碎片化科研创新点动态挖掘研究*

温有奎[1]，吴广印[2]
（1. 北京万方软件股份有限公司，北京 100038；2. 中国科学技术信息研究所，北京 100038）

摘要：从海量科技信息中挖掘出科研创新点碎片已成为大数据环境下知识挖掘与服务的一个关键问题，也仍然是迄今为止非结构化知识发现的一个难题。文章提出一种碎片化科研创新点动态挖掘方法。通过对学术成果的要素和条件分析，建立学术成果创新要素的关键变量和语义关系，给出学术成果创新点的本体模型；基于模型的理论指导，实现科技文献中科研创新点碎片的动态挖掘系统。该方法有利于过滤海量科技文献的创新点，发现文献中的知识关联关系，提高文献知识挖掘的效率，为科研工作者快速方便地直接获取科研动态信息提供技术支持。

关键词：碎片化；创新点；本体建模；动态挖掘
中图分类号：TP311
DOI：10.3772/j.issn.1673—2286.2014.07.005

1 引言

大数据的到来为科技信息服务机构提出了新的挑战，如何解决科研工作中的信息淹没而又知识贫乏的困境，研究新的知识挖掘技术成为当前信息服务业最为关键的问题。早在20世纪中叶科学家就在积极地探讨科学知识分裂现象，寻找直接挖掘所需知识的方法，但一直没有很好的解决方案。20世纪60年代，美国情报学家Swanson教授对科学知识碎片（Fragmentation of Science Knowledge）理论提出新的看法：（1）客观知识总量与人类吸收能力存在巨大的差距；（2）跨学科的信息传递变得更加困难；（3）跨学科间存在潜在未被发现的关联。他首次提出并验证了利用文献间存在知识碎片的推理发现新知识的方法[1]。为此，许多学者也做了大量的研究工作[2-5]，李国杰院士从知识创新的源头提出，我国已经走过论文数量增长历史阶段，现在应该是强调论文质量的时候了[6]。李怀祖教授的文章强调，创新是一篇论文的灵魂[7]，称得上科学研究成果的论文，一定要有新发现、新假设或新理论。韩客松博士抽样统计国内中文期刊中自然科学论文的标题与论文主题的基本符合率为98%[8]。至2012年末，非结构化数据占有比例达到整个数据量的75%以上[9]。大数据和云计算的出现加剧自然科学成果的传播速度，也加剧了信息淹没和知识贫乏的速度，信息检索需求已向更深层次的知识发现需求发展。因此，本文提出一种从学术论文成果中对科研创新点进行动态挖掘的方法。通过对构成学术论文成果的要素和条件的分析，建立学术论文成果创新点表现要素的关键变量和语义关系，构建创造性学术论文成果的本体模型，基于该模型的理论指导，实现对科技文献中创造性成果的动态挖掘方法，并利用关联规则为用户关注的创新点自动推荐关联关键词，为用户发现新的创新点提供帮助。

2 创新点挖掘的本体模型构造

2.1 创新点要素与判定

2.1.1 创新点要素分析

本文提出的创新点动态挖掘的思想，是把海量科技期刊论文中的创新点以短语片段形式抽取出来并加以聚类，旨在解决科技期刊论文关注点的快速、简明、

* 本研究得到国家科技支撑计划课题“跨媒体科技文献数字资产管理及内容复用关键技术研发与应用示范”（编号：2012BAH90F03）资助。

直接、准确的检索问题。以创新点过滤海量科技文献，探索以创新点动态挖掘科技文献知识的新方法，为建立科研创新点的语义关联推理建立基础。

学术成果的创新点是科学研究活动的灵魂，是科学发现与理论创新成果的核心，是科研工作者关注和跟踪的关键信息。有研究者提出科学发现与理论创新成果应当满足六项要素[10]：（1）新颖性；（2）创造性；（3）自洽性；（4）包容性；（5）简明性；（6）可实验检验性。上述六项要素，1-4项是必须同时满足的条件，5-6项则视具体情况而定。

2.1.2 创新点的判定[10]

（1）新颖性的判定：指科学发现与理论创新成果向社会公开之日以前，没有同样的科学事实和科学理论在国内外出版物上公开发表过，或者以其他方式为公众所了解。

（2）创造性的判定：指作者独自创作完成的，而不是剽窃抄袭他人的；同公开之日以前的所有科学事实和科学理论比较，该科学发现与理论创新成果有实质性的突破和显著的进步；科学发现与理论创新成果可以是既有成果的改进与发展，但必须与既有成果有显著的不同并有实质性的突破，论述中应当引证既有成果的论文资料。

（3）自洽性的判定：是一个理论能够成立的必备条件。指建构一个科学理论的若干个基本假设之间，假设与一系列结论之间，各个结论之间必须相容，不相互矛盾，逻辑推理和数学演算正确无误。

（4）包容性的判定：指新的科学理论应当能够解释已有的实验事实，新的科学理论应该在一定的条件下回归到已经被实践所证明、在同样条件下成立的相应的现有科学理论。

（5）简明性的判定：应当从尽可能少的基本假设出发描述尽可能多的认识对象，包罗尽可能多的科学结论。

（6）可实验检验性的判定：指自然科学理论可诉诸实验的检验。

2.2 创新点三要素的鉴别

本体（Ontology）是对客观存在的一个系统的解释或说明，它关心的是客观现实的抽象本质。本体应用在计算机领域可以构造对象模型，以及对象的关系和属性。我们利用语义网构建一个学术论文创新点挖掘的本体模型，这个模型有助于对无结构和半结构化文本知识的理解和挖掘。

学术论文完善地论述创新点，一般要回答三方面的问题[7]：（1）创新点是什么；（2）为何要提出此创新点；（3）回答这个创新点是否成立的质疑。为了回答上述问题，论文应有三方面的内容，即创新点的表述、创新点的理论和实际背景评述以及创新点的论证。表述反映论文的贡献所在，背景评述衬托出论文的价值，论证则表明创新点的可信程度，三者缺一不可。为此我们将这三点假设为鉴别创新点的三要素，即创新点存在的必要充分条件。

2.3 创新点的本体模型

2.3.1 创新点的本体模型构建

创新点分布在论文的整个结构中，表现为主题中的创新点、技术背景中的创新点、技术方法中的创新点、论文结论中的创新点和总体创新点。由于写作要求，每种创新点功能表现出了独特的知识本体结构。建立学术论文创新点的知识本体模型，是实现学术论文创新点智能识别和动态挖掘的关键理论。

一般学术论文对创新点的描述由五大部分组成，既展现出一种层次关系，又表现出一种网状关联关系，学术论文的创新点本体模型见图1。

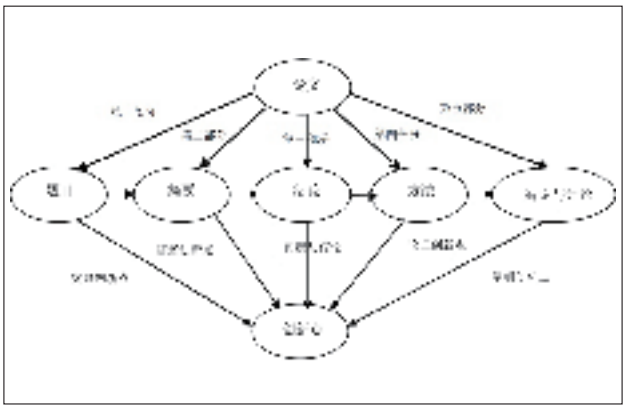

图1 学术论文创新点本体模型图

2.3.2 创新点本体模型的解释

（1）题目凸显创新点

学术论文的标题反映主题的创新点，是文章的灵魂。有人抽样统计国内中文期刊中自然科学论文的标题与论文主题的基本符合率为98%[6]。好的论文题目能明白无误地告诉读者创新点是什么，具有画龙点睛的功能。

（2）摘要陈述创新点

摘要是向读者叙述本文的创新点和它的价值，包含三个组成部分：①问题说明：指明论文所要解决的问题，并令读者意识到此问题的价值所在。②创新点：研究工作的贡献，即与众不同的新发现。创新点应占摘要中大部分篇幅。③论证途径的说明：创新点论证过程不必叙述，但作者若在论证方法技术上有改进和创新之处则可以写上。

（3）绪论论述问题与假设

论文首先都要点题，目的是为了阐明问题。绪论可细分为几部分：①问题提出及背景；②文献综述；③假设表述，有时还加上关键术语界定内容。

（4）方法论证创新点

方法部分要从方法论角度详细描述论证过程，使读者可以根据论文描述的方法，独立地重复此项论证和验证工作。方法部分应包括三项内容的描述：研究主体、论证框架及步骤和数据分析。

（5）结果和讨论阐明创新点

主要阐明假设验证结果，亦即论文的主要创新。结果应描述新发现取得的过程，尽管分析结果是围绕研究者的假设展开的，但分析过程的描述应该避免主观的议论，只是摆事实、数据和论据，强调叙述过程的客观和科学性。结果表述中也包括和同类分析结果的比较，揭示和前人分析结果的不同，以衬托出本研究工作的创新之处。

3 科研创新点碎片挖掘

3.1 建立创新点动态挖掘模板

3.1.1 创新点动态挖掘模板的结构

根据图1给出的学术论文创新点本体模型图的理论，我们建立了论文创新点动态挖掘模板。模板由5个模块组成。

（1）问题模块（用户关注点，也即论文的创新点，来自标题）

（2）方法模块（解决关注点问题所提出的新方法，来自文摘的创新点）

（3）结论模块（新发现、阶段性结论，来自文摘的创新点）

（4）作者模块（论文作者）

（5）时间模块（论文发表的时间）如果我们将S定义为结构，s_1、s_2、s_3、s_4、s_5分别表示问题类、方法类、结论类、作者类、时间类变量，我们就得到了创新点动态挖掘模板的结构函数，$S=(s_1,s_2,s_3,s_4,s_5)$。

3.1.2 创新点模板内部的语义关系

上述5个模块之间构成了5种语义关系。这5种语义关系可描述为：问题由作者提出，作者采用了方法，方法解决了问题，问题得到了结论，结论验证了方法。由此我们建立了由提出、采用、解决、得到、验证5种特征词组成的语义关系。

如果我们把V定义为语义，v_1、v_2、v_3、v_4、v_5分别表示提出类、采用类、解决类、得到类、验证类变量，我们就得到了创新点模版内部的语义关系函数，$V=(v_1,v_2,v_3,v_4,v_5)$。

其中，问题、作者、方法、结论这4个变量是关键变量。时间是依从变量，发生在关键变量的过程中。且问题、方法和结论是三个基本变量，具有直接关联关系，而作者是间接关系。

通过以上结构和语义变量的分析，我们得到了创新点动态挖掘模板是一对由两类特征词组成的结构关系图G：

$G=\{S,V\}$，其中$S=(s_1,s_2,s_3,s_4,s_5)$，$V=(v_1,v_2,v_3,v_4,v_5)$。

其中，$v_1=(v_{11},v_{12},v_{13},v_{14}\cdots)$。如$v_1$又可以写成：$v_{11}$（提出了），$v_{12}$（给出了），$v_{13}$（设计了），$v_{14}$（分析了）等。同样，$v_2$,$v_3$,$v_4$,$v_5$都有各自的同义词表示方式，因此，语义类型是由这五种语义特征词汇的聚类。创新点动态挖掘模板的关联关系如图2所示。

3.2 创新点的动态挖掘模式

模式是基于模式逻辑抽取的核心，文本模式是一个实例概念的形式和一般定义，而模式元素又是在模式中可能应用的文本实体类型。我们采用的创新点动态挖掘模式是一种模式匹配方法。为确定模式的元素，我们对描述创新点的要素和判定的特征词做了统

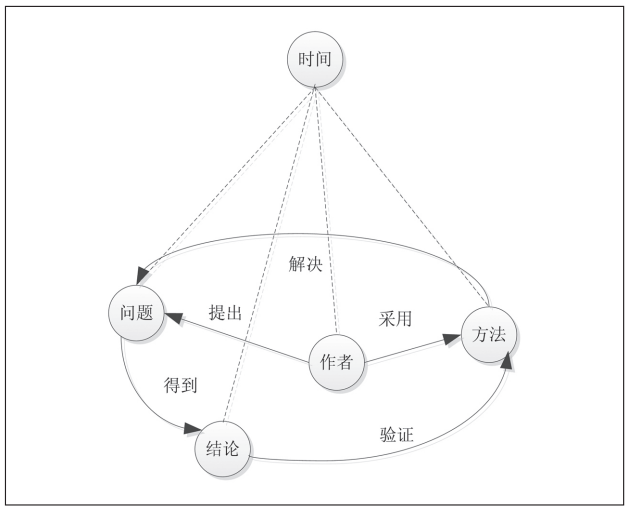

图2 创新点两类变量之间的关联关系动态挖掘模板

计分析，统计结果表明[2]绝大多数科技论文的创新点都是由“提出”这样的特征词引导出的，占了统计结果的71.8%，其次经常出现的特征词分别是“设计”、“研究”、“介绍”等，大约各占3.6%、3.6%和3.2%，这些特征词出现的频率和“提出”相比相差甚远。统计结果还表明论文创新点有特征词引导的大约占98.4%。统计结果证实了本文2.3.2提出的学术论文的文摘具有（1）问题说明、（2）创新点、（3）结果说明三个组成部分的基本规律，根据基本规律寻找相应特征词的匹配模式，就是本文的创新点动态挖掘模式的基础。

3.3 科研关注点挖掘算法设计

基于创新点的科研点挖掘方法建立在创新点挖掘模式的基础上，主要算法思想由三部分组成，如图3所示。（1）点的搜索与确定，（2）创新点的识别与判定，（3）点的关联关系推荐，（4）特征分类与子句提取，（5）语义关系的关联，（6）点的聚合，（7）报告生成。为了用户获取报告的方便性，报告的生成分为三部分：（1）文本格式，（2）表格格式，（3）参考文献格式。

4 挖掘结果分析

4.1 挖掘结果的输出格式

（1）挖掘结果的文本输出格式如表1所示，所举例子以“大数据”为关注点的挖掘结果的一个实例。

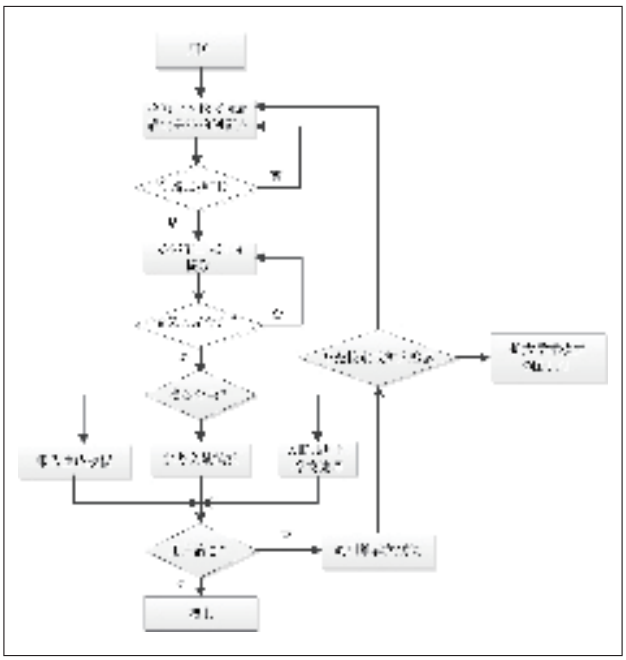

图3 科研关注点挖掘算法流图

表1 挖掘结果的文本输出格式

研究领域“计算技术、计算机技术”
作者“焦李成，高新波，李洁”[1]2004：提出一种基于克隆选择的模糊聚类新算法，通过改进距离测度函数将数值特征与类属特征相结合，从而实现具有混合属性特征数据的聚类分析，通过引入克隆选择算法（CSA）实现目标函数的全局优化。

（2）挖掘结果的列表格式输出如表2所示。列表格式为将来的语义关联推理建立了知识发现的基础。

（3）挖掘结果的参考文献输出格式如表3所示，这里所谓的参考文献是将挖掘出的关注点的文献作为文献来源，便于用户作为参考文献使用。

4.2 挖掘结果分析

4.2.1 计算领域的实验结果分析

（1）实际数据库特征词统计分析

实验选取几个不同领域的用户关注点，首先对实际数据库中特征词统计分析，对比检验表达创新点的特征

表2 挖掘结果列表输出格式

文献号	作者	提出方案	结果表明	时间	研究领域
1	焦李成	基于克隆选择的模糊聚类新算法	实现目标函数的全局优化	2004	计算技术
2	冯延蓬	基于Markov决策过程的任务调度算法	实现集群中节点的最优调度	2014	计算技术
3	贺　琪	混合云存储中海洋大数据的迁移算法	保证了数据的访问速度	2013	计算技术
4	冯延蓬	基于大数据的分布式云计算模型	解决了安防领域中的信息孤岛问题	2014	计算技术
5	刘继伟	工程机械GPS远程智能监控系统	便于工程机械设备的监管	2014	自动化

表3 挖掘结果的参考文献输出格式

[1]作者：焦李成，高新波，李洁.标题：一种基于CSA的混和属性特征大数据集聚类算法，作者单位：西安电子科技大学，刊名:电子学报，英文刊名：ACTA ELECTRONICA SINICA，年，卷(期)，页码：2004，003，357-362.

表4 “云计算”实际数据库特征词统计结果

统计结果			
关键词	拥有关键词的句子数	总句数	比例
表明	332	5692	5.83%
得到	94	5692	1.65%
方法	436	5692	7.66%
分析了	275	5692	4.83%
给出了	138	5692	2.42%
结论	15	5692	0.26%
解决	422	5692	7.41%
介绍了	190	5692	3.34%
设计了	103	5692	1.81%
提出	1061	5692	18.64%
问题	915	5692	16.08%
研究了	71	5692	1.25%
验证	166	5692	2.92%
证明	104	5692	1.83%
所有词	2698	5692	47.4%

词在不同领域的响应，观察其使用特征词的规律。

实验关注点：“云计算”，通过标题挖掘，得到实际数据库中文摘数1936条，总句数5692条，统计结果如表4所示。“统计结果”中关键词即为我们采用的特征词。

第一，分析表达提出创新点特征句所占总句子的比例：“提出，分析了，给出了，设计了，研究了”这几个特征词的句子数加起来，占总句子数的28.95%，如果把创新点的边界范围扩大，再加入“介绍了”，创新点的特征句与总句子数的比例达到32.23%。

第二，特征句占总文摘数的比例：“提出，分析了，给出了，设计了，研究了”这几个特征词句出现1648句次，占总文摘数1936条的85.12%，再加入“介绍了”，则会占总文摘数的94.93%.

第三，表现解决问题和结论的特征词比例：“表明，证明，解决，得到，验证，结论，采用”特征词，出现的数量占总句子1936条的19.9%，占总文摘的58.52%。再加入“问题”，则会占总文摘数的77.1%。

（2）挖掘结果的统计分析

根据以上方案，若挖掘创新点的特征词选取为“提出，设计了，研究了，给出了，介绍了，分析了”，得到的创新点1105条，与总文摘数1936条相比，挖掘率为1105/1936=57%。加入选择的方面词“表明，证明，解决，得到，验证，结论，采用，方法，问题”，得到的创新点454条，占总文摘1936条的23.5%。

4.2.2 医学领域挖掘结果的分析

（1）实际数据库特征词统计分析

实验关注点：“基因重组”，实际数据库中1489条。

按照“提出，分析了，给出了，设计了，研究了”特征词和“表明，证明，解决，得到，验证，结论，采用”方面词，挖掘结果不理想。我们通过对“基因重组”实际数据库统计分析得出结果如表5所示。

由统计结果可以看出，表达创新点的词汇发生了变

表5 “基因重组”实际数据库特征词统计分析结果

关键词	拥有关键词的句子数	总句数	比例
表明	452	10772	4.20%
采用	403	10772	3.74%
成功	913	10772	8.48%
得到	315	10772	2.92%
方法	1287	10772	11.95%
分析	366	10772	3.40%
构建	2034	10772	18.88%
获得	678	10772	6.29%
检测	1111	10772	10.31%
结果	1357	10772	12.60%
结论	869	10772	8.07%
利用	427	10772	3.96%
明显	408	10772	3.79%
试验	256	10772	2.38%
通过	572	10772	5.31%
效果	203	10772	1.88%
研究	832	10772	7.72%
影响	210	10772	1.95%
有效	263	10772	2.44%
证实	358	10772	3.32%
作用	486	10772	4.51%

化，基于统计信息，我们修改挖掘策略。

（2）挖掘结果的统计分析

修改挖掘策略，第一，特征词采用“构建，检测，利用，通过，研究，分析”，方面词采用“表明，采用，成果，得到，获得，结果，结论，作用，验证”。

由“提出方法”得到1249条，占总文摘数的1249/1489=83.8%，“结果表明”904条，占总文摘数的904/1489=60.7%。

4.2.3 期刊挖掘结果的分析

（1）实际数据库特征词统计分析

选择中国电子学会主办的高级学术刊物《电子学报》作为期刊论文创新点的统计分析实例。《电子学报》为中国自然科学核心期刊之一，刊登电子与信息科学及相邻领域的原始性（Original）科研成果。1962年创刊，现每年来稿2000篇左右，刊登约500篇[11]。我们对《电子学报》的10,533条文摘进行表达创新点特征词统计分析。统计分三步进行，第一步：对全部文摘进行分词、词频统计、选取词频突出的特征词；第二步：研究特征词在文摘句中出现的句法结构；第三步：统计词频突出的特征词的句子。表6给出部分统计数据。

可以看出，《电子学报》文摘表达创新点的用词非常有规律。《电子学报》文摘10,533条，使用“提出”作为引导创新点的句子达10,024条，占总文摘数的95.1%，其文摘的平均句子长度为44903/10533=4.23，具有结构完整、层次清楚、信息量大、表达准确的特点。

（2）挖掘结果的统计分析

表6 《电子学报》部分特征词的统计数据

统计结果			
关键词	拥有关键词的句子数	总句数	比例
表明	5087	44903	11.33%
方法	10606	44903	23.62%
分析	6063	44903	13.50%
获得	1101	44903	2.45%
结果	5920	44903	13.18%
提出	10024	44903	22.32%
通过	4836	44903	10.77%
问题	5572	44903	12.41%
有效	4769	44903	10.62%
所有词	29265	44903	65.17%

数据经过本挖掘软件处理后，得到10333条文摘。使用“提出，分析，通过，表明，结果，有效，获得”作为引导创新点的特征词，挖掘结果：“提出方法”得到9363条，占总文摘数的/10333=90.6%，“结果表明”7454条，占总文摘数的/103533=72.1%。

4.3 基于关联规则的关键词推荐

用户关注点的确定是一个反馈实验的过程，因此，基于关联规则的关键词推荐对用户选择关注点能起到帮助作用。系统以“大数据”为例，给出的关联规则推荐的关键词如表7和图4所示。

表7的最小支持度大于3%，最小置信度大于10%。图4显示了关联程度高的几个节点的分布。

通过对不同领域创新点的挖掘实验，证明我们建立的创造性学术成果的本体模型对科研关注点挖掘具

表7 系统给出“大数据”的关联规则推荐的关键词

推荐词1	次数1	推荐词2	次数2	共词	支持度	置信度
云计算	362	物联网	139	81	4%	22%
数据分析	383	数据挖掘	243	53	5%	14%
互联网	233	海量数据	221	31	3%	13%
big data	263	数据挖掘	243	34	4%	13%
云计算	362	big data	263	43	5%	12%
数据分析	383	云计算	362	39	5%	10%
云计算	362	移动互联网	150	36	4%	10%
数据分析	383	海量数据	221	38	4%	10%
云计算	362	数据挖掘	243	35	4%	10%

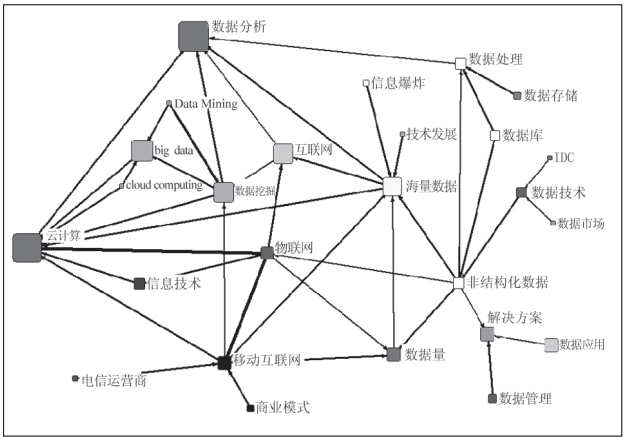

图4 关联程度高的几个节点的分布

有一定的指导意义，同时也证明我们提出的碎片化科研点挖掘，具有直接性、客观性、简便性。虽然正如我们对“大数据，云计算，基因重组……”等关注点的各种统计分析、实验和参数的调整，结果表明不同领域表达创新点的特征词不尽相同，但一般文摘写作格式为问题强调、方法提出、过程说明、结果表明四个层面。而真正的创新点就在于提出新方法和取得的结果。由此，围绕方法和结果描述的特征词和方面词的各种聚类形式的软件自适应统计分析将是优化挖掘效率的关键，这方面的实验我们已在进行。

5 结语

本文的贡献在于建立了创造性学术成果创新点的本体模型，该本体模型第一次展示了创造性学术成果的创新点的分布结构和表达方式，为提高创新点的挖掘提供了理论根据。根据这一理论基础建立了科研创新点的知识挖掘系统，系统提供了多个模板和模板之间的语义关系，为创新点报告的直接、准确、快速、简练的生成提供了有效的方式。学术成果创新点的挖掘为非结构化文本的挖掘提供了一种方法，为科研关注点的挖掘服务提供了技术手段，为科技工作者在大数据中快速准确地获得有用知识提供了帮助，同时也为信息检索向知识挖掘服务开创了一种实验方法，我衷心希望有更多的人加入探讨和实验。

参考文献

[1] 温有奎.基于碎片重组的动态数字出版模型研究[J].数字图书馆论坛,2014,119(4):2-8.

[2] 温有奎,温浩,徐端颐,等.基于创新点的知识元挖掘[J].情报学报,2005,24(6):663-668.

[3] 王昕红,凌永祥.博士学位论文创新性评议书的调查分析[J].高等工程教育研究,2004,24(3):54-56.

[4] 周露阳.论审评学术论文创新因素的指标体系[J].编辑学报,2006,18(1):68-70.

[5] 李如森,彭彩红,赵福荣.科研成果的创新性在科技论文中的表达[J].大连轻工业学院学报,2001,20(2):154-156.

[6] 李国杰.SCI不是评价科研成果的唯一标准:由论文数量高速增长引发的思考[EB/OL]. [2014-06-12]. http://www.cas.cn/html/Dir/2006/10/24/14/47/81.htm.

[7] 李怀祖.管理学科博士论文撰写探讨[J].学位与研究生教育,

2000(3):21-27.

[8] 韩客松,王永成.中文全文标引的主题词标引和主题概念标引方法[J].情报学报,2001,20(4):212-216.

[9] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6):647-657.

[10] 谭暑生.科学发现与理论创新成果评价标准[J].发明与创新综合科技,2006(1):38-39.

[11] 电子学报[EB/OL]. [2014-03-12]. http://baike.sogou.com/v10747114.htm.

作者简介

温有奎，男，1951年生，管理学博士，教授，北京万方软件股份有限公司特聘专家，研究方向：文本挖掘、语义网推理。E-mail：wykui123@126.com。
吴广印，男，1965年生，中国科学技术信息研究所研究员，研究方向：非结构数据库管理系统和中文信息检索。

Dynamic Mining of Fragmented Scientific Research Innovation Points

WEN YouKui[1], WU GuangYin[2]
（1. Beijing Wanfang Software Co., Ltd., Beijing 100038, China; 2. Institute of Scientific and Technical Information of China, BeiJing 100038, China）

Abstract: Innovation fragments excavated from the mass of information in science and technology have become a key issue in large data mining and knowledge services, which remains a problem so far in unstructured knowledge discovery. This paper presents a fragmented innovation dynamic mining method. Through the analysis of the elements and conditions of academic achievement, we establish key variables and semantic relationships of innovative elements in academic achievements, and give an ontology model of innovation in academic achievement. Based on theoretical models, we achieve a dynamic mining system of science and technology research and innovation literature debris. This method is conducive to innovation filtration of massive scientific literature. We also find the association between knowledge of scientific literature, improve the efficiency of knowledge mining literature, and help researchers access dynamic information quickly and easily.

Keywords: Fragmentation; Innovation; Ontology modeling; Dynamic mining

（收稿日期：2014-06-20）