·科学论坛·

地理学学科语义网及其在科学基金智能辅助指派中的应用

冷疏影[1]* 俞肇元[2] 胡 勇[2] 郑袁明[1] 宋长青[1]

(1 国家自然科学基金委员会地球科学部,北京 100085; 2 南京师范大学,南京 210023;

[摘 要] 本文介绍了地球科学部一处基于学科语义网的科学基金智能辅助指派工作的阶段性成果,分析了学科语义网的构建原则与策略、基本环节、技术实现途径、质量控制效果、以及未来优化策略,最后总结了基于学科语义网的辅助指派研究带来的启示。

[关键词] 地理学;语义网;科学基金;智能辅助指派

DOI:10.16262/j.cnki.1000-8217.2015.01.014

在国家自然科学基金领域交叉广泛、知识更新迅速、领域评审专家规模日益增长的新形势下,为提升基金项目分组评审与专家遴选匹配的指向性与智能化水平,更好地体现科学基金评审过程的科学性、专业性和高效性,国家自然科学基金委员会(简称自然科学基金委)地球科学部一处(简称地学一处,同地理学科)自 2013 年起开始探索在学科关键词体系基础上构建学科语义网,以研究方向、关键词的规范表达实现申请书、函评专家特征信息的智能提取,并实现申请书与专家信息的学科语义标定、分组聚类与智能匹配。经过 2014 年的评审实践,基于学科语义网的函评专家指派的优势已初步显现。

1 学科语义网智能辅助指派的背景与需求

地学一处自 2011 年起在国家自然科学基金申报及评审中全程引入地理学学科方向分类与关键词体系进行辅助指派[1],进而基于学科关键词体系,在申请书信息规范和专家信息准确收集两个方面进行地理学科基金智能辅助指派系统的构建[2]。经过 3 年(2012—2014)的试点工作,学科关键词体系对基金申请与辅助信息收集起到了重要的规范化作用。申请书方面,80%以上申请书选择了学科提供的关键词,70%以上学科关键词被使用。专家信息方面,已建立了以学科-信息中心、学科-专家互动式为主的、涵盖了专家主持项目、主要研究方向、研究关键词、发表成果及历史评议的专家综合信息库,并建立了地理学科同行评议专家推荐指标体系[3]。学科关键词体系在国家自然科学基金评审全流程中的规范化应用在提升指派效率的同时,也为基于机器学习的智能辅助指派提供了基础。

地理学科研究领域的综合性、交叉性与复杂性使得单纯基于学科关键词的辅助指派难以满足当前基金评审对专业性、交叉性以及准确性等方面的要求。主要表现在:

(1) 单纯的地理学科关键词体系难以全面涵盖地理学所有研究领域。近年来,仍有 20%左右的申请书采用了非学科体系中的自填关键词。申请书中未选择具体研究方向而选择"其他方向"的申请书也占 20%左右。

(2) 申请数量和评议专家队伍庞大,相关信息更新耗时,难度日益增大。如 ISISN 系统专家库中大量文献成果信息难以得到及时补充,散布于互联网及文献资料中的同行评议专家信息难以汇总等。

(3) 对精细化的申请书分组指派支撑仍存不足。随着学科交叉的日益增强,同一申请书往往可能跨越多个学科领域与研究方向,导致基金申请时方向选择相对困难,也可能出现研究内容差异较大的申请书选择了类似的关键词。由于地学研究的日益专业化,缺乏语义关联的关键词体系难以满足科研基金的精细化分组和指派需求。

收稿日期:2014-09-03;修回日期:2014-10-09

* 通信作者,Email:lengsy@nsfc.gov.cn

(4) 智能化处理能力仍相对较弱。传统的匹配和指派方法基于词汇和统计规律进行，关键词组合只考虑出现频度的引导作用，未提供以学科关键词为主题牵引的词汇之间的“语义”关系。而词汇学科含义具有中英文对照、同词多义、领域差异及层级关联等问题，导致直接基于学科关键词的辅助指派在智能性与准确性上仍存瓶颈，所获取的知识的可重用性也相对较差。

语义网是在传统互联网基础上，通过对相关信息进行语义标定，利用不同数据之间的语义关联来实现机器对知识的理解、管理、检索、处理与分析。传统意义上的语义网构建一般需要经过海量信息的资源描述、语义关系标注、数据/知识/信息的抽象与分类体系建立、领域本体构造表达、服务发现与应用等步骤，并通过 Web 服务与 RDF、XML、OWL 等技术体系与规范加以组织和集成[4]。其中准确的分类体系、语义关系标注和领域本体构建是语义网质量的关键。一个可用、规范的语义网往往需要经过多年持续的人、物力投入，并需要经过长时间的优化。当前语义网多用于分类体系完备、领域边界明确且已拥有大量信息化数据支撑的知识管理系统上，部分应用于商业领域。受制于领域本体构建和语义标注的复杂性，具有实用价值的学科语义网仍不多见。尤其是以宏观性、综合性、交叉性为特色的地理学科，很难在当前技术条件下构建本体结构完善、语义标注完备的学科语义网。

面向申请书智能辅助指派的目标，地学一处于 2013 年开始，尝试在原有学科关键词与智能辅助指派系统的基础上，通过构建学科语义网建立多源词汇与学科关键词体系之间的准确联系，建立申请书及专家信息中非结构化文本信息间确切的语义关联，利用词汇间的语义关联描述语义的领域概念组合并进行层次搜索。基于专家知识融合多元文献与基金申请书信息，利用机器学习模型梳理多元信息之间的语义关联，以多元信息校正申请书、专家信息中对关键词描绘的不足，实现知识空间中申请书与专家的一致性处理与匹配，为未能准确选择研究方向的申请书提供合适的研究方向标定。在此基础上进行申请书和专家信息的研究领域与方向标定，实现对申请书快速分类、批量浏览、分组与自动专家匹配，进而进行自然科学基金的趋势分析，实现领域知识的机器识别与可重复利用。

2　学科语义网构建的整体思路与关键问题

学科语义网构建的主要目的是建立已有的多源信息(申请书、专家信息、互联网信息)与已有学科体系之间的语义关联，通过不同词汇之间的语义关系实现对申请书的方向标定、智能分组与同行评议专家指派。学科语义网构建必须满足专业性、规范性、高效性、完备性以及可支撑分析等原则，要求具备有充分数量的专业语料库、具有明确研究方向表征的词汇语义关联、准确高效的存储与检索方法以及可支撑多源信息文本(申请书、专家信息)研究方向标定、分组聚类及同行评议专家智能指派的综合性分析方法。学科语义网的构建及应用的主要流程包括基于多源信息的语料库及语义词典构建、语义网构建与处理、申请书及专家信息的专业分词与停词、语义规范化处理与语义网检索、研究方向标定与匹配等。技术路线见图 1。

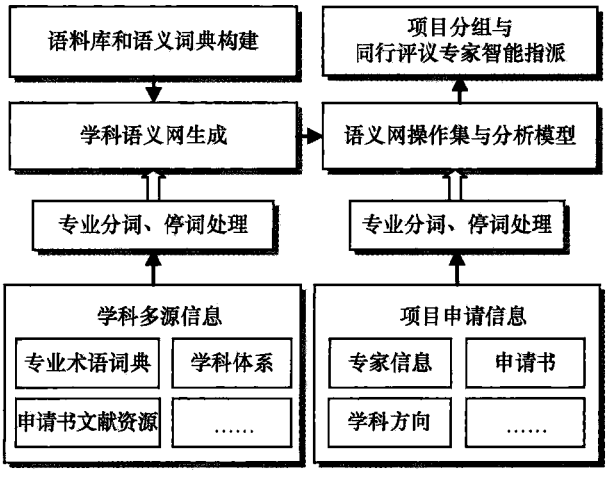

图 1　学科语义网构建的技术路线

学科语义网构建及应用的关键问题如下：

(1) 语料库和语义词典构建是学科语义网生成的基础，也是智能指派能否实现的基准，但需要解决语料库词汇的专业性、关联性、全面性、指称性和唯一性问题，以及厘清语料库词汇之间的关联共现关系、语义表征关系和层级隶属关系[4]。

(2) 申请书及专家的非结构化文本分词处理是实现其语义分析，智能匹配的前提。分词效果的好坏对后续的方向标定等影响显著，关键在于如何构建具有专业指向性和学科属性特征的专业停词方法，并利用同义词表实现对词汇表征的规范化处理。

(3) 语义网的高效使用及基于此的申请书与专

家信息相似度匹配，核心是解决数据存储与检索效率，以及不同研究方向概率分布的修正与标定问题。

(4) 对于项目分组与同行评议专家的智能指派，主要是探索语义网和研究方向共同约束下专家研究方向信息的自动解译与标定，及实现智能指派的专家推荐问题。

3 学科语义网实现与应用

3.1 基于多源信息集成的语料库及语义词典构建

CNKI 的主题词可以实现对文献标题、关键词、摘要进行综合检索，具有全面性、模糊性和概括性的特点。文献记录自身的专业性保证了所检索词汇的全面性、专业性和完备性，有助于在短时间内构建高质量的语料库，并可支撑语料库的持续更新。从专业术语词典、历年申请书以及文献资料中进行专业词汇初选，建立初始的学科语料库。以地学一处编制的《地理学学科方向分类与关键词(修订版 2014)》中 5902 个关键词为主题词，检索 CNKI 学术文献数据库中 1999—2013 年与地理学密切相关的 132 种学术期刊，建立学科关键词与检索文献论文关键词的语义关联，并进行规范化处理。在此基础上通过整合申请书及信息系统中专家自填关键词，得到规范的关键词对 219 万多对，224 个研究方向均具有丰富的关联词汇。

3.2 研究方向约束的专业分词、停词处理

稳健、高效的分词算法可以有效解决分词歧义，基于词汇的专业含义构建高质量的分词词典和同义词典则可强化申请书的研究方向指征。以语义网节点词汇作为分词词典，采用最长匹配与概率权重叠加多源冗余分词的策略实现对申请书、专家信息的分词及标志性主导关键词的提取，并采用同义词表进行分词结果的规范化处理。经过几轮的算法调优，分词词典及停词词典优化，最终形成包含 247 319 词的关键词分词词典，经优化后的分词处理效率约为 6—10 项/分钟，可满足基金项目处理的需求。为进一步提升词汇的学科边界，还构建了面向 224 个研究方向的专业停词表。通过细致核实专家对不同领域的分词结果，构建具有强针对性的专业停词表可以很好地保证语义网的专业性、准确性和针对性，并可起到了约束学科边界，凸显学科特征的作用，且所构建的停词表和停词方法在后一年新申请书分词处理过程中均可直接复用。利用列表 Hash 编码和向量化处理技术处理后停词效率约为 13 450 词/分钟。

3.3 基于学科语义网的词汇检索与研究方向标定

基于 JAVA 开发了地理学科关键词语义网络管理与分析系统，实现了语义网的管理，编辑，统计参数计算与可视化等基本功能。采用稀疏矩阵及向量化运算策略实现词汇序列的快速检索。通过对语义网统计特征参数的计算获取语义网词汇之间的群聚、关联、共现等关系，实现对语义网词汇的快速检索、筛选和层次构造。由于直接基于词汇的相似性度量可能导致不同研究方向的申请书出现高相似度的混淆情况，我们基于向量空间模型从词汇、语料及研究方向三个层次构建了申请书语料的语义相似性度量模型，利用语义网约束研究方向的边界，从而实现研究方向相似度的计算。针对可能存在的语义歧义情况，对已有部分样本进行人工标定，而后利用 Dynamic Text Regression 学习模型[5]进行研究方向校正。

3.4 同行评议专家信息处理与智能指派

通过对原始申请书及专家信息数据梳理构建相关语料的词频矩阵，基于该词频矩阵，利用语义网进行研究方向匹配、标定与关联后构建方向打分矩阵，计算出申请书/专家之间的相似性程度，实现分组、匹配关系。在此基础上，我们还收集 6 年的专家函评信息和 4 年的专家拒绝指派评议(简称“拒指”)信息，设计了函评意见回复实时性、评议人评审结果差异性、函评意见平均长度、拒指率、拒指实时性等评价指标，构建“历史评议信息权重”的评估模型，以关键词匹配度与历史评议信息权重的乘积得到最终的综合专家推荐度。考虑与 ISISN 系统的衔接以及每一位评审专家评审项目数量的限制，采用对申请书分组匹配专家的方式。对于 20%左右申请“其他方向”的交叉项目，在 1 级申请代码权限(D01)下对此类项目一并处理，然后将产生分组匹配专家结果或未参与分组匹配专家的项目一并反馈 2 级申请代码(如 D0101)或 3 级申请代码(如 D010101)管理权限下。而对其余 80%填写了具体研究方向的申请书，在每个 2 级申请代码下，通过对其研究方向打分矩阵进行 Spherical K-Means 聚类获得不同项目的分组，对具有严格分组的项目再进行专家匹配，实现基于分组的同行评议专家智能指派。

4 学科语义网辅助智能指派的效果评估

学科语义网在申请书项目分组及同行评议专家

匹配方面无疑具有独到优势，极大地提高了学科工作人员同行评议专家指派效率，在此不做过多阐述。以下将主要针对学科语义网在智能辅助指派中的关键作用，即申请书研究方向标定进行评估分析。

以 2012—2013 年度申请书中填写了具体研究方向的申请书为训练样本（10 875 份），2014 年的申请书为测试样本（3622 份），进行学科语义网研究方向标定测试。分别计算语义网标定的最主导研究方向与学科工作人员人工核实的研究方向之间的一致比例（第一方向正确率）以及语义网标定的前三个主导研究方向中包含学科工作人员人工核实方向的比例（前三方向正确率），依次作为学科语义网研究方向标定效果的评估指标。其中 2012—2013 年的训练数据的第一方向与前三方向正确率分别是 51.69% 和 72.70%。2014 年测试数据中，第一方向和前三方向正确率分别为 40.69% 和 62.87%。如考虑将语义网标定方向与原申请书自填方向进行对比，2014 年测试数据的第一方向和前三方向正确率则上升至 53.31% 和 73.33%。

由于不同分支学科的学习样本及语义网词汇存在偏差，学科语义网对不同研究方向的正确率标定也存在偏差。为考察学科语义网对不同研究方向标定的偏差程度，对 2012—2014 年地学一处 224 个研究方向的申请项目分别计算第一和前三方向正确率。其中前三方向正确率超过 75% 的研究方向共 85 个，大部分研究领域方向标定的正确率均相对较高。由于不同研究方向上的申请数量差异较大，为避免样本量影响，以学科工作人员人工标定方向为基准，选取申请项目数超过 128 项的 30 个研究方向的申请书样本，统计各研究方向正确率分布（图 2(a)、(b)）。

除旅游地理、全球环境变化及其影响、资源管理与可持续发展三个方向外，其余项目的前三方向正确率均超过 70%。其中，第一方向正确率较高的研究方向的学科边界明确，学科体系较为成熟，关键词特征属性明显，基于语义网可以直接进行研究方向的准确判定。而正确率较低的研究方向大致存在两类，一类学科综合性很强，关键词特征属性和学科意义相对不明显（如：旅游地理、资源管理和可持续发展等）。另一类为学科交叉性强，研究对象多样且关键词特征属性多维（如：重金属污染与修复、污染物环境行为等）。

由于机器学习模型对学科领域的判定是硬性的、排他的，对于上述两类情况要求有更多具有学科指征的关键词加以标定和约束。考虑到地理学科的综合性和宏观性，可以在机器学习模型中引入软约束规则或利用模糊分类，并通过样本量的积累，持续优化学科语义网及其中的机器学习模型，不断提升研究方向标定的正确率。

对申请书按研究方向的相似性进行聚类和分组指派可以更好的挖掘申请书的隐藏及共性信息，提高专家评审结果的可比性，降低拒评率。基于学科语义网对 2014 年申请书进行聚类分组，依据 2 级申请代码＋第一研究方向＋第一关键词的分组原则可将申请书分为 2029 组。经过学科工作人员校验、确认和调整分组后，得到最终指派分组 1181 组。主要的调整包括申请书研究方向组别调整与分组合并。其中调整组别为将分组中研究方向差异明显的个别项目移动至更适合的分组。而分组合并则是将方向类似的分组合并成一组以便进行项目指派。主要的研究方向组别调整和分组合并的统计见图 2(c)、(d)。

总体上，不同方向上分组调整/合并的比例多不超过项目总体数量的 1/3，显示了语义网分组与最终指派结果之间具有较好的对应性。其中组别调整较多的研究方向中，旅游地理、污染物环境行为、城市空间、生态遥感等方向的语义网标定正确率也相对较低，主要通过人工校验纠正研究方向标定的偏差；而分组合并的研究方向主要发生在进一步精细划分了的研究方向（如污染物环境行为、地理信息系统建模、植物地理、区域发展等）。这显示出学科语义网可以揭示申请书之间细微的差异，但也导致不同申请书之间较小的学科差异被识别为不同的组别，在敏感性上仍可进一步优化。总体上，基于关键词标定的聚类结果有效约束了学科的边界，不同研究方向上申请书区分明显，很好地标定了整个聚类组的研究方向。但受语义网研究方向标定的粒度、精度以及指派过程中专家申请项目限制等因素，一定程度上仍需学科工作人员进行适当的手工调整。考虑到分组调整和分组合并的研究方向具有一定的重叠，在现有成果基础上通过持续优化语义网提升基金项目分组的合理性与有效性是可行的。

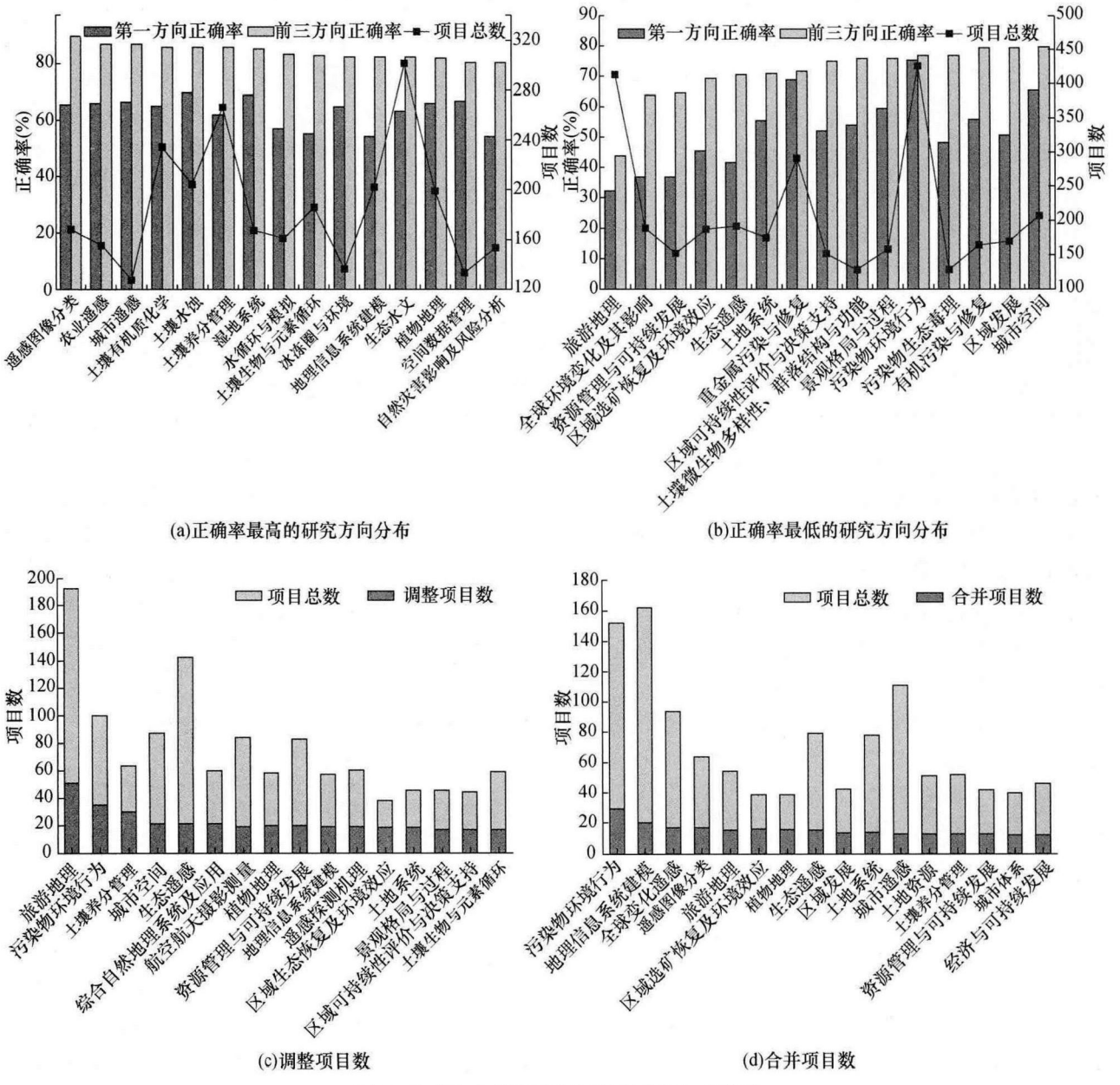

图 2 研究方向标定结果评价及分组调整统计

5 结论与展望

当前自然科学基金学科领域交叉广泛，选题更新迅速。近年来，国家自然科学基金委在信息科学部、医学科学部及地球科学部等进行了智能辅助指派的试点工作，从多种途径探索了学科体系建立、关键词与研究方向关联、专家分组匹配以及辅助指派系统设计等理论与方法问题。受学科体系及学科边界的影响，不同学科进行智能辅助指派的实施途径和实现方法的确存在差异。信息化方法对学科细分明确、学科边界明确的领域处理效果相对较好，但对于像地理学这类研究内容广泛、学科交叉明显的综合性学科，仍需要通过多角度的整合领域专家知识，总结相关专业词汇的学科语义特性，不断提升智能辅助指派的水平。

学科语义网是实现无层级、表述多样的多源、开放的申请书和专家信息与具有层级结构、表述相对规范且稳定的学科体系对接的中间媒介。语义网可以发挥研究方向控制下的学科关键词的中心作用，实现对申请书和专家主要研究特征的智能提取及标定。以学科关键词体系为基础的学科语义网，是实现学科知识支撑与持续更新、申请书学科语义特征深度挖掘、以及研究方向标定与同行评议专家智能指派的关键环节。试点实验表明，学科语义网可以有效地约束关键词的学科指向、实现申请书和专家

特征信息的智能提取、建立申请书与评议专家的智能对接。在国家自然科学基金同行评议专家指派过程中，一个专业、高效、完备的学科语义网有望发挥数据支撑、专业保障、智能决策及结果客观的四大优势。

基于语义网和机器学习的智能辅助指派是科学基金管理研究的前沿课题。在 2014 年的国家自然科学基金项目同行评议专家指派中，我们已经成功尝试并理清了基于学科语义网的智能辅助指派的关键环节。但受制于语料库、专业停词表和同义词表的质量，仍需要学科工作人员和专家进行较多的人工校验工作。基于机器学习的研究方向标定依赖于大量高质量申请书作为学习样本，对于新增或样本量较小的研究方向具有一定的误判率，仍需要多年的样本量积累以提升标定质量。后续的研究主要包括：①在语义网词汇中补充词汇属性约束，优化兼顾词汇属性和语义特征的语义网关键词汇筛选算法；②利用已有语料训练分词算法，降低分词歧义，增强分词冗余的专业指向性；③基于申请书及专家信息进行研究方向标定约束模型构建，进一步降低研究方向误判；④研究学科语义网在 ISISN 智能指派中的应用及实现模式。

致谢　本文工作得到国家自然科学基金（项目资助号：J1324003）资助。感谢以下专家在地理学学科语义网构建及应用过程中的支持和帮助：袁林旺、闾国年、王永君（南京师范大学），刘志刚（北京师范大学），赵小蓉（中国农业大学），李本纲、陶澍、沈泽昊（北京大学），林耿（中山大学），陈崇成（福州大学），张运林（中国科学院南京地理与湖泊研究所），周成虎、程昌秀、陈洁（中国科学院地理科学与资源研究所），傅伯杰（中国科学院生态环境研究中心），冯章献（东北师范大学），黄建毅（北京联合大学），王均平（首都师范大学），卫泽斌（华南农业大学）。

参 考 文 献

[1] 冷疏影，赵小蓉，刘志刚，等. 国家自然科学基金委员会地球科学部一处学科方向分类与关键词编制工作初探. 中国科学基金. 2012. (3)：170—174

[2] 冷疏影. 同行评议辅助指派实验系统研究取得阶段性成果. 中国科学基金. 2013. (3)：160—166

[3] 李东，郝艳妮，何贤芒. 国家自然科学基金同行评议专家信息库的梳理与重构设计. 中国科学基金. 2013. (3)：209—213

[4] Allemang D, Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL-Second Edition. Walham: Morgan Kaufmann. 2011

[5] Matt Taddy. Multinomial Inverse Regression for Text Analysis. Journal of the American Statistical Association. 2013, 108(503): 755—770

A semantic web of geography and its application in computational intelligence supporting peer review system of NSFC

Leng Shuying[1]　Yu Zhaoyuan[2]　Hu Yong[2]　Zheng Yuanming[3]　Song Changqing[1]

(1 *Department of Earth Sciences, National Natural Science Foundation of China, Beijing* 100085;
2 *Nanjing Normal University, Nanjing* 210023;
3 *Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing* 100085)

Abstract　The initial results of semantic web-based computational intelligence supporting peer review system of geography of National Nature Science Foundation are presented in this paper. The construction principles and strategies, process, techiniques, quality control effects and optimization strategy are also analyzed in the paper. We finally summarize some guide from the research of semantic web-based computational intelligence supporting peer review system.

Key words　geography; semantic web; NSFC; computational intelligence supporting peer review system