

基于汉语科技词系统的新兴 科研趋势监测研究*

□ 刘红霞 张运良 乔晓东 / 中国科学技术信息研究所 北京 100038

摘要: 及时了解科研动态,把握学科方向对科研机构、企业和整个国家都非常重要。在科学研究的全生命周期中,新兴趋势阶段代表了科研发展方向,因此对其监测具有重要的意义。对科研新兴趋势的研究也将是数字图书馆的一种知识服务。文章在已有的新兴趋势监测模型基础之上,以期刊文献资源为基础,设计了兴趣度和有用性的计算方法及新兴科研趋势的评价标准。文章以自主开发的汉语科技词系统为例,说明了知识组织系统在主题识别、主题表示和主题确认等工作中的作用,为以汉语科技词系统为代表的知识组织系统在新兴趋势监测中的应用提供了基本解决方案。

关键词: 新兴科研趋势, 新兴趋势监测, 知识组织系统, 汉语科技词系统, 知识服务, 数字图书馆
DOI: 10.3772/j.issn.1673-2286.2010.05.003

1 引言

各学科领域的科学研究都是动态演进、不断发展的,政府、科研机构和企业的相关人员需要对科学技术的动态发展过程和总体构架有全面的了解,进而能够科学地分析科学技术发展的现状,预测科学技术发展的趋势,合理配置各种科研资源,有效评估科研成果,从而寻找出提高科研效率的方法和途径。为了实现以上目标,需要对整个科研发展情况做监测,而其中最为重要的是对新兴科研趋势的监测,具体而言包括新兴科研趋势的评估指标、评估标准和评估技术。在新兴科研趋势监测中,需要解决主题提取、主题表示和主题确认等关键问题,而引入知识组织系统将会提升处理效果。本文以中国科学技术信

息研究所开发的汉语科技词系统为例,探索基于知识组织系统的新兴科研趋势监测。

2 新兴科研趋势监测研究

2.1 科学研究的生命周期及新兴趋势监测

任何事物都有发生、发展直至消亡的过程,因此在对不同事物发展阶段的研究中,引入生命周期的概念都能够获得更清晰的区分。关于生命周期,比较有影响的是Raymond Vernon于1965年提出的产品生命周期理论;Anderson和Tushman于1990年提出了技术周期理论^[1]。Moore在技术周期理论基础上提出了新兴技术周期理

论,将新兴技术生命周期按市场占有率(Market Share, MS)和技术采用生命周期曲线的斜率划分有为始创阶段(Innovators)、分化阶段(Chasm)、爆发阶段(Tornado)、平稳阶段(Main Street)、下降阶段(Decline)和淘汰阶段(Obsolescence)等六个阶段,见图1。对于生命周期各个阶段的评估是根据市场占有率的最大值 $\xi_{\max}$ 以及市场占有率对时间的变化导数给出的。在市场占有率方面关键的几个临界点包括选取了 $\xi_{\max/2}$,在市场占有率对时间的导数上取的几个临界点包括1和0。各阶段具体的判别标准如式(1)-(6)所示^[2]。需要说明的是这里I取值是1,而 v 表示的含义是逻辑并,与我们通常的表示含义不同。

* 本文系“十一五”国家科技支撑计划“知识组织系统的集成及服务体系研究与实现”(2006BAH03B03)、中国科学技术信息研究所重点工程项目“汉语科技词系统建设与应用工程——新能源汽车领域完善及领域扩展”(2009KP01-3-2)的研究成果之一。全XML流程多样性出版服务研究。

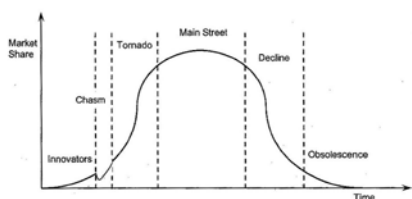


图1 Moore六阶段新兴技术生命周期图

$0 < \frac{\Delta \xi}{\Delta t} < I \cup MS < \frac{\xi_{\max}}{2}$: 始创阶段 (1)

$\frac{\Delta \xi}{\Delta t} < 0 \cup MS < \frac{\xi_{\max}}{2}$: 分化阶段 (2)

$\frac{\Delta \xi}{\Delta t} > I$: 爆发阶段 (3)

$-I < \frac{\Delta \xi}{\Delta t} < I \cup MS < \frac{\xi_{\max}}{2}$: 平稳阶段 (4)

$\frac{\Delta \xi}{\Delta t} < -I$: 下降阶段 (5)

$0 < \frac{\Delta \xi}{\Delta t} < I \cup MS < \frac{\xi_{\max}}{2}$: 淘汰阶段 (6)

本文关注的是科学技术研究的生命周期，从一定程度上看与技术生命周期有类似之处，但是判断的标准不是通过市场占有率来衡量，而是通过从科技期刊文献中衍生出的相关的指标来衡量。

本文将对某一主题（Topic）的科研分为六个阶段：（1）萌芽阶段：主题研究相关的研究开始出现，并且时间较短，影响较小，在一段时间内有一定的上升趋势。

（2）分化阶段：主题研究时间短，在数量上有一定的规模，但是发展趋势不明朗，影响较小。

（3）新兴阶段：研究的时间适中，在总量上超过一定水平，且出现持续的增长趋势，影响日益扩大。（4）热点阶段：研究时间相对较长，总体研究数量较多，影响较大，发展速度放缓。（5）衰退阶段：研究的时间相对较长，影响有较大幅度的下降，影响减弱。

（6）消亡阶段，已经研究较长时间，总体研究大幅度下降，基本上没有影响。

本文希望在尽可能早的阶段，包括新兴阶段甚至平稳发展阶段的早期（上升期）发现这种趋势，即对新兴科研趋势做监测。Apirl Kontostathis等于2003年提出新兴趋势（Emerging Trend, ET）的定义是兴趣度和有用性随着时间不断增长的主题领域^[3]。这些主题领域必然具备逐渐引起人们兴趣并被越来越多的研究者讨论的特征。这一定义是后续监测模型及本文研究的重要基础。

2.2 新兴科研趋势监测模型

Le Minh Hoang在上述Apirl Kontostathis新兴趋势定义的基础上，提出新兴科研趋势监测的模型，如图2所示^[4]。模型中各参数的含义如下：M=Model（模型），D=Documents（文本集），E=emerging trend（新兴趋势），T=topic（主题），TR=topic representation（主题描述），TI=topic identification（主题识别），TV=topic verification（主题

确认），f=interest function（兴趣度函数），g=utility function（有用性函数）。通过对文本集或语料库作处理，从而完成主题特征的抽取（识别）、主题的表示（表示为一系列时序特征）以及最终的主题确认，判别是否是新兴趋势。该模型中有三个关键模块：（I）主题识别：抽取出文本集中相应的特征，通常是名词和名词性短语，从另外一个角度看，往往是能够表达领域主题特性的术语。（II）主题描述：对于特定的主题，以表示兴趣度、有用性以及有关的综合参数的时间序列来表示主题的发展状况。（III）主题确认：主要是利用兴趣度、有用性和衍生的计算结果，来确认该主题是否处于新兴趋势阶段。

3 汉语科技词系统简介

汉语科技词系统是中国科学技术信息研究所推出的一个知识组织系统，在构成上包含协同构建平台、辅助构建工具、应用服务接口和知识内容等四个组成部分。

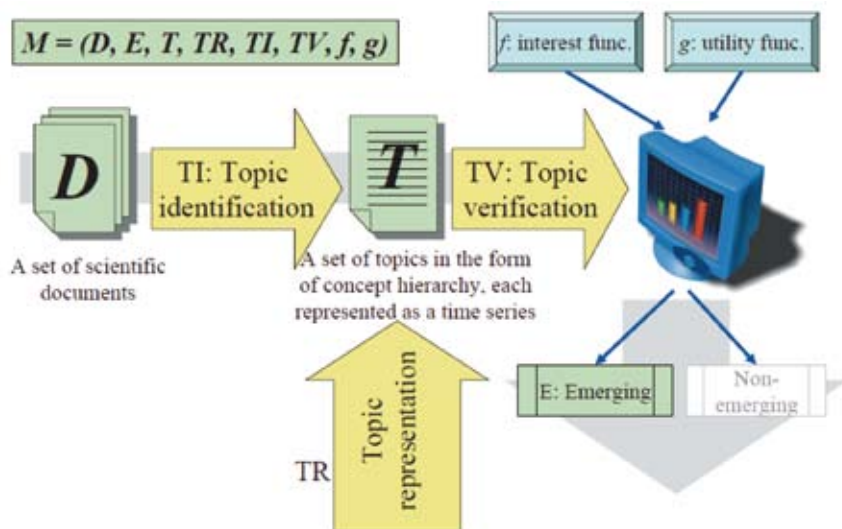


图2 新兴趋势监测模型

其中协同构建平台^[5]能够协调各加工单位、加工者的工作,实现基本的加工、审核和发布功能。辅助构建工具通过对语料分析处理,得到有关术语的频率和分布等信息,同时通过共现计算和语义计算,能够给出有关术语的关键推荐术语。此外,还可以根据领域特点发现一些未登录词。应用服务接口是针对应用,由于汉语科技词系统主要是为计算机服务的,因此提供了两层的应用程序接口^[6],包括数据访问层和业务逻辑层,其中,数据访问层解决的是对汉语科技词系统各类知识的读取功能,业务逻辑层主要针对信息检索、文本分类等几类特殊的应用做API的集成封装和优化,从而使得词系统更易用。汉语科技词系统的复杂程度介于主题与本体之间,主要知识内容设计了包括词汇中英文基本信息、定义、关系、属性、分类、概念描述等知识类型。其中关系的类型在等同、层级和相关三个基本类下又进行了细化,划分为11个一级类,52个二级类。关系分为15个一级类,76个二级类^[7]。以新能源汽车词系统为例,目前有大约5000个核心词,50000个基础词,有大约25000条关系,15000条属性。关于“双燃料天然气汽车”词条的一个示例如图3所示。本文主要使用的是关系。谈到关系,必须说明汉语科技词系统的三个特点。一是从同义关系处理上结合了同义词环和叙词表的特点,包含同义词环,如图4所示。二是根据自然语言与受控语言相结合的方式分配和布置,即知识部署在同义词环中自然语言词形最接近的词条之间,如“铜铟硒太阳能电池”作为子类部署在“太阳能电池”下,而不是同义的“光伏器

件”下。三是特定的关系和属性可以继承,即可以获得一些没有直接建立的关联。如等同关系词可以从同义词环中继承子类、父类、继承部件组成等关系,也可以继承特点、优点、缺点等属性。如对于图

4所示的同义词词环,配置于“三元催化器”的部件“网底架”“催化层”“金属外壳”等部件都可以为“三元催化转换器”“三元催化转化器”“三元触媒转换器”“三元型催化式触媒转换器”所共享。

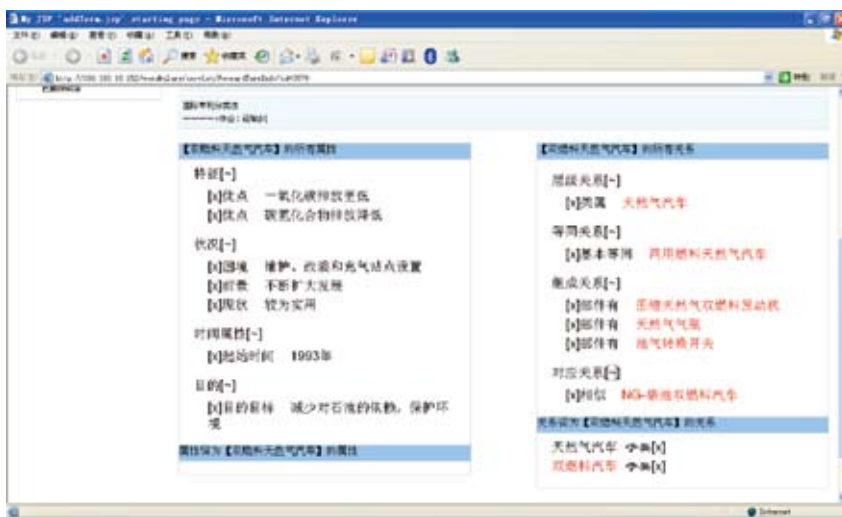


图3 新能源汽车领域汉语科技词系统示例:双燃料天然气汽车

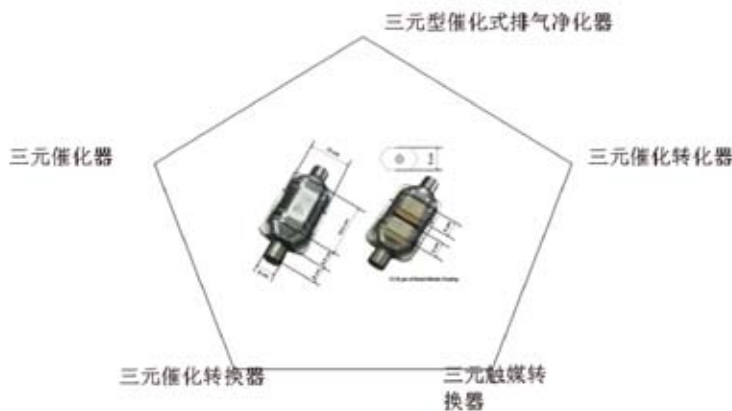


图4 同义词环示例:三元催化器

4 新兴科研趋势监测中的关键问题研究

4.1 用汉语科技词系统解决主题识别问题

主题识别的第一步是将文献文本做分词处理。由于通常我们能

够获取的都是二次文献,所以我们分词的范围是标题和摘要。从应用的角度,需要尽可能地把术语分出来。为此,我们的词库采用领域词库加上通用词库的方式。在分词过程中要优先领域词库,这里我们可以把汉语科技词系统等知识组织系统中的词条(包含核心词和基础

词) 导出来作为领域词典。

常用的分词算法有三大类, 分别是基于词典的机械匹配算法、基于统计的算法, 以及词典和统计相结合的算法^[8]。由于不具备标注语料条件, 因此本文采用基于词典机械匹配的方法, 并在分词后作后续处理, 以弥补机械匹配方法的不足。在匹配分词过程中, 不可避免产生一些碎片, 这些碎片中可能包含着我们关注的主题, 因此对碎片的处理关系到主题识别的成败。在碎片处理上可以采取三种策略, 第一种是对于单篇或者多篇文献范围内相同的连续碎片多次反复出现, 可以认定为是固定短语或者新词; 第二种是采取碎片展示并人工查看确认的方式, 将确定的词和短语识别出来; 第三种是综合频率和人工校对的方式。对于本文来说, 仅仅出现一次的主题没有意义, 因此本文采用第一种策略。

4.2 主题表示中的主题词筛选

在新兴趋势监测研究中, 需要把主题表示为时间序列, 而一个单一的主题, 需要以若干的主题词来表示, 什么主题词是我们需要的和关注的主题词, 就涉及对主题词的权重和筛选问题, 最后会得到一个入选主题词序列。在研究新兴趋势的时候, 很多可能的新兴主题词频次不太高, 在上节的识别过程中, 可能会被遗漏。当前获得的主题词的维度较大, 为了进一步降低主题词的维度, 可以采用不同的降维方式, 如根据词频(TF)或者逆文档频率(IDF), 或者二者乘积(TFIDF)选取阈值, 取得权重后TFIDF值更大的主题词。TFIDF权

重公式如式(7)所示, 其中 $f_{\text{tfidf}}(t_i)$ 表示主题词 t_i 的TFIDF权重值, m_i 表示 t_i 的词频, N 表示文档集的文档总数, N_i 表示包含主题词 t_i 的文档数量。

$$f_{\text{tfidf}}(t_i) = m_i \times \log\left(\frac{N}{N_i}\right) \quad (7)$$

此外, 我们还有一些特征有助于确定主题词, 这些主题词可能相应的计算指标没有达到筛选阈值, 但是却非常重要。具体的识别特征至少有以下几类, 这些识别出来的主题词需要进入入选主题词序列并加以标记, 做特殊处理。

- A. 附带定义等解释说明的;
- B. 后面有括号外文说明的;
- C. 特殊词汇判断, 如有前置的“提出”“引入”等词汇。

在一个新的科学或者技术话题开始研究的时候, 不同的研究者可能从不同的角度来对其命名, 如“计算语言学”与“自然语言处理”。这种普遍存在的不规范的命名, 自然会削弱基于频次的主题词筛选的准确性。因此需要进行同义合并, 以减少主题词数量, 准确确定主题词词频, 进而确定主题发展状况。在归并的时候, 关键是相似度计算。我们可以基于字面匹配, 基于编辑距离计算, 也可以基于一些其他的匹配方式, 并且可以利用通用的语义资源。这里我们利用通用语义资源, 结合汉语科技词系统中的一些同义关系和双序列比对算法来进行相似度计算和同义词识别^[9]。通过汉语科技词系统中的同义词环, 我们可以得到一些同义要素。如通过图4所示的同义词环, 我们可以得到“催化剂-催化转化器”“催化剂-催化转换器”“转换器-转化器”“触媒-催化”等同义要素, 这些要素如果看作单个的词, 则可能不是同义, 如触媒和催

化, 但是在构词上可能是同义的。而且由于我们的这些同义要素和处理的语料是同源的, 因此其适用性很好。

这里需要说明的是通过双序列比对识别出来的同义词主题词, 如果单独计算已经进入入选主题词序列, 则本计算的结果仅在下节扩展指标计算的时候使用; 如果单独计算没有进入入选主题词序列, 则合并计算后达到入选最低指标, 可以进入入选序列。通过以上处理, 通常最后的入选序列会略大于开始预设的数量阈值。

4.3 指标体系和判断标准

根据文献[3]和文献[4]进行新兴的研究趋势判断, 需要利用兴趣度和有用性两个指标, 其中兴趣度主要考虑主题词频率及出现位置, 有用性主要用在摘要中出现的而在标题和关键词中没有出现的主题词的相对值来表示, 也就是说如果论述的核心并不是本主题, 但是本主题对应的词大量出现, 则体现了本主题对其他主题的较高的支持度, 因此会提高本主题有用性。从计算的角度并结合汉语科技词系统, 最小的主题仅仅包含主题词本身, 也用主题词本身来表征, 更多的情况是主题词可以用它本身、同义词环和相关的子类、部件、实例、材料或者成分等主题词共同构成。表征虽然还是这个主题词本身, 但是在计算的时候需要考虑大量的相关主题词。主题具体的计算公式如式(8)-式(20)所示, 其中 $f_{\text{exp}}(t_i)$ 表示由主题词 t_i 表示的主题的全部扩展主题词, $f_{\text{eq}}(t_i)$ 表示主题词 t_i 对应的同义词环的全部主题词, 包括通过上节介绍的方法未归并的新同义

词词环, $f_{inc}(t_i)$ 表示主题词 t_i 对应的子类、部件、实例、材料或者成分等全部主题词。 $f_{int}(t_i, T_X)$ 表示在 T_X 时间段内对主题词 t_i 代表主题的基本兴趣度度量函数, 也即 T_X 时间段内对主题词 t_i 的兴趣度度量函数。 $f_{intex}(t_i, T_X)$ 表示在 T_X 时间段内对主题词 t_i 代表的主题的扩展兴趣度度量函数, 也即 T_X 时间段内对主题词 t_i 及其扩展主题词的兴趣度度量函数。其中 t_{ipt} 表示一个在文档 d_p 的关键位置(标题、关键词、重要标识词)的主题词 t_i 的频率, t_{ipo} 表示一个在文档 d_p 的其他位置的关键词 t_i 的频率, α 是一个可调节参数, 并且通常取一个 ≥ 5 的数, $f_{imp}(d_p)$ 表示文档 d_p 所在期刊的影响因子, $|f_{exp}(t_i)|$ 表示 $f_{exp}(t_i)$ 的元素数量, β_{ki} 是一系列可调参数, 通常反映了不同的扩展主题词对主题的权重, 在不考虑关系类型的情况下, 可以认为所有的 $\beta_{ki}=1$ 。 $f_{use}(t_i, T_X)$ 表示在 T_X 时间段内对主题词 t_i 代表主题的基本有用性度量函数, $f_{useex}(t_i, T_X)$ 表示在 T_X 时间段内对主题词 t_i 代表的主题的扩展有用性度量函数。 t_{ipos} 实际上是 t_{ipo} 的一个子集, 过滤掉在关键位置和其他位置都出现的情况的频次信息。 F 表示所有时间段内任意主题 t_i 的基本兴趣度度量结果集合, F' 表示所有时间段内任意主题 t_i 的扩展兴趣度度量结果集合, G 表示所有时间段内任意主题 t_i 的基本有用性度量结果集合, G' 表示所有时间段内任意主题 t_i 的扩展有用性度量结果集合, $\|F\|$ 表示集合 F 中的最大值。 $f_{ire}(t_i, T_X)$ 和 $f_{ireex}(t_i, T_X)$ 分别表示基于基本度量方法和扩展度量方法的趋势函数, γ 和 γ' 是可调价的参数, 其数值代表评价对兴趣度和可用性的权值, 在认为两者同等重要的情况下, 可选取 $\gamma=\gamma'=1$ 。

$$f_{exp}(t_i)=\{t_i\} \cup f_{eq}(t_i) \cup f_{inc}(t_i) \quad (8)$$

$$f_{eq}(t_i)=\{t_j|t_j \text{与} t_i \text{在同义词环内}\} \quad (9)$$

$$f_{inc}(t_i)=\{t_j|t_j \text{是} t_i \text{子类、部件、实例、材料或成分}\} \quad (10)$$

$$f_{int}(t_i, T_X)=\sum_{p=1}^q (\alpha t_{ipt}+t_{ipo}) \times f_{imp}(d_p) \quad (11)$$

$$f_{intex}(t_i, T_X)=\sum_{k=1}^{|f_{exp}(t_i)|} \beta_{ki}+f_{ipt}(t_i) \quad (12)$$

$$f_{use}(t_i, T_X)=\frac{\sum_{p=1}^q t_{ipos} \times f_{imp}(d_p)}{\sum_{p=1}^q t_{ipt} \times f_{imp}(d_p)} \quad (13)$$

$$f_{useex}(t_i, T_X)=\sum_{k=1}^{|f_{exp}(t_i)|} f_{use}(t_k) \quad (14)$$

$$F=U_{T_X}\{f_{int}(t_i, T_X)\} \quad (15)$$

$$F'=U_{T_X}\{f_{intex}(t_i, T_X)\} \quad (16)$$

$$G=U_{T_X}\{f_{use}(t_i, T_X)\} \quad (17)$$

$$G'=U_{T_X}\{f_{useex}(t_i, T_X)\} \quad (18)$$

$$f_{ire}(t_i, T_X)=\frac{f_{int}(t_i, T_X)}{\|F\|} + \gamma \frac{f_{use}(t_i, T_X)}{\|G\|} \quad (19)$$

$$f_{ireex}(t_i, T_X)=\frac{f_{intex}(t_i, T_X)}{\|F'\|} + \gamma' \frac{f_{useex}(t_i, T_X)}{\|G'\|} \quad (20)$$

结合 $T_X \in \{T_1, T_2, \dots, T_n\}$ 的 $f_{ire}(t_i, T_X)$ 和 $f_{ireex}(t_i, T_X)$ 可以得到一系列折线, 通过拟合可以得到各主题的发展曲线, 即类似图1的全部或部分。判断主要利用曲线导数以及兴趣度和有用性绝对值来进行。基于基本度量和扩展度量的判别公式, 分别如式(21)-(23)和式(24)-(26)所示, 其中 δ 、 ε 、 I 和 δ' 、 ε' 、 I' 都是可调节参数。其含义是当一个主题其兴趣度和有用性都达到一定水平, 并且其趋势曲线处于上升阶段, 但是上升速度不超过一个固定阈值的情况下, 可以认为是处于新兴趋势阶段。各参数可以利用已有的数据集, 通过机器学习的方式给出来。

$$f_{int}(t_i, T_X) > \delta \quad (21)$$

$$f_{use}(t_i, T_X) > \varepsilon \quad (22)$$

$$0 < \frac{df_{ire}(t_i, T_X)}{dT_X} < I \quad (23)$$

$$f_{intex}(t_i, T_X) > \delta' \quad (24)$$

$$f_{useex}(t_i, T_X) > \varepsilon' \quad (25)$$

$$0 < \frac{df_{ireex}(t_i, T_X)}{dT_X} > I' \quad (26)$$

5 问题讨论

本文主要使用期刊文献进行新兴科研趋势监测, 但是期刊文献通常有3-12个月的出版周期, 甚至可能会更长, 从一定程度上讲, 我们的分析是滞后于实际发展的。解决这个问题基本途径是引入一些周期更短的文献来源, 如学术会议和Web资源。引入新的资源, 并不意味着放弃原有资源, 因此需要讨论综合利用期刊、学术会议和Web资源实现监测的问题。此外, 如果引入到统一的框架, 需要对学术会议和Web资源作影响因子的考虑, 而且有一些会议文献没有关键词, Web资源可能既没有关键词, 又没有摘要, 其处理策略和计算方法都需要调整。

本文采用的兴趣度和有用性的度量方法是比较简单的方法, 虽然在利用知识组织系统上作了扩充, 但是还有待进一步研究。从理论上还可以增加多个兴趣度和有用性度量的要素, 如作者单位、基金支持状况等, 但是其影响究竟如何还有待于继续探讨。理论上, 如果作者所在单位属于比较领先的科研机构, 或者有基金支持, 那么可能会为其研究的主题带来更高的权重; 从另一个角度看, 如果作者所处单位属于一般的科研机构或者没有基金支持, 则其论文发表难度大, 能够发表, 说明其水平高, 应该有更高的权重, 而两者还存在一定的矛盾之处。

在新兴科研趋势的判断过程

中,对于式(23)和式(26),我们希望寻找到一个尽可能小的 $I_{\min}$ 或 $I'_{\min}$,从而尽早发现新兴趋势。在现实中,也许我们并不需要明确地自动给出处于新兴科研趋势的主题,而是在一个标准更低的、数量更大的范围内给出参考,可以由专家进行判断,从而可能有助于更早发现趋势。此外,本文虽然采取一些办法对固定主题词的筛选方式进行补充,但是仍然存在一定的主题

词被过滤掉了,因此在计算能力满足的情况下要尽可能扩大入选主题词的范围。

6 结论

新兴科研趋势研究对于国家、企业、学者的科研跟踪和科研开展都是非常必要的。本文基于科研主题的生命周期理论以及检测模型,结合汉语科技词系统的资源,从兴

趣度和有用性角度提出了基于汉语科技词系统的监测方法。在分词阶段、同义词归并阶段和主题扩展阶段都不同程度地使用了汉语科技词系统,从理论上阐明了方案的可行性和以汉语科技词系统为代表的知识组织系统在新兴科研趋势监测上的应用前景。本文的研究还存在一些待解决的问题,对这些问题的研究将是本文的下一步工作。

参考文献

- [1] 王爱国,何庆梅. 基于生命周期理论的高技术企业技术创新战略研究[J]. 山东理工大学学报(社会科学版),2006,22(3):35-39.
- [2] MEADE P T, RABELO L. The technology adoption life cycle attractor: Understanding the dynamics of high-tech markets [J]. Technological Forecasting and Social Change, 2004, vol.71; no.7: 667-684.
- [3] KONTOSTATHIS A, GALITSKY L M, POTTINGER W M, ROY S, PHELPS D J. A Survey of Emerging Trend Detection in Textual Data Mining [M]// BERRY M W. A Comprehensive Survey of Text Mining. Springer-Verlag, 2003: 185-219.
- [4] HOANG L M. Emerging Trends Detection from Scientific Online Documents [D]. Japan Advanced Institute of Science and Technology, 2006.
- [5] [EB/OL]. [2010-3-5]. <http://www.vocgrid.org/vocabulary/>.
- [6] 史新. 科技词汇知识的网络服务模式研究与原型实现[D]. 北京:中国科学技术信息研究所,2009.
- [7] 张运良,朱礼军,薛春香,乔晓东. 汉语科技词系统构建分析报告[R]. 中国科学技术信息研究所,2009.
- [8] 何国斌,赵晶璐. 汉语文本自动分词算法的研究[J]. 计算机工程与应用,2010,46(3):125-127,130.
- [9] XU S, ZHU L J, QIAO X D, XUE C X. A Novel Approach for Measuring Chinese Terms Semantic Similarity based on Pairwise Sequence Alignment[C]. The Fifth International Conference on Semantics, Knowledge and Grid, Zhuhai, China, 2009.

作者简介

刘红霞 (1983-), 女, 中国科学技术信息研究所硕士研究生, 主要研究方向为新兴趋势监测及其指标体系建设。通讯地址: 北京市海淀区复兴路15号中国科学技术信息研究所 100038

张运良 (1979-), 男, 中国科学技术信息研究所/信息技术支持中心, 博士, 助理研究员, 研究方向: 知识组织、自然语言处理, 已发表论文20余篇。通讯地址: 北京市海淀区复兴路15号中国科学技术信息研究所 100038。E-mail: zhangyl@istic.ac.cn

乔晓东 (1965-), 男, 研究员, 硕士生导师, 中国科学技术信息研究所/信息技术支持中心主任, 中国互联网协会理事, 中国情报学会计算机应用分会副主任委员, CILIP会员, 研究方向为信息资源管理和信息服务, 发表论文50余篇, 为本文通讯作者。通讯地址: 中国科学技术信息研究所信息技术支持中心 100038

The Study on Emerging Research Trend Detection Based on Chinese S&T Vocabulary System

Liu Hongxia, Zhang Yunliang, Qiao Xiaodong / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: It is very important for research institutions, enterprises and the governments to master the research trend. In the life cycle of research, emerging trend phase is of great concern because it represents the direction of research and development. It will also be a knowledge service for future digital library. In this paper, with existing emerging trend detection model and journal resources, we propose the computation algorithms of interest and utility and also evaluation criteria of emerging research trend. And we explain the role of KOS in topic representation, topic identification and topic verification of ETD systems with the Chinese S&T vocabulary system of ISTIC as an example, and provide basic solutions for ETD with KOS.

Keywords: Emerging research trend, Emerging trend detection, Knowledge organization system, Chinese S&T vocabulary system, Knowledge service, Digital library

(收稿日期: 2010-04-15)