基于“主语-谓语-宾语”三元组的知识发现研究
——以诱导多能干细胞领域为例*

隗玲[1,2]，胡正银[2]，庞弘燊[3]，覃筱楚[4]，郭红梅[5]，方曙[2]
（1.山西财经大学信息管理学院，太原 030006；2.中国科学院成都文献情报中心，成都 610041；
3.深圳大学图书馆，深圳 518060；4.中国科学院广州生物医药与健康研究院，广州 510530；
5.中国科学院文献情报中心，北京 100190）

摘要：本文提出基于“主语-谓语-宾语（Subject-Predication-Object，SPO）”三元组的生物医学领域知识发现框架，对该框架的关键技术和实施流程进行研究。首先，基于UMLS语料库，利用SemRep工具从生物医药文献中抽取SPO三元组；其次，基于领域知识组织体系，结合自定义词表和清洗规则对SPO进行清洗和筛选；再次，利用NetMiner分别绘制以Subject和Object为中心节点，Predication为边的语义网络图；最后，结合专家解读，实现领域知识发现。本文以诱导多能干细胞领域为例进行实证研究。结果显示，SPO三元组可细粒度地揭示科技文献的知识内容，基于SPO的语义网络能直观地支持领域知识发现，该框架具有兼容、高效、易实施等优点。

关键词：知识发现；SPO；知识组织；语义网络
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2017.09.005

1 引言

文本知识发现（Knowledge Discovery in Text，KDT）是以可信的方式，从文献中识别和提取有用、新颖、潜在有用和最终可理解的模式的过程。信息抽取是KDT的核心技术之一，其目的是从文本中自动抽取实体、实体属性以及实体间的语义关系等信息作为知识发现的基础知识单元[1]。SPO三元组是一种以“主语-谓语-宾语”形式来表示文献中知识单元及其语义关系的知识表示方式，具有语义表示能力丰富、结构简单、技术成熟等优点。通过对SPO三元组中的“主语-谓语-宾语”进行聚类、分类、重构、降维等文本挖掘操作，结合可视化分析工具，可快速、清晰、直观地揭示领域知识主题、重要概念及其关系，被广泛应用于知识组织、语义网络、本体映射、科技文献挖掘与知识发现等领域[2-4]。

诱导多能干细胞（induced Pluripotent Stem Cells，iPSC）技术可通过对成熟细胞进行“重新”编程培育出新的干细胞，拥有与胚胎干细胞相似的分化潜力，可分化为多种类型的细胞，有望用于多种疑难病症的治疗，对于药物筛选、再生医学与发育生物学的研究均具有重要意义，是生物医学领域重要的前沿技术。基于SPO对iPSC领域科技文献中蕴含的知识单元进行深度信息揭示，形成知识单元语义网络，可以多维度、细粒度地呈现iPSC的知识脉络，实现领域知识发现。

2 研究现状

2.1 KDiBL常用语料库与工具

随着生物医学文献数量的快速增长和生命科学研究的交叉发展，生物医学领域知识发现（Knowledge Discovery in Biomedical Literature，KDiBL）已成为一

* 本研究得到中国科学技术信息研究所ISTIC-EBSCO文献大数据发现服务联合实验室基金项目“基于SemRep与SKOS的科技文献语义知识组织应用示范研究”资助。

个重要研究领域[1]。统一医学语言系统（Unified Medical Language System，UMLS）是美国国家医学图书馆（the United States National Library of Medicine，NLM）自1986年研究和开发的生物医学一体化超级叙词表系统[5]。其融合多个生物医药、卫生健康等领域词表，采用字符串-术语-概念的组合方式对生物医学领域的术语进行规范，并提供计算机处理的互操作接口，是KDiBL常用的标准语料库[2,5]。NLM基于UMLS开发出一系列自然语言处理工具。其中，MetaMap是一款将自由词向UMLS概念映射的工具[2,6]，可标记出文本中包含的UMLS概念，作为一项基础性文本处理工具被广泛应用于KDiBL的各领域。SemRep是NLM语义知识表示项目的重要成果之一，是一款基于UMLS和MetaMap的生物医学文献语义知识抽取与表示工具[2,7]，SemRep可从海量生物医学文献自动抽取SPO结构来揭示文献的知识内容。其中，SPO三元组的主语和宾语是UMLS中的概念，谓语来自于UMLS语义网络中的语义关系。UMLS语义网络包含133种语义类型和54种语义关系。以“DNA-ADMINISTERED_TO-Pluripotent Stem Cells”为例，DNA为主语，其语义类型为实体物质，语义关系为ADMINISTERED_TO，宾语为Pluripotent Stem Cells，其语义类型为解剖要素[5]。

医学主题词（Medical Subject Headings，MeSH）是NLM开发和维护的综合型词汇表[8]，用来描述生物医学主题或特性。MeSH由主题词变更表、字母顺序表、副主题词和树形结构组成，在文本挖掘过程中常用于词表清洗和语义关系计算。树形结构表将表中所有主题词按照学科性质和语义关系进行层次分类，表示概念间的隶属关系，越底层的概念越具体，所包含的信息颗粒度越细。NLM提供的生物医学文献数据库PubMed包含基础医学、临床医学、医疗保健、微生物等多个领域的海量文献[9]，富含医学疾病和生物信息知识，已成为生物医学文献知识发现的核心数据库。

2.2 KDiBL研究进展

基于上述语料库与分析工具，KDiBL研究有了更进一步的发展，不仅可以基于文档词频统计信息和高频动词识别重要关系开展医学发现，而且可以根据概念间的语义关系挖掘知识内容。Reeve等首先利用 UMLS识别生物医学文献中的名字词组，将其转化为UMLS概念和语义类型，并基于概念间的语义类型关联关系生成词汇链，然后结合概念出现的频次和词汇链的3个特征识别强词汇链，最终形成文献知识主题[10]；Kilicoglu等开发的Semantic MEDLINE自动摘要系统利用SemRep对文献集中包含的谓语进行数据挖掘，将其划分成疾病治疗、药物相互作用、药物基因组学和疾病遗传因素4个研究主题，并通过分析其语义关系和频次生成语义网络图[11]；Fiszman等对循证医学文献开展知识发现研究，对53种药物的干预效果进行识别[12]；Workman等为膀胱癌寻找对应的基因信息，采用3种统计指标对重要的语义述语进行知识抽取，将结果与相关标准进行对比，最终验证了其方法用于管理基因数据库的优越性[13-14]；Zhang等利用中心度指标抽取语义网络的关键节点，对5种不同学科疾病的伴发疾病、发病部位、治疗药物和治疗措施进行知识发现[15]；Cairelli等通过对大脑神经损伤语义关系网络按照关系频次和概念关联度进行裁剪，从海量科学概念中发现17种有助神经损伤诊断的潜在生物标记[16]。

总之，UMLS及其相关语料库与工具集已成为KDiBL研究的基础性资源，SPO三元组是揭示生物医药文本信息的基础知识单元。

3 基于SPO的生物医学领域知识发现框架

本文以药物基因组学领域知识发现为例，将基于SPO的生物医学领域知识发现框架进行描述，如图1所示。该框架由语义关系架构和知识发现流程两部分组成，其中语义关系架构定义领域的知识组织体系，主要用于指导SPO抽取和清洗，是该框架的关键技术；知识发现流程则描述实施过程。

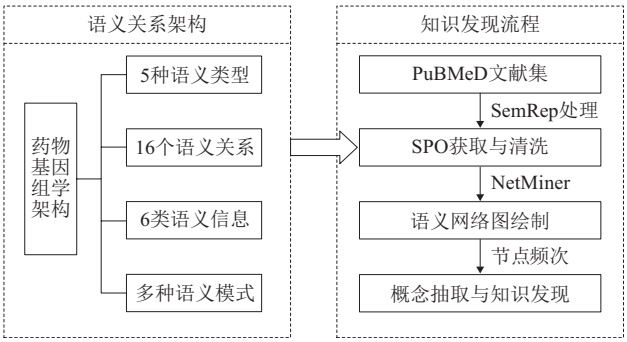

图 1 基于SPO的药物基因组学领域知识发现框架

3.1 药物基因组学语义关系架构

UMLS语义网络能较全面地揭示生物医学涉及概

念间的各种关系，为语义抽取和知识发现提供支撑。为有针对性地分析生物医学不同领域概念间的语义关系，学者提出可将医学概念划分为几种主要的语义类型，并结合医学理论确定若干个谓词代表重要的语义关系。不同的谓词与不同语义类型的概念组合成不同的语义模式，相应的语义模式集合共同表达某类语义信息，即具体领域的某个核心研究内容。这样一种对概念、关系及其组合进行分类和定义的模式，被称为架构[5]。

Fismzan等提出疾病治疗学的架构，将该领域的研究内容分为伴发疾病、发病部位、治疗药物、治疗措施[17-18]，与之相关的语义关系为COEXISTS_WITH、LOCATION_OF、TREATS和PREVENTS；Fiszman等接着提出物质相互作用研究的架构，将物质分为药物、化学品、生理学、病状[18]，与之相关的语义关系为AFFECTS、CAUSES、COMPLICATES、DISRUPTS、ISA、TREATS、PREVENTS、INTERACTS_WEITH；随后，Fismzan等再次提出药物基因组学和疾病基因伦理学研究的架构[19-20]。疾病基因伦理学的架构将语义类型分为基因表象、解剖要素和疾病过程，与之相关的语义关系为AFFECTS、ASSOCIATED_WITH、AUGMENTS、CAUSES、DISRUPTS、COEXISTS_WITH、INHIBITS、PREDISPOSES、STIMULATES。药物基因组学架构定义了5种语义类型和16种语义关系、6类语义信息和对应的多种语义模式。该架构在UMLS标准语义类型和语义关系的基础上，将语义类型与语义关系的组合定义为语义模式，一组语义模式包含多个具有相近语义关系的SPO，描述一类语义信息具体内涵。通过定义语义模式与语义信息类型，可将大量SPO所蕴含的药物医学信息进行分组归类，从而体现其知识主题。

3.2 基于SPO的生物医学领域知识发现流程

Fiszman等提出SPO获取和处理的原则：根据相关性标准参考领域架构，定义具体子领域相关的核心语义SPO；根据连接性标准识别与核心SPO相关联的其他SPO；根据新颖性标准剔除一般的、无具体信息的SPO，这些SPO中的主语或谓语一般位于靠近UMLS词表的根节点的位置；根据显著性标准剔除频次低于平均值的SPO[17]。

本文以诱导多能干细胞领域iPSC为实证对象，参考药物基因组学架构和Fiszman等[17]的数据获取与处理原则，制定知识发现流程。

（1）SPO获取与清洗。根据确定的检索表达式在PubMed数据库中检索文献并下载相关的SPO，从中筛选出主语或宾语为iPSC的SPO，形成分析使用的初始数据集。确定清洗规则和清洗流程，在初始数据集对SPO进行筛选和剔除。首先，利用预设的药物基因组学相关语义搭配模式对语义述语进行筛选，保留架构范围内的述语，删除其他述语；其次，根据新颖性和重要性原则，剔除含义宽泛、对信息抽取无意义的概念组成语义述语；最后，合并重复的SPO，一篇文献中多个相同的SPO视为同一SPO。另外，还可根据SPO出现的频次对其进行过滤，设置阈值去掉出现频次较低的SPO。

（2）语义网络图绘制。根据SPO中主语和宾语基于谓语的共现关系，绘制其语义网络图。

（3）概念抽取与知识发现。对语义网络中谓语的iPSC概念按照出现频次排序，抽取排序靠前的概念形成知识主题。

4 实证分析

4.1 数据来源与数据处理

在PubMed数据库中以“Regenerative Medicine”为关键词进行检索，选取被Semantic Medline数据库索引且文献类型为“Journal Article”，时间为2010年1月1日—2014年12月31日，检索日期为2017年3月28日，检索获得10 687篇文献。

利用SemRep抽取每篇文献题目和摘要文本中的语义述语，得到65 042个原始SPO三元组。为聚焦于诱导多能干细胞技术，首先以主语或宾语为“iPSC”在原始SPO三元组数据集中进行筛选，获得相关语义述语782个，其中以“iPSC”为主语的SPO数量为634个，以“iPSC”为宾语的SPO数量为148个；其次，依据药物基因组学框架设定谓词和领域主题词进一步筛选，保留758个SPO，不属于框架设定的谓词有NEG_PART_OF，非领域主题词有notch、complex、research personnel、material等；再次，以“iPSC”为主语的SPO数量为618个，以“iPSC”为宾语的SPO数量为140个；最后，在所抽取的语义述语中，有些主语或宾语概念的含义过于宽泛，不能为概念抽取提供有意义的信息。如location、surface、central、enviroment、part、place、intermediate、generalized、landscapsce等，需要将其

剔除，共得到用于分析的SPO数量为698个，其中以“iPSC”为主语的SPO数量为603个，以“iPSC”为宾语的SPO数量为95个。

4.2 iPSC文献知识发现

本节在上述SPO统计的基础上，进一步对SPO中谓语进行分析，并以谓语为边绘制SPO语义网络图。

4.2.1 Prediction分析

iPSC领域的语义类型主要集中于实体物质、解剖要素、生物有机体和病理学四个方面。从概念类型数量角度看，实体物质和解剖要素的相关概念居多；从概念出现频次角度看，生物有机体的相关概念居多（见表1）。该领域的语义述语谓语类型不多，共有AFFECTS、DISRUPTS、AUGEMTS、ADMINSTERED_TO、PRODUCES、LACATION_OF、PART_OF种，其语义信息主要聚焦于药理作用和生物特征两方面（见表2）。其中，表示生物特征的语义模式包含数量众多的由谓语LACATION_OF和PART_OF连接的SPO三元组；表示药理作用的语义模式中，SPO三元组中谓语出现的频次依次为PRODUCES、AUGEMTS、ADMINSTERED_TO、AFFECTS和DISRUPTS。从语义类型和语义模式可初步判断，该领域的研究重点聚焦于诱导多能干细胞的生成，具体内容为使用重组编码来源物质借助各种辅助物质生成iPSC。

相较药物基因组学架构，iPSC语义类型缺少生物医学过程，语义信息类型缺少遗传病因、物质关系、临床作用和过程并发。

表 1 诱导多能干细胞领域语义类型

语义类型	语义模式
实体物质	genes；c-myc genes|MYC；MicroRNAs；Proteins；BCL1 Oncogene；CDX2 gene；SOX2；Membrane Proteins；KLF4；Monoclonal Antibodies
解剖要素	Pluripotent Stem Cell；human tissue；dental pulp；kidney；structure of anterior cerebral artery；individual；entire hair follicle；Sendai virus；heart；human tissue
生物有机体	human；mus；house mice；Rattus Norvegicus；Canis Familiaris；Oryctolagus Cuniculus；primates；homo sapiens；infraclass eutheria；monkeys；animalia；family suidae
病理学	Parkinson Disease；Neurodegenerative Disorders；Neoplasm；Amyotrophic Lateral Sclerosis；Spinal Muscular Atrophy；Down Syndrome；Neuropathy

表 2 诱导多能干细胞领域语义信息类型和语义模式

语义信息	语义模式	示 例
药理作用	{Substance}AFFECTS{Anatomy} {Substance}DISRUPTS{Anatomy} {Substance}AUGMENTS{Anatomy} {Substance}ADMINISTERED_TO {Anatomy} {Substance or anatomy}PRODUCES {Anatomy}	{FMOD, MicroRNAs, Platelet Factor}AFFECTS{PSC} {Acid sphingomyelinase, N-glycolyl neuraminic acid} DISRUPTS{PSC} {antigens, BMP, CCND1 gene, MYC gene} AUGMENTS{PSC} {DNA, MicroRNAs ,TRANSCRIPTION FACTOR} ADMINISTERED_TO{PSC} {TRANSCRIPTION FACTOR, POU5F1, MYC} PRODUCES{PSC}
生物特征	{Anatomy}PART_OF {Living Being OR Anatomy} {Anatomy}LOCATION_OF {Substance OR Pathology} {substance}PART_OF{anatomy}	{PSC}PART_OF{human, mus, Canis Familiaris, dental pulp,human tissue} {PSC}LOCATION_OF{PAX3 gene, Parkinson Disease, Down Syndrome} {Activins, GFP, SALL4, BCL1A}PART_OF{PSC}

4.2.2 语义网络分析

NetMiner是将社会网络分析和可视化探索技术相结合的工具，允许使用者以可视化和交换的方式探查网络数据，分析网络潜在的模式和结构，并具有高级的图形特性[21]。本文使用NetMiner工具绘制基于谓语共现关系的有向语义网络图。网络节点表示语义概念，节点形状表示概念的语义类型。语义关系用节点间连线的标签标识，连线具备宽度和方向两个属性，宽度表示对应语义术语的频次，方向由主语指向宾语。

图2中iPSC为主语，其他节点为宾语，二者的语义关系主要有PART_OF、LOCATION_OF、PRODUCES三种，对应的语义模式为{Anatomy}PART_OF{Living Being or Anatomy}、{Anatomy}LOCATION_OF{Substance or Pathology}、{Substance or Anatomy}PRODUCES{Anatomy}。前两种语义模式表示诱导多能干细胞领域蕴含的生物特征，第三种语义模式揭示该领域涉及的药理作用。其中，五角星形节点隶属于实体组生物有机体或解剖要素，主语iPSC与这些宾语形成的语义关系为PART_OF。生物有机体出现频次较高的概念有human、house mice、mus，解剖要素出现频次较高的概念为dental pulp。现阶段用于PSC重组编码通常用的细胞为人体皮肤细胞，其次为动物成纤维细胞和牙髓细胞；用于实验的动物对象主体为鼠类，其次是家兔、家猪、猴子等；涉及的人体病灶组织有心脏、肝脏等。圆形节点隶属于实体组物质要素或病理学，主语iPSC与这些宾语形成的语义关系为LOCATION_OF。物质要素中出现频次较高的概念为TRANSCRIPTION FACTOR、POUSP1、SOX2、MYC等。物质要素组成员为各种转录因子、基因、蛋白质、酶等生成诱导多能干细胞的辅助因素。病例症状组成员有帕金森病、神经退行性疾病、脊髓性肌萎缩、唐氏症、神经病等，显示现阶段诱导多能干细胞研究所针对的疾病类型。十字形节点也隶属于实体组生物有机体或解剖要素，主语iPSC与这些宾语形成的语义关系为PRODUCE。十字形节点内容与圆形节点内容有较大重叠度，为了区分两种语义关系，特此用两种不同的形状标识节点。此处需要说明的是箭头指向是由主语指向宾语，但不代表语义关系一定也是主语指向宾语，图2中的PRODUCES关系和图3中的PART_OF关系需要反向理解。

图3中iPSC为宾语，其他节点为主语，分别用三角形、菱形、圆形、五角星形和十字形五种形状表示，二者对应的语义关系为AFFECTS、ADMINISTERED_TO、DISRUPTS、AUGMENTS和PART_OF。对应的语义模式有{Substance or Anatomy}AFFECTS OR ADMINISTERED_TO OR DISRUPTS OR AUGEMENTS{Anatomy}、{Substance or Anatomy}PART_OF{Anatomy}。其中，除语义关系DISRUPTS表示某些蛋白质或酸性鞘磷脂酶会破坏诱导多能干细胞的生成外，其他多种基因、转录分子等有助于诱导多能干细胞的生成或使用。语义关系PART_OF表示各种基因或蛋白质是诱导多能干细胞的组成部分，与图2中的LOCATION_OF和PRODUCE关系形成互补。

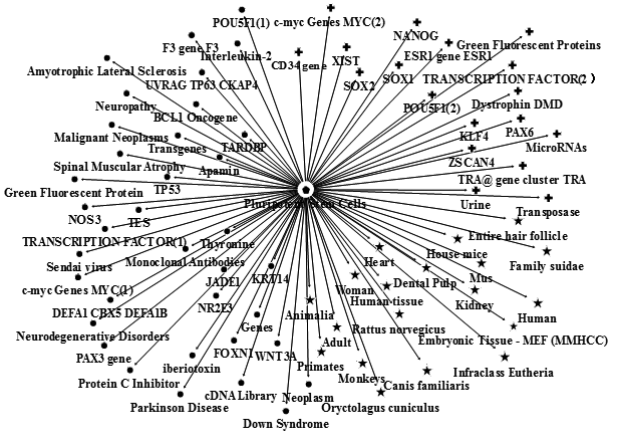

图 2 Pluripotent Stem Cells为主语的语义网络

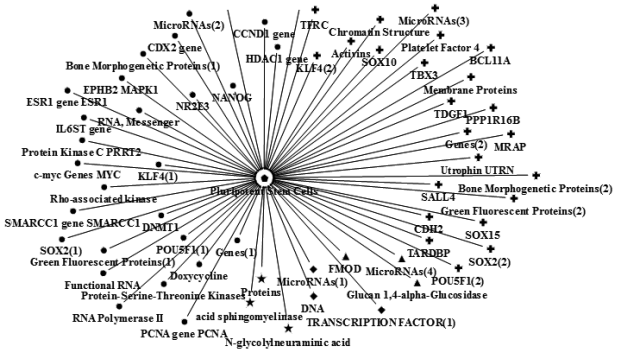

图 3 Pluripotent Stem Cells为宾语的语义网络

4.2.3 知识主题分析

本节结合专家解读，对iPSC领域的知识主题进行描述和分析。通过分析发现该领域的知识主题集中于生物特征和药理作用两大类，而生物特征可细分为iPSC生成的影响要素（或辅助要素）、iPSC针对的疾病类型、iPSC实验涉及的人体病灶组织和iPSC实验的

对象四个方面，而药理作用聚焦于多种要素对iPSC生成的影响作用。

（1）iPSC生成的影响要素（或辅助要素）。诱导多能干细胞是通过对成熟细胞进行“重新编程”培育出的干细胞。在细胞重组过程中，对源细胞的基因表达和转录调控是两个非常重要的环节。因此，语义网络中出现大量基因类、蛋白质类概念，其中TRANSCRIPTION FACTOR、genes、MicroRNAS、OKSM为出现频次较高的重要概念。两个重要环节中的任何因素都可对iPSC生成产生影响。

（2）iPSC针对的疾病类型。诱导多能干细胞实验采用最多的成熟细胞为人体皮肤细胞和猪皮肤干细胞，最终生成的诱导产物有心肌细胞、肝脏细胞、人类红细胞和血小板、胰岛素分泌细胞和神经干细胞等，分别用于治疗心脏病、肝脏疾病、贫血、糖尿病和神经变性疾病等。iPSC细胞用于以上疾病的细胞治疗或基因治疗。此外，也有研究者使用牙髓细胞作为诱导多能干细胞的来源细胞将其重组后生成神经组细胞，用于治疗自闭症。

（3）iPSC实验涉及的人体病灶组织。心脏、肝脏、大脑是iPSC实验涉及的较典型的人体病灶组织，此外，还有眼睛、胰腺器官等。iPSC在治疗影响再生能力较差的器官、组织的毁灭性疾病和神经变性疾病方面有巨大的潜力。

（4）iPSC实验的对象。目前有关诱导多能干细胞的研究基本处于实验研究阶段，诱导生成的各种细胞一般用于鼠类、家兔、家犬、猴子等动物。

多种要素对iPSC生成的影响作用。诱导多能干细胞药理作用方面的摘要主题聚焦于多种要素对PSC生成的影响作用。除三个节点概念（Proteins、acid sphingomyelinase和N-glycolylneuraminic acid）对PSC的生成有抑制作用外，其他概念如小分子核糖核酸、特别的蛋白质、转录分子及各种基因等对PSC的生成均有正向促进作用。如NANOG细胞周期蛋白可抑制PSC重编程过程的反复性，提高重编程效率；转录因子SOX2在多能干细胞形成的过程中扮演重要角色，是干细胞多能性的一个指示器，具有影响干细胞维持或分化的能力。

5 结语

本文利用UMLS和SemRep从iPSC领域文献中抽取SPO三元组，参考药物基因组学语义关系架构对SPO三元组数据集进行清洗和筛选，构建富含语义信息的iPSC领域SPO语义网络，挖掘iPSC领域知识主题的内涵。研究结果显示，该框架具有细粒度、高效、直观等优点。该框架可以兼容生物医学领域多种架构，帮助领域专家快速、直观地发现海量文献中非结构化文本信息所蕴含的知识主题；基于SPO的语义网络能直观细致地揭示概念间的语义关系，从微观层面深入揭示文献内容，细粒度地揭示医学文献的知识内容。研究的不足在于，分析使用的数据集来自药物基因组学文献，数据内容不够完整，对知识发现结果有所影响。未来，将进一步完善iPSC领域数据集，通过对语义网络进行子团识别和聚类分析，开展渐进式知识发现研究。

参考文献

[1] 李清.一体化医学语言系统的语义相似度及推理研究[D].哈尔滨:哈尔滨工业大学,2012.

[2] 白海燕,王莉,梁冰.UMLS及其在智能检索中的应用[J].现代图书情报技术,2012(4):1-9.

[3] 胡正银.基于个性化语义TRIZ的专利技术挖掘研究[D].北京:中国科学院大学,2015.

[4] KESELMAN A,ROSEMBLAT G,KILICOGLU H,et al.Adapting semantic natural language processing technology to address information overload in influenza epidemic management[J].Journal of the American Society for Information Science & Technology Jasist,2010,61(12):2531-2543.

[5] NCBI.UMLS® Reference Manual[EB/OL].[2017-05-31].http://www.ncbi.nlm.nih.gov/books/NBK9676/.

[6] ARONSON A R,LANG F.An overview of MetaMap:historical perspective and recent advances[J].Journal of the American Medical Informatics Association,2010,17(3):229-236.

[7] ARNOLD P,RAHM E.Semrep:a repository for semantic mapping[EB/OL].[2017-09-08].https://dbs.uni-leipzig.de/en/publication/title/semrep_a_repository_for_semantic_mapping.

[8] NCBI.Introduction to MeSH[EB/OL].[2017-05-31].https://www.ncbi.nlm.nih.gov/mesh.

[9] NCBI.PubMed Centra[EB/OL].[2017-05-31].https://www.ncbi.nlm.nih.gov/pubmed.

[10] REEVE L H,HAN H,BROOKS A D.The use of domain-specific concepts in biomedical text summarization[J].Information Processing & Mana-gement,2007,43(6):1765-1776.

[11] KILICOGLU H,FISZMAN M,RODRIGUEZ A,et al.Semantic MEDILINE: a web application for managing the results of PubMed searches[EB/OL].[2017-05-31].https://www.researchgate.net/publication/228617741.

[12] FISZMAN M,DEMNER-FUSHMAN D,KILICOGLU H,et al. Automatic summarization of MEDLINE citations for evidence-based medical treatment:a topic-oriented evaluation[J].Journal Biomedical Informatics,2009,42(5):801-813.

[13] WORKMAN T E,FISZMAN M,HURDLE J F,et al.Biomedical text summarization to support genetic database curation:using Semantic MEDLINE to create a secondary database of genetic information[J]. Journal of Medical Library Association Jmla,2010,98(4):273-281.

[14] WORKMAN T E,HURDLE J F.Dynamic summarization of bibliographic-based data[J].Bmc Medical Informatics and Decision Making,2011,11(1):1-10.

[15] ZHANG H,FISZMAN M,SHIN D,et al.Degree centrality for semantic abstraction summarization of therapeutic studies[J]. Journal of Biomedical Informatics,2011,44(5):830-838.

[16] CAIRELLI M J,FISZMAN M,ZHANG H,et al.Networks of neuroinjury semantic predications to identify biomarkers for mild traumatic brain injury[J].Journal of Biomedical Semantics,2015,6(1):25.

[17] FISZMAN M,RINDFLESCH T C,KILICOGLU H.Abstraction summarization for managing the biomedical research literature[C]// Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics.[S.l.]:[s.n.],2004:76-83.

[18] FISZMAN M,RINDFLESCH T C,KILICOGLU H.Summarizing drug information in Medline citations[J].AMIA.Annual Symposium proceedings.AMIA Symposium,2006:254-258.

[19] AHLERS C B,FISZMAN M,DEMNER-FUSHMAN D,et al. Extracting semantic predications from MEDLINE citations for pharmacogenomics[J].Pac Symp Biocomput,2007,12:209-220.

[20] WORKMAN T E,FISZMAN M,HURDLE J F,et al.Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information[J]. Journal of the Medical Library Association Jmla,2010,98(4):273-281.

[21] Cyram.NetMiner v4.3.0[EB/OL].[2017-05-08].http://www.netminer.com.

作者简介

隗玲，女，1981年生，博士研究生，研究方向：科学计量、知识发现，E-mail：weiling@mail.las.ac.cn。
胡正银，男，1979年生，博士，研究方向：知识组织、技术挖掘，E-mail： huzy@clas.ac.cn。
庞弘燊，男，1985年生，博士，研究方向：信息可视化、知识组织，E-mail： phs@szu.edu.cn。
覃筱楚，女，1988年生，硕士，研究方向：生物信息学，E-mail：qin_xiaochu@gibh.ac.cn。
郭红梅，女，1985年生，博士，馆员，研究方向：文本挖掘、科学计量分析，E-mail：guohm@mail.las.ac.cn。
方曙，男，1957年生，博士，研究方向：科学计量、科技政策，E-mail：fangsh@clas.ac.cn。

Study on Knowledge Discovery Based on "Subject-Predication-Object" Predications: A Case Study of Induced Pluripotent Stem Cells

WEI Ling[1,2], HU ZhengYin[2], PANG HongShen[3], QIN XiaoChu[4], GUO HongMei[5], FANG Shu[2]
(1.School of Information and Management, Shanxi University of Finance and Economics, Taiyuan 030006, China;
2.Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041, China; 3.Shenzhen University Library, Shenzhen 518060, China;
4.Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China;
5.National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper summarizes a set of knowledge discovery framework to make studies on knowledge discovery in biomedical literature based on Subject-Predication-Object (SPO) predications, and studies the key technology and implementation process of the framework. First, SPO predications were extracted from the biomedical literature by using UMLS corpus and SemRep; then, according to the knowledge organization system, vocabulary and cleaning rules were self-defined, the SPOs were cleaned and filtered; next, semantic network diagrams were constructed by NetMiner, which included subjects and objects as the center nodes and predications as the edges; finally, combining the diagrams and experts' interpretation, domain knowledge discovery was achieved. In this paper, an empirical study was conducted to investigate the field of pluripotent stem cells. Research results show that, SPO predications can reveal the knowledge content of scientific literature, and SPOs-based semantic networks can intuitively support domain knowledge discovery. The framework is compatible, efficient and easy to implement.

Keywords: Knowledge Discovery; SPO; Knowledge Organization; Semantic Network

（收稿日期：2017-07-20）