情报分析与研究

基于 SOM的领域热点主题探测*

陆 伟　彭 玉　陈 武
（武汉大学信息资源研究中心　武汉 430072）

【摘要】针对学科领域中热点研究主题探测，尝试综合运用共词分析方法与自组织映射（SOM）方法，在词频统计的基础上，分析高频主题词在文献中的共现，并作为输入数据利用 SOM Toolbox进行 SOM聚类分析，得到领域热点研究主题。以传统医药领域为例进行实证分析，结果表明该方法对领域中热点主题探测有一定效果。

【关键词】自组织映射　热点主题　共词分析　传统医药
【分类号】G202

Hot Research Topics Detection Based on SOM

Lu Wei　Peng Yu　Chen Wu
（Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China）

【Abstract】According to detection of hot topics in a research field, the paper proposes a method combining co-word analysis and SOM together. By analysing the co-occurrence of high-frequency keywords in the literature as input data and using SOM Toolbox for SOM clustering, the collection of hot research topics is obtained. At last a case study is done by taking traditional medicine as an example, and experimental results show that this method is efficient in the process of hot research topics detection.

【Keywords】SOM　Hot research topics　Co-word analysis　Traditional medicine

1 引　言

热点研究主题是学科领域中聚焦了大量研究者关注和研究的一些主题，这些主题可能数量很少，却集中了领域中大部分的研究力量和资源，对热点主题的研究能够摒弃领域中一些并不能对全局产生影响的边缘研究主题，从而迅速简洁地展示整个学科领域的知识结构。热点研究主题的探测，对研究者、研究机构、相关政府部门的选题、科研立项有重要的指导意义，对学科本身的发展也十分重要。

如何准确地探测领域中的热点主题一直是情报学研究的一个问题。科学研究具有高度的动态性，新的研究主题不断出现，已经形成的研究主题通过分裂或融合形成新的主题，各个主题的重要性或者受关注程度也在不断增加或减少。因此热点主题探测的结果对准确性、实时性具有较高的要求。传统的探测方法依赖于研究者查阅领域相关的文献或成果，通常要耗费大量的人力、物力和时间，而且由于目前文献、相关研究成果数量正在迅猛增长，已经很难继续采用这种方法。综合运用信息计量学、数据挖掘等方法实现自动化的热点主题探测，已经成为相关研究的趋势。

收稿日期：2010－11－22
收修改稿日期：2010－12－20
*本文系教育部人文社会科学规划项目“专家专长智能识别与检索系统实现研究”（项目编号：09YJA870021）的研究成果之一。

2 国内外研究现状

国内外研究者很早就开展了科研领域热点主题识别与趋势预测的相关研究，Price在1965年就提出研究前沿（Research Front）的概念，用以描述某研究领域中引用周期较短暂的一类文献[1]。目前国内外的相关研究主要对领域内学术文献采用定量分析的方法，这些方法可以根据以主题词或文献为研究单元而分为两类。

2.1 以主题词为研究单元的相关研究

如果某一关键词或主题词在其所在领域的文献中反复出现，则可反映出该关键词或主题词所表征的研究主题是该领域的研究热点[2]。1997年加拿大蒙特利尔大学的Dale教授向加拿大国家研究理事会（NRC）提交了一份关于国际纳米科技研究现状的分析报告，这份报告在NRC提供的79个关键词的基础上，通过分析它们的词频总结了纳米科技论文和专利在全球范围内的产出分布[3]。

在词频统计的基础上，一些研究者进一步分析高频主题词在文献中的共现，以它们的关联强度为基础进行共词分析，在同簇中的主题词通常具有较高的相关性，因而使研究者能够较为容易地确定一些多义的主题词的具体含义和指向内容，从而有效降低研究者的认知负担。马费成等对CNKI数据库中近10年以来数字信息资源领域发表的期刊论文的关键词进行共词分析，并借助多元统计学方法中的因子分析法和系统聚类法，研究各主题词间的关系，探讨了国内数字信息资源的研究现状与热点[4]。Courtial利用此方法描述了科学计量学的学科结构和动态发展变化[5]。

一些学者认为热点研究主题是增长势头不断加强、未来可能成为主流研究对象的主题，这些学者采用的方法以Kleinberg的突发检测算法为代表。Kleinberg在2003年提出话题的突发监测（Burst Detection）算法，最初应用于新闻和电子邮件的主题突发性探测[6]。Chen在其设计的CiteSpace II中，在构建论文共引网络的基础上，利用Kleinberg算法对每个簇抽取其中文献的关键词，并作为该簇代表的热点主题的标签[7]。笔者认为这类主题可以称为趋势主题，与热点主题有较大区别，因而在本研究中没有采用该算法。

2.2 以文献为研究单元的相关研究

以文献为研究单元的相关研究主要分为文献主题分布统计分析和引文分析两类。在某领域已经存在一个权威的分类体系的前提下，用文献主题分布统计分析方法探测热点研究主题，具有简单和易于理解的特点。李文兰等对1993－2002年情报学期刊论文研究主题分布进行了统计分析[8]，以中图分类法为基础，统计G35下子类别的文献数，从而反映不同主题的研究热度。

引文分析能较好地反映学科的知识结构，因此被广泛用于热点主题探测的相关研究。根据引用关系的不同，引文分析又分为具体的三种形式：共被引分析、文献耦合、直接引用分析。Shibata等对比分析了三者用于探测领域研究前沿的优劣，通过实验得出结论以直接引用计量文献间联系具有一定优势[9]。张倩等以Web of Science网络数据库为数据源，对2004年SCI和SSCI共同收录的23种图书情报杂志刊载的参考文献进行共被引聚类分析，结合图书馆学、情报学专业知识对聚类结果进行分析解释，从而动态揭示近年来该学科的研究热点[10]。在引文分析的基础上，运用自组织映射（Self－Organizing Map，SOM）、多维尺度分析（Multi Dimensional Scaling，MDS）、路径寻找网络（Path Finder Network Scaling）等方法，描绘出某一学科领域的知识图谱（Knowledge Mapping），能够将学科内部结构以可视化的方式清楚地展现出来。德雷塞尔大学的Chen教授做了大量此方面的研究，他与McCain等以科学计量学1981－2001共20年间文献为数据，可视化展现了被引次数在一定阈值以上的文献的共被引网络[11]。近年来，国内也开始关注知识图谱，并取得了一些研究成果。大连理工大学刘则渊教授在此方面做了大量研究[12,13]。

因为领域内的研究成果大多会以学术论文的形式展现出来，所以目前国内外的相关研究基本上都以学术论文为数据源，而忽视了其他数据类型，如专利数据、科研项目、领域内奖励等。这些数据也是进行热点主题识别、趋势预测研究重要的数据源，但目前以这些数据为对象进行热点主题识别研究的成果很少，宋旭昌以iSchool成员科研立项为数据源用词频分析方法研究了2007年iSchool的科研热点[14]；Courtial等利用从专利标题抽取的主题词进行共词分析，确定发明的主题并预测可能的发展趋势[15]。

3 基于共词分析与SOM的热点主题探测方法

本文综合共词分析方法与自组织映射（SOM）作为

热点研究主题探测方法。共词分析是目前相关研究中最成熟的方法之一，它能够将表征相同主题的关键词归纳到同一个类中，从而更易总结出热点研究主题，而且避免了引文分析的时滞性的缺点。自组织映射作为一种聚类算法，能够将高维数据映射到低维空间并保持较好的拓扑结构，对噪声数据不敏感并且聚类结果不会因为初始值的选择发生很大变动，可视化的聚类结果结合人的感知能力将使理解数据间的关联和模式变得更加简单。

本方法的具体分析过程如图 1所示：

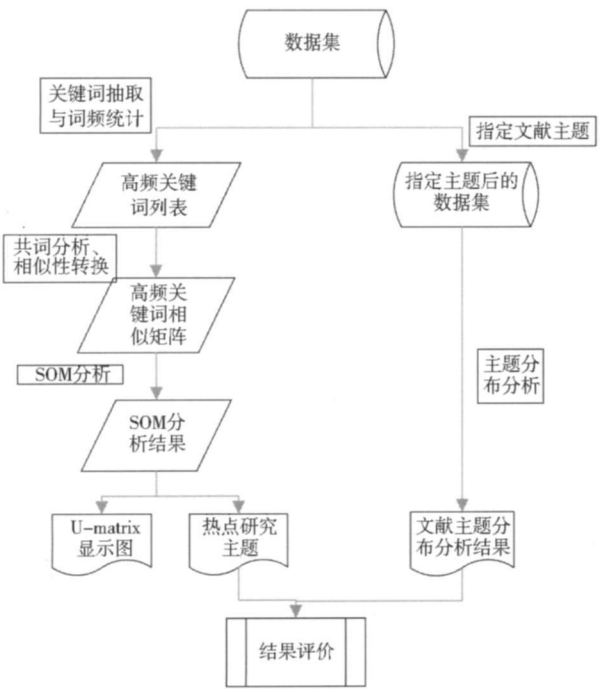

图 1　基于共词分析与 SOM的热点主题识别流程

图 1中左分支为本研究方法的分析过程，右分支为文献主题分布分析过程，这是为了评价研究方法的有效性而进行的对比实验，在本文中不再详细描述其具体分析过程。热点主题探测整体分析过程如下：

3.1　数据集获取与预处理

为了研究领域的热点主题，必须获取数据集以代表该领域并作为后续研究的基础。本文采用首先由专家人工选取相关关键词，然后以其为检索词从学术数据库中检索并下载的方式得到文献，经过去重、去除非学术论文的记录等预处理操作后，存入本地数据库作为数据集。这种方法最大限度地覆盖了所研究的领域范围，避免了研究主题的遗漏。

3.2　高频词提取

作为共词分析基础的主题词，可以是论文提供的关键词，或者是利用自然语言处理技术从题名、摘要和正文中提取的词。在本文的实证分析中，笔者将论文提供的关键词作为分析对象。

从数据集的文献中抽取关键词，统计各关键词的出现频率，得到关键词列表，并按频率从高到低排序。从该列表中选取部分关键词作为高频词。目前有三种方法较广泛地用于高频关键词的确定：结合研究者的专业知识背景人工确定高频词阈值，该方法主观性太强且受每年文献绝对数量的影响；利用 Donohue的高低频词分界公式确定阈值，该方法的有效性在学术界目前还存在争议；按词频高低顺序选择一定数量的关键词，使这些关键词的频率之和与所有关键词的总频率的比值达到一定阈值，该方法不受每年文献绝对数量变化的影响，在相关研究中取得了较好的结果，所以笔者采用了该方法。

3.3　共词分析

对高频词列表中的关键词两两间的共词关系进行统计并生成共词矩阵，在本文的实证研究中仅统计关键词在特定字段中而不是全文中的共现情况。共词矩阵中的元素值是词对的共现次数，没有经过规范化处理，高频词相对于其他频率不高的词在规模上的优势将会影响相似性分析结果的准确性。因此共词矩阵的元素值不能直接作为词对的相似度，笔者采用余弦系数将共词矩阵转换为相似性矩阵。

3.4　SOM分析

SOM由 Kohonen在 20世纪 80年代提出，近年来被应用于众多领域并取得了很好的研究成果。它是一种两层结构（输入层、输出层）的无监督竞争式学习的神经网络，通过递归的竞争学习，输入层中的对象分别映射到与之最合适的输出层结点中，其中具有较高相似度的对象所对应的输出层结点在理想状态下是相同的，或者具有较小的欧氏距离。

SOM的输出层如图 2所示。

同一个结点对应的输入层对象一般可以归为同一个聚类中；相邻结点的相似度较高，因此它们对应的输入层对象也可能是同一类。输出层结点的颜色代表其 U-matrix值，U-matrix是 Ultsch在 1992年定义的，其大小与原输出层一致，每个元素的值等于该结点的权

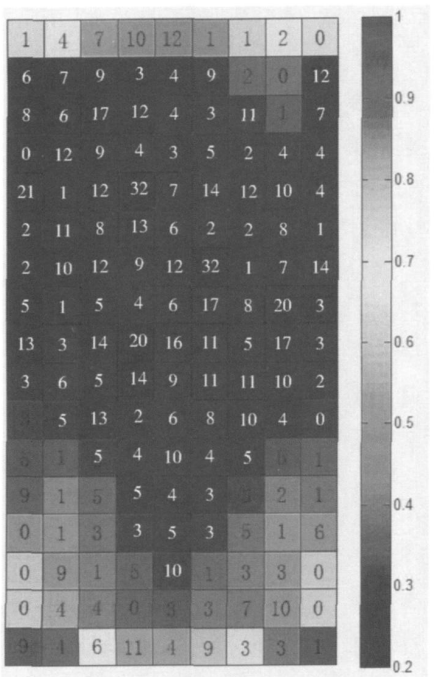

图 2 2008年传统医药领域高频词 SOM分析结果

向量与所有直接相邻结点的权向量之间的距离之和除以出现的最大的值[16]。因而在 SOM输出层中,颜色值较大的结点表示这个结点与相邻结点的距离较大,这些结点可能是聚类的边缘;颜色值较小的结点表示这个结点与相邻结点的距离较小,这些结点可能处于聚类的内部。

将关键词相似矩阵作为输入数据进行 SOM分析,生成 U-matrix显示图,结合图中结点的颜色及结点下关键词的意义,人工将所有结点归纳为不同的聚类,每个聚类代表该领域内的一个热点研究主题。

3.5 结果评价方法

对于聚类效果,笔者认为,某个聚类下与聚类主题相关的关键词占此聚类总关键词数的比例越大,则聚类效果越好。定义聚类 C_k的隶属度为 $M(C_k)$,其计算方式如下:

$$M(C_k)=\frac{\sum_{i=1}^{n} m_i}{n},\ m_i=1,0 \qquad (1)$$

其中,n为 C_k下的关键词数,m_i为第 i个关键词与 C_k的相关性,值 1,0分别表示相关、不相关。当以下几种情况中至少有一种发生时,$m_i=1$。

(1)该词与聚类主题意义相近,如 Tumor(肿瘤)与主题癌症意义相近。

(2)该词与聚类主题的某一方面相关,如 Bone-marrow-transplantation(骨髓移植)与主题白血病相关。

(3)该词意义过泛,而不能看出与聚类主题的相关性,但由其组成的词组与聚类主题相关,如 Expression的意义不明显,但 Gene Expression与主题癌细胞基因相关。

对于聚类内容,笔者通过比较本方法的研究结果与文献主题分布的分析结果,来评价研究结果的客观性及是否与领域宏观发展状况相符。如果文献主题分析采用的分类体系是该领域内的权威分类体系,其分析将能够较好地反映领域宏观情况,并具有很高的可信度和客观性,因而本文以其作为评价聚类内容的基准。

4 以传统医药领域为例的实证分析

以 2008年的传统医药领域为例,对本文提出的基于共词分析与 SOM的热点主题探测方法进行实现和实证分析,以评价本研究方法的有效性。

4.1 实验结果

数据集由 2008年传统医药领域的 22 991篇文献构成,共有不重复的关键词 80 735个,总词频 259 397次。按词频从高到低选取了 984个关键词作为高频词,它们的词频之和占总词频的 35%。统计高频词的共现情况,生成 984×984的相似矩阵,将其作为输入数据,利用芬兰赫尔辛基大学信息与计算机科学实验室在 Matlab环境中开发的 SOM Toolbox[17]进行 SOM分析。生成的 U-matrix图(见图 2),有 17行 9列共 153个结点,结点中的数字表示映射到其中的关键词的数量。

通过查看 SOM结点内的关键词,发现 153个输出结点中有 56个结点的主题较为明显。由于 SOM分析的特点,相邻输出结点的主题可能是相同或者相近的,所以结合这些主题本身以及对应 SOM结点的相邻程度,将这 56个结点归纳为 17个聚类主题。本文以该 17个聚类主题作为 2008年传统医药领域热点研究主题,如表 1所示。其中,G(i,j)表示图 2中第 i行第 j列的 SOM输出结点。

4.2 结果分析

17个热点研究主题的隶属度如表 2所示。

最终的平均隶属度超过了 0.6,由此可以看出该 56个节点与其所属聚类主题的相关性较高,17个热点研究主题都对应了一定数量的相关关键词。因而笔者

表 1　2008年传统医药领域热点研究主题

聚类主题	输出层节点
C1癌症、癌细胞、癌细胞基因	G(1, 1)、G(1, 2)、G(1, 3)、G(2, 1)、G(2, 2)、G(2, 3)、G(2, 9)、G(3, 1)、G(3, 2)、G(3, 3)、G(4, 2)
C2抗癌药物与癌细胞的抗药性	G(1, 4)、G(1, 5)、G(2, 4)、G(3, 4)、G(4, 4)
C3肿瘤、放化疗	G(2, 5)、G(2, 6)
C4白血病、干细胞	G(3, 6)、G(3, 7)、G(4, 6)
C5艾滋病	G(5, 7)
C6心脏病、动脉硬化	G(4, 8)、G(4, 9)、G(5, 8)、G(5, 9)、G(6, 1)
C7植物成分提取、植物提取物的各种作用	G(5, 4)、G(5, 5)、G(6, 4)、G(6, 5)、G(6, 6)、G(7, 4)
C8医学实验与其安全性	G(6, 8)、G(7, 9)
C9糖尿病	G(7, 2)、G(7, 3)
C10卫星图像、生物多样性、生态系统	G(7, 6)、G(7, 7)
C11神经系统、大脑皮层	G(8, 3)、G(9, 3)、G(9, 4)、G(9, 5)、G(10, 3)、G(10, 4)
C12推拿、徒手治疗	G(8, 8)
C13精神病、心理分析	G(10, 5)、G(10, 6)、G(10, 7)
C14听觉	G(11, 3)、G(12, 3)
C15女性更年期症状与治疗	G(11, 7)
C16瑜伽等疗法	G(13, 6)、G(13, 7)
C17护理学	G(16, 7)、G(16, 8)

表 2　2008年传统医药领域热点主题探测隶属度

聚类编号	节点数	热点关键词数	相关关键词数	不相关关键词数	隶属度
C1	11	89	49	40	0.550562
C2	5	41	29	12	0.707317
C3	2	13	8	5	0.615385
C4	3	19	10	9	0.526316
C5	1	12	6	6	0.5
C6	5	24	17	7	0.708333
C7	6	69	50	19	0.724638
C8	2	23	11	12	0.478261
C9	2	22	9	13	0.40909
C10	2	33	23	10	0.69697
C11	6	74	44	30	0.594595
C12	1	20	14	6	0.7
C13	3	31	17	14	0.548387
C14	2	18	14	4	0.777778
C15	1	10	8	2	0.8
C16	2	9	4	5	0.444444
C17	2	17	10	7	0.58824
总计	56	524	323	201	0.616412

总结出的这 17个主题能够代表其所在的聚类及包含的关键词。

MeSH是医学领域最权威的分类体系，本文以其为基础进行文献主题分布分析，得到的分析结果中与前文总结的 17个热点主题有许多相互印证之处，如表 3所示。

本文的研究结果与文献主题分布分析结果有很多吻合之处，因此本研究的结果正确反映了传统医药领域的大致研究状况，利用本方法识别领域的热点主题

表 3　SOM聚类结果与文献主题分布分析结果对比

SOM聚类编号与主题	对应的 MeSH编号与主题	MeSH词表对主题的解释	MeSH主题在文献主题分布结果中的排名
C1癌症、癌细胞、癌细胞基因	E02.190.701.884 Tissue Therapy	组织、细胞疗法	1
C4白血病、干细胞	E02.190.701.884 Tissue Therapy	组织、细胞疗法	1
C7植物成分提取、植物提取物的各种作用	D20.215.784.500.350 Drugs, Chinese Herbal	中国中草药	2
C7植物成分提取、植物提取物的各种作用	D26.667 Plant Extracts	制药，从植物中去除一定成分得到药剂	6
C7植物成分提取、植物提取物的各种作用	G01.273.118.299 Ethnobotany	植物及它们的成分用于传统医药	23
C13精神病、心理分析	E02.190.525.249 Imagery (Psychotherapy)	精神图像（Mental Images），用于精神障碍的治疗	5
C13精神病、心理分析	E02.190.525.217 Hypnosis	人类催眠	9
C13精神病、心理分析	E02.190.888.249 Color Therapy	光线疗法，用于治疗生理或心理疾病	15
C14听觉	E02.190.888.080 Acoustic Stimulation	声音刺激神经系统反应	7
C16瑜伽等疗法	E02.190.525.890 Tai Ji	太极	25

得到的结果可信。利用本研究方法得到的热点研究主题更加具体和细致，聚类主题中的关键词可以帮助研究者了解相关的主题中正在被研究的具体问题或者方法。例如 C7（植物成分提取、植物提取物的各种作用）与 D26.667（Plant Extracts），通过查阅 C7下的关键词，可以发现植物提取中经常被提取的成分、经常用到的植物部件以及植物提取物的用途等。

5　结　语

本文将共词分析与自组织映射综合运用于领域热点主题探测，取得了较好的研究成果，但尚存不足之处。

（1）本研究以主题词为研究单元，而以文献为研究单元的方法也具有其特点和长处，可以在未来的研究工作中综合运用以充分利用各自的优点；

（2）本文的研究仅采用了学术文献作为研究对象，各种其他数据源如专利也可以揭示领域的热点研究主题，在以后的研究中需要探讨如何综合多种类型的数据源。

（致谢：本文得到了武汉大学“70”后学者团队计划和汇海科技——武汉大学移动商务平台联合实验室资助，特此感谢！）

参考文献：

[1] Price D J. Networks of Scientific Papers[J]. Science, 1965, 149 (3683): 510－515

[2] 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[J].情报学报，2006，25(2)：163－171.
[3] Dalle R，Gauthier E，Ippersiel M P. The State of Nanotechnology Research：Report to the National Research Council of Canada[R]. 1997.
[4] 马费成，望俊成，陈金霞，等. 我国数字信息资源研究的热点领域：共词分析透视[J].情报理论与实践，2007，30(4)：438－443.
[5] Courtial J P. A Coword Analysis of Scientometrics[J]. Scientometrics，1994，31(3)：251－260.
[6] Kleinberg J. Bursty and Hierarchical Structure in Streams[C]. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2002，91－101.
[7] Chen C M. CiteSpace Ⅱ：Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature[J]. Journal of the American Society for Information Science and Technology，2006，57(3)：359－377.
[8] 李文兰，杨祖国. 情报学研究主题分布的文献计量学分析[J].情报科学，2005，23(3)：396－400.
[9] Shibata Naoki，Kajikawa Yuya，Takeda Yoshiyuki，et al. Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation[J]. Journal of the American Society for Information Science and Technology，2009，60(3)：571－580.
[10] 张倩，潘云涛，武夷山. 基于Web of Science数据的图书情报学研究聚类分析[J].情报杂志，2007，26(2)：82－84.
[11] Chen C M，McCain K，White H，et al. Mapping Scientometrics (1981－2001)[C]. In Proceedings of the 65th Asist Annual Meeting. Medford：Information Today Inc，2002，25－34.
[12] 刘则渊. 科学学理论体系建构的思考——基于科学计量学的中外科学学进展研究报告[J]. 科学学研究，2006，24(1)：1－11.
[13] 梁永霞，刘则渊，杨中楷，等. 引文分析领域前沿与演化知识图谱[J]. 科学学研究，2009，27(4)：516－522.
[14] 宋旭昌. 2007年iSchool科研热点分析[J]. 情报探索，2008，12(9)：118－121.
[15] Courtial J P，Callon M，Sigogneau A. The Use of Patent Titles for Identifying the Topics of Invention and Forecasting Trends[J]. Scientometrics，1993，26(2)：231－242.
[16] 安璐.基于自组织映射的期刊主题研究[D]. 武汉：武汉大学，2009.
[17] Laboratory of Computer and Information Science. SOM Toolbox[EB/OL]. [2010－03－24]. http://www.cis.hut.fi/projects/somtoolbox/.

(作者 E-mail：pengyucic@126.com)