于系统中的数据集,结合系统设置的抽取规则,从 4 个维度得出创新性综合评分。

数据集训练模块: 数据集训练主要采用机器学习的方法,结合创新性评价指标,构建训练模型,对数据集进行评分。

论文综合评分模块: 对系统中已发表的论文和新提交的论文,结合数据集训练的模型,对学术论文的创新性进行综合评分。然后以该综合评分为依据,通过机器学习对该学术论文是否发表进行评价。

目前,各实验模块可以完成的功能有: ①基础数据集的管理与维护,可以对基础数据进行添加、修改、删除管理。②自定义评分规则,包括评分规则描述、评分的分数、评分的优先级。③可以根据自定义的评分规则,基于正则表达式匹配的方式,对理论维度和方法维度进行评分,并显示匹配的评分规则。④根据语义相似度计算,可单独计算出研究问题维度与结论维度的创新性评分。⑤研究问题维度与结论维度的创新性评分可以针对不同年份的数据集进行,可以显示出与之对应的语义相似度的值,以及实验数据的分析与实验结果。⑥结合 4 个维度及系数,可以计算出论文的综合评分。⑦结合论文创新性评分,程序将生成论文的创新性分布,根据设定的阈值给出采纳或不采纳的评价结果。

4.4.3 部分实验结果及分析

随机选取 2016、2017、2018 年发表在《图书情报工作》的论文(目前库中所拥有的论文为 2015 – 2018 年) 和发表在其他期刊的论文进行创新性评分计算,部分评分结果见表 2。

从表 2 的部分随机结果来看,多数论文的创新性随着计算年的增加而变小,这符合创新性扩散的一般规律。随着数据量加大、学习与训练次数增多,创新性维度的判断将更加精确。

4.5 结论

创新性是学术论文是否录用的重要标准,发现论文在论点(即研究问题) 、理论、论据/数据、方法、结论、价值等方面的重要创新点或重要贡献是判断学术论文内容是否具备创新性的依据。本文以知识元研究为基础,在学术论文内容分析的基础上,构建了反映学术论文内容创新性的 4 个维度的学术论文知识元本体

表2　随机判断的部分论文的得分情况

编号	发表年	论文来源	对比年		
		是否数据库中论文	2015	2015－2016	2015－2017
1	2016	是	1.448	1.210	－
2	2016	是	1.694	0	－
3	2016	是	1.686	1.686	－
4	2016	是	1.448	1.204	－
5	2017	是	1.448	1.211	1.211
6	2017	是	1.694	1.694	1.694
7	2017	是	1.694	1.694	1.694
8	2017	是	1.448	1.448	0.49
9	2018	是	1.694	1.204	1.204
10	2018	是	1.448	1.204	0.966
11	2018	是	1.932	1.932	1.686
12	2018	是	1.694	1.448	1.448
13	2018	是	0.966	0.483	0.483
14	2018	否，发表在他刊	1.448	1.448	1.448
15	2018	否，发表在他刊	1.694	1.448	1.448
16	2018	否，发表在他刊	1.448	1.210	0.973
17	2018	否，发表在他刊	1.694	1.210	1.210
18	2018	否，发表在他刊	1.932	1.694	1.694
……					

模型，确定了4个维度的知识元抽取规则，利用Word2Vec和朴素贝叶斯方法对学术论文理论与方法的创新性进行分类，并采用SVM模型构建知识元抽取规则库。在学术论文知识元库构建基础上，提出学术论文研究问题创新性、理论创新性、方法创新性、结论创新性智能化评价的基本方法，构建学术论文创新性智能化评价过程。

最后，本文以《图书情报工作》2015－2017年发表的学术论文为实验数据库，依照抽取规则对这些学术论文的知识元进行抽取，对理论与方法的知识元进行机器学习分类，使得理论与方法知识元成为自带权重的知识元类别。对抽取的4个维度知识元进行进一步词向量训练，建立语料库。以2018、2017、2016年的学术论文为试验数据，对其创新性进行识别与判断，最后的评分结果具有一定的可行性，基本上反映了论文创新扩散的过程，即创新性递减的过程。

通过对评分结果的进一步分析，发现评分系统在评分过程中存在一些问题，如由于理论与方法的规则设置较为严格，部分论文的方法创新性得分为0，需要进一步调整计算的方法。在研究结论的创新性计算方面，也需要做进一步调整，结果结论元数据抽取规则也需要进一步完善，以便获得更好的结果。

参考文献：

[1] 陈建青. 对我国学术论文创新性评审的几点思考[J]. 青年记者，2013(18)：33－35.

[2] 王文彦. 论创新的层次性[J]. 河南师范大学学报(哲学社会科学版)，2006(1)：218－219.

[3] 逯万辉，谭宗颖. 学术成果主题新颖性测度方法研究——基于Doc2Vec和HMM算法[J]. 数据分析与知识发现，2018(3)：22－29.

[4] UZZI B，MUKHERJEE S，STRINGER M，et al. Atypical combinations and scientific impact[J]. Science，2013，342(6157)：468－472.

[5] 李冲，苏永建. 学术评价：量化模式的反思与超越[J]. 自然辩证法研究，2017，33(2)：59－63.

[6] SELVARAJOO K. Measuring merit：take the risk[J]. Science，2015，347：139－140.

[7] 彭琳，杜杏叶. 学术期刊开放式同行评议实施调查[J]. 中国科技期刊研究，2018，29(11)：1114－1121.

[8] 贺颖. 同行评议专家遴选问题研究[M]. 北京：中国社会科学出版社，2016.

[9] VIEIRA E S，GOMES J A N F. The peer-review process：the most valued dimensions according to the researcher's scientific career[J]. Research evaluation，2018，27(3)：246－261.

[10] 沈阳. 一种基于关键词的创新度评价方法[J]. 情报理论与实践，2007(1)：125－127.

[11] 贺婉莹. 基于机器学习的论文学术创新力评价研究[D]. 南京：南京大学，2019.

[12] 索传军. 知识转移视角下的学术论文老化与创新研究[J]. 图书情报工作，2014，58(5)：5－12.

[13] 杨京，王芳，白如江. 一种基于研究主题对比的单篇学术论文创新力评价方法[J]. 图书情报工作，2018，62(17)：75－83.

[14] 阮光册，夏磊. 基于Doc2Vec的期刊论文热点选题识别[J]. 情报理论与实践，2019，42(4)：107－111，106.

[15] 邱均平. 评价学：理论·方法·实践[M]. 北京：科学出版社，2010.

[16] 邱均平，赵岩杰，罗力. 科学评价中的论文分类方法研究[J]. 情报学报，2011(5)：554－560.

[17] 徐建强，崔慧洁，李小平，等. 一种新型学位论文智能评价系统[J]. 计算机工程，2013，39(7)：224－227，232.

[18] 索传军，盖双双. 知识元的内涵、结构与描述模型研究[J]. 中国图书馆学报，2018，44(4)：54－72.

[19] 温有奎，徐国华，赖伯年，等. 知识元挖掘[M]. 西安：西安电子科技大学出版社，2005.

[20] BROOKES B C. The developing cognitive viewpoint in information science[C]// International workshop on the cognitive viewpoint. Ghent：Ghent University，1977：195－203.

[21] BROOKES B C. The foundations of information science part I. Philosophical aspects[J]. Journal of information science，1980，2(3/4)：125－133.

[22] 温有奎，焦玉英. 基于知识元的知识发现[M]. 西安：西安电子

科技大学出版社，2011.
[23] 温有奎，徐国华. 知识元链接理论[J]. 情报学报，2003，22(6)：665－670.
[24] 文庭孝，侯经川，龚蛟腾，等. 中文文本知识元的构建及其现实意义[J]. 中国图书馆学报，2007，33(6)：91－95.
[25] 袁名依，谢深泉. 基于知识元本体的知识统一表示[J]. 现代计算机(专业版)，2008(5)：46－48，57.
[26] 姜永常，杨宏岩，张丽波. 基于知识元的知识组织及其系统服务功能研究[J]. 情报理论与实践，2007(1)：37－40.
[27] 方龙，李信，黄永，等. 学术文本的结构功能识别——在关键词自动抽取中的应用[J]. 情报学报，2017，36(6)：599－605.
[28] 王忠义，沈雪莹，黄京. 基于知识元的中文文本层级分割[J]. 图书情报工作，2019，63(7)：105－115.
[29] 杜慧，徐学可，伍大勇，等. 基于情感词向量的微博情感分类[J]. 中文信息学报，2017，31(3)：170－176.
[30] 李惠富，陆光. 多类型分类器融合的文本分类方法研究[J]. 计算机应用研究，2019，36(3)：752－755.

作者贡献说明：
李贺：负责论文选题、框架及修改、定稿；
杜杏叶：负责论文初稿写作与修改。

Research on Intelligent Evaluation for the Content Innovation of Academic Papers

Li He[1] Du Xingye[1,2,3]

[1] School of Management Jilin University, Changchun 130022

[2] National Science Library, Chinese Academy of Sciences, Beijing 100190

[3] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/significance] Innovation is the key factor of academic paper evaluation. Based on the knowledge element theory and machine learning theory and algorithm, this paper studies how to intelligently evaluate the innovation of academic papers from the content of paper. [Method/process] Firstly, we constructed 4 knowledge element ontologies of academic papers including 'research problem ontology', 'theory ontology', 'method ontology' and 'conclusion ontology', and proposed the model of innovation evaluation. Secondly, we put forward the rules of knowledge element extraction. Word2vec and naive Bayes were used to classify the innovation of theories and methods of academic papers, and SVM model was used to build the rule base of knowledge element extraction. At last, on the basis of the construction of knowledge Meta base of academic papers, we proposed the basic methods of intelligent evaluation of research questions, theories, methods and conclusions of academic papers. We also constructed the process of intelligent evaluation of innovation of academic papers. [Result/conclusion] The feasibility of the methods is verified by the experiment and could provide the references for the realization of intelligently evaluation of academic paper.

Keywords: academic papers evaluation knowledge element content innovation intelligent evaluation