基于科技文献中词语的科技发展监测方法研究

Study on Science and Teclnology Watch Based on Terms from S&T Documents

魏晓俊[1,2]

(1.中国科学院研究生院 北京 100049; 2.中国科学院文献情报中心 北京 100080)

摘 要 词是科技文献的基本内容单元。因此通过词来对科技文献进行研究分析,可揭示科技的发展动态。目前这些方法主要包括:词频分析方法、基于词网络关系的共词分析,Kleinberg 突发词监测方法、短语差异分析方法。

关键词 词频统计 共词分析 突发监测 短语差异 科技文献

词是科技文献中承载科技概念的最小单位,分析科技文献中词所带来的信息变化,可以把握科技发展动态。根据对词分析的角度不同,本文将它们分为词频分析方法、基于词网络关系的共词分析方法、基于词频变化率的 Kleinberg 突发监测方法、基于短语差异的分析方法。通过研究各种方法的设计原理和机制,科技情报研究人员可以根据不同的分析目的来进行选择。

1 词频分析方法

1.1 词频分析方法的含义 词频分析方法是利用能够揭示或表达文献核心内容的关键词或主题词在某一研究领域文献中出现的频次高低来确定该领域研究热点和发展动向的文献计量方法[1]。

1.2 词频统计方法的基本设计 在文献数据库中,作者和数据库标引员采用关键词来标识文章中的主要研究内容,如果关于某一问题的研究多,则相应的关键词出现次数也多。因此选出高频关键词作为研究的热点主题,并进行深一步研究。高频词阈值的确定主要有两种方法:一种是结合研究者的经验在选词个数和词频高度上平衡,该方法具有一定的主观性;另一种是结合齐普夫第二定律辅助判断高低频词的界限[2]。通过揭示高频词在各年中的分布变化,可揭示学科的发展热点和脉络,包括:统计各年收录的高频词的差异;统计各个高频词在各年中的频次变化和频次排序变化;与内容分析相结合,分析高频词所承载的科技内容,将这些关键词进行相应的分类与组织,揭示领域研究的热点主题。

1.3 词频统计的优势与不足 采用词频统计方法监测科技发展的优点是操作相对简单,揭示科技发展的方式比较直接。但是仍存在着一些不足主要表现在以下几点:a.词频阈值的确定比较主观,不同的研究者有不同的标准,可能导致研究结果不一致;b.词频阈值通常是固定的,而词的出现频次具有波动性,因此某些研究从长期来看是属于热点,但是可能在某一年的波动略在词频阈值下方,有可能被忽略掉,导致分析的误差;c.关键词带有辅助检索的任务,它主题范围一般比较大,比如:在关于知识管理研究热点分析中[3],抽取出来的高频词中包括知识管理、知识、企业、管理等泛义词,在揭示领域深度和微观层变化上还有一定的差距;d.高频词在形成研究主题的过程中,需要较多的人工干预,需要专家根据自己的知识背景将高频词分成特定的研究主题。

2 共词分析方法

2.1 共词分析的含义 共词分析由 Callon 等人于 20 世纪 70 年代末到 80 年代初提出,其原理是对一组词两两统计它们在同一篇文献中所出现的次数,以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化[4]。

2.2 共词分析的基本设计 随着信息可视化研究的发展,共词分析常采用 MDS(MultiDimensional Scaling,多维度标尺分析)、SOM(Self - Organized Mapping 自组织映射)、PFNet(PathFinder NETwork,路径搜寻网络法)展现词与词之间的网络结构,并将网络中存在着关系紧密的词的集合提炼成主题,进而可研究主题与主题之间的关系,以及主题的内部结构。

为研究主题的非线性发展,揭示各个研究主题在领域中所处的位置以及自身的发展程度,Law 在 1988 年提出的战略图(strategic diagram),它是共词分析中的常用工具。战略图是以中心度为横坐标,密度为纵坐标,两个轴的中位数或平均数为原点构建二维图[5]。

中心度(Centrality)量度一个主题和其它主题的相互影响程度。一个主题与其它主题联系的数目和强度越大,这个主题在整个研究工作中就越趋于中心地位。对于特定的主题,中心度的计算可以通过该主题的所有主题词或关键词与其他主题的主题词之间连接的强度来核算。这些外部连接的总和、平方和的开平方等都可以作为该主题的中心度。密度(density)是量度字词聚合成主题的联系强度,也就是该主题的内部强度。它表示该类维持自己和发展自己的能力。某一主题的密度的计算可以有多种方式,首先计算本主题中每一对主题词或关键词在同一篇文献中同时出现的次数,通过计算这些内部连接的平均值、中位数或平方和,可得出这个主题的密度。

构建的战略图共分为四个区域,见图 1。在这四个区域中,

作者简介:魏晓俊,女,1982 年生,硕士研究生,研究方向为学科情报。

分别代表着主题发展的四个状态，其中第一象限中的主题是领域的中心，同时发展较为成熟，是当前的研究热点；第二象限中的主题在研究中处于中心地位，但是发展还不够成熟，具有潜在的增长性；第三象限中的主题未具规模；第四象限中的主题研究比较集中，但是尚不处于领域的中心位置，尚未引起广泛的关注。

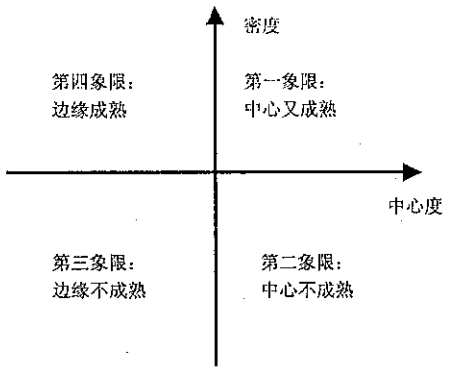

图1 战略图

2.3 *共词分析的优势与不足* 共词分析以词网络关系所形成的主题为监测对象，能够较好地揭示科技领域中主题的发展状态。它可以减少专家在主题形成过程中的参与。但是由于共词分析为了达到较好的聚类效果，通常选用高频词，使得某些尚处于低频的热点不能早期发现出来。

3 基于词频变化率的Kleinberg突发词监测方法

3.1 *Kleinberg突发词监测方法介绍* Kleinberg于2002年提出话题的突发监测(burst detection)算法[6]，它关注焦点词——相对增长率突然增长的词。Kleinberg突发监测算法认为话题的报道数量不是平滑增长，而是在不同水平之间跃迁。Kleinberg基于概率机对不同时间段上词出现的频次进行建模。概率机的状态确定了某时间点上词出现频次的期望值，而概率机的状态改变由概率模型控制。词突发时，概率机处于高频状态。只要给定了文献集合，确定状态的个数、状态差异的大小，以及状态改变的成本，利用Viterbi动态建模法对状态改变的概率模型求最优解，便可得出概率机状态变化的最优时序序列。

突发词监测与统计单位时间里(通常是一年里)词出现的频次不同。前者主要是从关注词自身的发展变化出发，关注单个词发展的阶段性，而后者主要是对领域中各个词的增长势头进行比较。由于科技领域中的局部热点变化不一定会引起全领域的注意或者研究，但又是领域发展中不可缺少的部分，比如关于某学科的教育研究，不一定会引起全领域范围的讨论，但是它的研究本身也会不断发展。因此基于单个词的词频增长率变化更有可能涉及到领域局部热点的变化。

突发词监测热点与在一个时间段里的词频阈值监测热点的出发点不同。前者认为增长势头不断加强的词是大家越来越关注的，它正在聚集越来越多的力量，在揭示科技发展上更具及时性，虽然它还未达到词频阈值的要求，但是未来的发展势头好，这些词有可能是低频词但却具有情报意义。而后者是在当前科学领域研究中，已经聚集了较强力量的研究，是已经反映出来的热点。

3.2 *针对科技文献的批量信息突发词监测算法设计* 针对科技文献是批量式报道，因此通常采用批量信息的突发词监测算法。它的设计如下：

在批量信息中，有的信息与考察话题相关，有的信息则无关，因此话题的突发是指它在批量信息中所占的报道成分显著增加；换句话说，在某个特定领域的相关论文集合中，某个话题占该领域的论文数的比例增加时，它越来越成为领域的焦点。现假定有 n 批数据，第 t 批数据一共有 d_t 篇文献，其中有 r_t 篇相关的文献。令 $R=\sum_{t=1}^{n} r_t$，R 是 n 批数据中与某个主题相关的所有论文数；$D=\sum_{t=1}^{n} d_t$，D 是 n 批数据中领域内的所有论文数。概率机为 $B_{s,\gamma}^{*}$，其中 s 是规模参数，控制概率机状态的分辨程度，即状态差异的显著程度，当 s 越大，两个状态的差异越大，词的突发程度高。γ 是控制概率机状态改变的代价参数，通常其缺省值为1。令某个状态为 $q_i(i\geq 0)$，相应的话题占文献集合的比例为 p_i。令 $p_0=\frac{R}{D}$ 是基状态；$p_i=p_0 s^i$ 是第 i 个状态下话题占总文献数的比例，但要满足 $p_i\leq 1$，所以 $B_{s,\gamma}^{*}$ 是有限状态概率机。因此将其限制在 k 个状态，那么就是 $B_{s,\gamma}^{k}$。假定状态的出现序列为 $q=(q_{i1},\cdots,q_{in})$，其中 q_{in} 表示话题在第 n 批数据中的状态为 q_i。在状态 q_i 下，文献流中话题的出现次数是服从概率为 p_i 的二次多项式分布，即 $\binom{d_t}{r_t} p_i^{r_t}(1-p_i)^{d_t-r_t}$。构建文献流出现序列 q 的贝叶斯条件式，得当第 t 批数据时，概率机仍然处在 q_i 的成本是 $\sigma(i,r_t,d_t)=-\ln[\binom{d_t}{r_t} p_i^{r_t}(1-p_i)^{d_t-r_t}]$。而从 q_i 到 q_j 的跃迁成本是 $\tau(i_t,i_{t+1})=(j-i)\gamma l_{nn}$。通过最小化成本，可以得出最优的状态序列，其中需要设定 k,s,γ 这3个参数的值。在批量信息分析中，重点是为了找出具有突发性质的词而不是揭示词突发的层次结构，所以 k 取值一般为2。为了对突发词的突发程度进行排序，设计了突发权重指数：$\text{weight}=\sum_{t=i_1}^{i_2}(\sigma(0,r_t,d_t)-\sigma(1,r_t,d_t))$，也就是说权重等于从非突发状态跃迁到突发状态的成本，从某种意义上说权重越大，突发的可信度越高。

3.3 *突发词监测算法的优缺点* 批量信息的Kleinberg算法的优点包括：a.可根据突发权重指数进行排序，识别出显著的突发事件。b.可以对所有的词进行突发分析，找出具有学科主题的词。Kleinberg算法只需对词语做少量的预处理，去掉非字母词，统一词语的大小写。由于停用词、泛义词发生突发情况少，Kleinberg算法发现科技领域中的主题词具有一定的抗干扰性。c.确定状态持续的时间。Kleinberg算法根据突发状态的时序序列，确定各个状态的持续时间。d.Kleinberg算法能够揭示词目前是处在突发还是非突发状态，对词的发展分析具有动态性和历史性。

Kleinberg算法也存在一些不足：首先Kleinberg算法中最优序列的确定受参数 k,s,γ 的影响，而参数确定是具有主观性的。同时Kleinberg算法是基于时间段的，也就说是基于回溯的，它的分析需要一定的时间积累。

3.4 *突发词监测算法的应用* 已有学者将Kleinberg算法应

用于科技情报研究领域。陈朝美在设计 CitespaceII 时，认为新兴的处于上升阶段的焦点词更能揭示学科的前沿问题，因此采用 Kleinberg 的突发算法来进行识别[7]。印第安那大学信息学院积极推动 Kleinberg 算法在情报研究中的作用，取得了一系列的成果，包括将 Kleinberg 算法嵌入到 Infovis cyberinfrastructure 框架中。Weimao Ke 和 Katy Börner 将突发词按照突发时间的先后顺序排序，揭示出 ACM 数据库中关于信息可视化研究热点的演变过程[8]（见表 1），从用户界面（user interface）和人类因素（human factor）的研究发展到数据可视化（data visualizatiou）研究，而信息可视化（information visualization）从 1998 年起，一直处于持续突发的状态。Ketan Mane 和 Katy Börner 将突发词引入到共词分析中[9]。

表 1 突发词分析

词	突发权重	突发时间段
Data visualization	3.7	1994～1995
focus+context	4.29	1999～2002
hierarchy	3.95	2000～2002
human factors	3.42	1983～1994
information visualization	13.083	1998～现在
user interface	3.457	1983～1991

4 短语差异分析

4.1 *短语差异分析含义* 短语差异分析是通过研究词如何组织成短语，比较短语各个组成部分的差异，由此来比较短语所揭示的科技概念的差异，这些差异代表着科技概念间的扩展、深化和演变。

4.2 *短语差异分析的设计* TermWatch 系统是由 Fidelia Ibekwe－SanJuan 和 EricSanJuan 基于短语差异分析而设计的科技发展监测系统[8]。该系统不仅根据语言分析来抽取词和短语，还以语言关系作为主题聚类的基础，通过揭示短语之间的差异关系，监测新词，判断新的研究热点。它的分析步骤包括：

a. 利用 INTEX 工具包，根据形态－句法的分析方法从文献数据库的全文中抽取候选词。利用介词、定冠词、逗号、连词等进行分割，首先根据抽取复杂的名词序列，然后将其分解成更简单的名词性短语，直到满意为止。抽取出来的短语一般是中粒的，能够揭示简单概念之间的组合关系。

b. 利用词频、语言分析和专家相结合，精选候选词。

c. 进入 TermWatch 系统核心模块，它包括三个子模块：识别短语之间的差异关系；根据差异关系聚类；利用 AISEE 可视化。其中最重要的是前面两个子模块。识别短语之间的差异关系子模块中将短语之间的差异关系定义为两种：扩展关系和替代关系。其中扩展关系又包括左扩展、右扩展、左右扩展和中间插入；替代关系包括修饰词替代和中心词替代，它们的定义分别如下：

M 代表修饰性词组，m 代表修饰词，h 代表中心词，t_1 和 t_2 代表短语。

如果 $t_1 = Mh$，$t_2 = M'm'Mh$；则 t_2 是 t_1 的左扩展。例如：t_1 为 bread manufacture，t_2 为 French bread manufacture。

如果 $t_1 = Mh$，$t_2 = MhM'h'$；则 t_2 是 t_1 的右扩展。例如：t_1 为 bakers' yeast，t_2 为 bakers' yeast preparation。

如果 $t_1 = M_1mM_2h$，$t_2 = M_1mm'M'M_2h$；则 t_2 是 t_1 的中间插入。例如：t_1 为 bread improvement，t_2 为 bread flavour improvement。

如果 $t_1 = M_1mM_2h$，$t_2 = M_1m'M_2h$；则 t_2 是 t_1 的修饰词替代。例如：t_1 为 protein content of bread，t_2 为 protein content of bun。

如果 $t_1 = Mmh$，$t_2 = Mmh'$；则 t_2 是 t_1 的中心词替代。例如：t_1 为 frozen dough baking，t_2 为 frozen dough method。

基于差异关系的聚类，它包括三个层次：

第一级是建立差异关系图。差异关系图给出了文献数据库中词与词之间的差异关系。图中利用虚线代表替换关系，箭头代表扩展关系，箭头指向扩展结果。在差异关系图中，短语与短语之间的关系强度是由该种关系的数目的倒数确定。比如在差异关系图中，右扩展关系只有 2 个，右扩展关系强度为 1/2。

第二级聚类是基于 COMP 关系。COMP 关系是短语中修饰词的各种差异关系，包括左扩展、中间插入、替换修饰词。COMP 关系将拥有相同中心词的短语聚集成一个成分，它们属于同一个概念簇。第二级聚类的重点是强调概念簇之间的关系。

第三级聚类是基于 CLAS 关系。CLAS 关系是短语的中心词差异关系，包括右扩展、左右扩展、替换中心词，在此基础上形成的类，类有同样的修饰词但是中心词不一样。中心词差异表明了中心词的变化，因此很有可能是概念的转变。

短语差异聚类结果的分析可利用下述指标：

类的大小：TermWatch 系统中没有事先限制类的大小，因此类的大小可以表明研究主题所包含的范围。

中心度：根据连接的数目确定中心主题和边缘主题。

差异指数：通过揭示词的差异关系变化，来揭示主题的动态发展。

内部差异指数：$\mathrm{Int}_j = \frac{R_j}{T_j}$，$R_j$ 是类 j 中的所有差异关系强度总和，T_j 是类 j 中所有词的个数。

外部差异指数：$\mathrm{Ext}_j = \frac{T_j^-}{T_j} \times \frac{T_j^+}{T}$，$T_j^-$ 是与类 j 外的词存在着差异的类 j 中的词的个数，T_j^+ 是与类 j 存在着差异关系的类 j 外的词的个数，T_j 是类 j 中所有词的个数，T 是所有的词的个数。

转化指数：$\mathrm{TRANS}_{ij} = \frac{V_{ij}}{N_{ij}^2 + 1}$，$\mathrm{TRANS}_{ij}$ 是类 i 指向类 j 的转化指数，V_{ij} 是类 i 和类 j 差异关系强度总和，N_{ij} 是类 i 和类 j 拥有的相同的词。

4.3 *短语差异方法的优点与不足* 短语差异方法由于是基于语言特征，而不是统计特征，所以更容易把低频的新热点揭示出来。短语差异方法的聚类是根据短语的中心词和修饰词的差异关系来构建的，所以聚类结果的主题比较明确。基于短语的分析下一步应该向监测词汇之间的语义关系发展，而不仅仅是词的形式变化。

5 总结

本文分析了利用词频分析、共词分析、突发词 （下转第 39 页）

业为主的城市更多的是要考虑环境因素这个指标，如果过分地去考虑经济因素，对这个城市的长远发展将极为不利。不同的地区之间在建立经济信息效益评价指标体系时也是可以相互借鉴的，特别是那些经济发达地区出现的问题众多、复杂而且具有典型性，对一些经济发展欠发达地区的经济信息效益评价指标体系的建立有很好的借鉴指导作用。

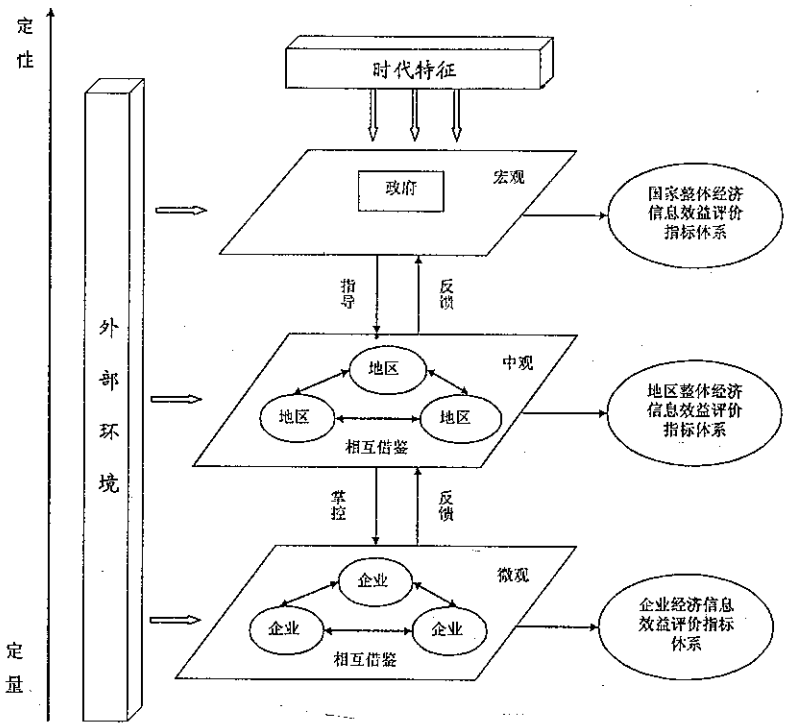

图1 经济信息效益评价指标体系模型

4.3.3 微观层次。在针对处于某个地区的不同企业，企业经济信息效益评价指标体系的建立不仅仅只是以自我为中心，其经济信息评价指标体系的建立要做到以国家的基本框架为指导，依附于地区的经济信息效益评价指标体系，反映企业的各自特征，要将这几个方面综合起来考虑，形成各企业的经济信息效益评价指标体系。各企业的经济信息效益指标体系制定也要相互借鉴，尤其是一些知名企业，他们的实践为企业经济信息效益评价指标体系的建立也有一定的指导意义。

从对经济信息分析的定性和定量的角度来看，三个不同的层次分析方法的侧重点也有所不同。笔者认为，在微观的层次上，对于经济信息效益的评价主要是定量的评价，用数据和指标来进行分析；在中观层次上的评价是定性和定量的结合，除了企业内部的定量评价，还要考虑对于外部的影响，此时只采用定量的方法就很难对其分析，需结合定性综合评价；而宏观层次上，尽管可以汇总各类数据，但很多事实都难以量化，在综合评价时更多的需要是人头脑的整体的思维能力、判断能力、决策能力，因此宏观层次更侧重于定性的方法。

5 结 语

在如今全球一体化的经济发展环境中，传统的经济信息效益评价指标已经不能够满足当前的评价需要，迫切需要建立一个全面的、多层次的、完善的评价指标架构。这不仅要从企业自身出发，更要立足于整个社会、国家；不仅考虑企业个体的利益，更要思考对于社会、国家的影响。只有这样，才可能做出全面、准确的评价。

参 考 文 献

1 梁前德，王光甫. 经济信息概论. 北京：中国商业出版社，1998
2 邱均平. 市场经济信息学. 武汉：武汉大学出版社，2001
3 乌家培. 经济信息与信息经济. 北京：中国经济出版社，1991
4 王光甫，丁玉国. 经济信息学. 郑州：河南人民出版社，1995
5 黄学忠. 经济信息与管理. 北京：人民出版社，1985
6 乌家培. 经济 信息 信息化. 大连：东北财经大学出版社，1996

（责编：愚加勃）

（上接第36页）监测分析和短语差异分析研究科技文献信息，从而监测科技发展的基本设计和优缺点。其中词频分析和突发词监测分析是以词出现的数量为基础的，但是两者的侧重点不同，词频分析主要是观察高频词的发展变化，而突发词监测分析是观察具有增长趋势加快的突发词的发展变化，突发词有可能是低频的但却具有情报意义，同时突发词监测也可以揭示主题发展过程中的阶段性。共词分析以词与词在文献中的共现为基础提炼出研究主题，减少了专家的参与，并且将主题作为研究对象。短语差异分析是以词汇特征分析为基础的，一方面可以观察到低频词的新热点；另一方面可以有效地揭示短语与短语之间的反映科技概念的发展变化上的关系。

参 考 文 献

1 马费成，张 勤. 国内外知识管理研究热点——基于词频的统计分析. 情报学报，2006；(2)
2 魏瑞斌. 基于关键词的情报学研究主题分析. 情报科学，2006；(9)
3 崔 雷，郑华川. 关于从 MEDLINE 数据库中进行知识抽取和挖掘的研究进展. 情报学报，2003；(4)
4 张 晗，崔 雷等. 生物信息学的共词分析研究. 情报学报，2003；(5)
5 Jon Kleinberg. Bursty and Hierarchical Structure in Streams. In：Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2002
6 Chen Chaomei. CiteSpace II：Detecting and Visualizing Emerging Trends And transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technology，2005；57(3)
7 Ke Weimao，Borner Katy，Viswanath Lalitha. Major Information Visualization Authors，Papers and Topics in the ACM Library. http://www.cs.umd.edu/hcil/InfovisRepository/contest-2004/6/unzip/contest-paper.pdf
8 Ketan Mane，Katy Börner. Mapping Topics and Topic Bursts in PNAS. http://arxiv.org/ftp/cs/papers/0402/0402029.pdf
9 Fidelia Ibekwe-SanJuan，Eric SanJuan. From Term Variants to Research Topic. http://fidelia1.free.fr/isko-hlt.pdf

（责编：梅）