冯 璐 冷伏海

共词分析方法理论进展*

摘 要 共词分析方法属于内容分析方法的一种。其原理主要是对一组词两两统计它们在同一篇文献中出现的次数,对这些词进行聚类分析,进而分析这些词所代表的学科和主题的结构变化。有基于包容指数和临近指数的共词分析方法、基于战略坐标的共词分析方法、基于数据库内容结构分析的共词分析方法。公式 5。图 1。参考文献 13。

关键词 文献计量学 共词 共词分析方法
分类号 G257

ABSTRACT Co-word analysis is a kind of content analysis. Its main principle is to make statistics of pairs of words appearing in the same documents and then analyze structural changes of disciples and subjects they represent. In this paper, the authors introduce some kinds of co-word analysis methods. 5 formulas. 1 figs. 13 refs.

KEY WORDS Bibliometrics. Co-word. Co-word analysis.
CLASS NUMBER G257

1 概述

共词分析方法最早被详细描述是在 20 世纪 70 年代中后期由法国文献计量学家开始的。共词分析经过 20 多年的发展,方法已经被广泛应用到许多领域。研究者利用共词方法基本原理概述研究领域的研究热点,横向和纵向分析领域学科的发展过程、特点以及领域或学科之间的关系,反映某个专业的科学研究水平及其发展历史的动态和静态结构,拓展信息检索领域以求帮助用户检索信息等等。到目前为止,共词分析方法产生了大量的应用成果。主要集中的领域有人工智能(Courtial 和 Law, 1989)、科学计量学(Courtial, 1994)、人文学科计算研究(Horton 等, 1998)、信息科学和信息系统的研究(Monarch, 2000)、信息检索(Ying Ding 等,2000)等等。

2 共词分析方法内涵

共词分析方法属于内容分析方法的一种。它的原理主要是对一组词两两统计它们在同一篇文献中出现的次数,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[1]。它利用大量文献中共同出现的关键词对有效地反映文本关键词之间的关联强度,减少了关键词的空间,用一套结构图有效地展示了关键词之间的关联。

共词分析方法的实施是要在理想化的状态下开展的,因此,学者们在研究过程中不断对共词分析的假设前提产生质疑。最早提出共词分析假设前提的是 Whittaker(1989)[2]。他当时指出,选择文献作为共词分析的假设前提主要是:作者都是很认真地选择他们的技术术语;当在同一篇文章中使用不同的术语时,就意味着它们之间有一些关系并不微不足道,它们一定是被作者认可或要求的;如果有足够的不同作者都对同一种关系认可,那么这种关系可以认为他们所关注的科学领域具有一定的意义。当关键词被用于分析时,第四个论据被提出来,即经过培训的标引者选择出来的用来描述文章内容的关键词,事实上是相关科学概念可以信赖的一个指标。之后,Law 和 Whittaker(1992)再次重申了上面假设中的两个,第一个是标引论文的关键词毫无疑问可以反映科学研究的现状,第二个是其他科学家接受的观点可以影响未来使用类似关键词标引发表的科学论文。

共词分析方法就是基于这样的一些假设而成立的。如果这些假设前提都成立的话,那么共词分析方法利用文章中词语对的共现频次来反映包含在文章中的概念结构就会成为可能。

* 本文得到国家软科学项目"科学发展趋势预测的理论与方法及其实践研究"(2003DGQ1B170)和国家自然科学基金项目"我国情报学学科发展、建设与前瞻性研究"(70373038)的资助。

3 共词分析方法演进

从共词分析方法发展至今，共词分析方法主要经历了3个阶段，即第一代基于包容指数和临近指数的共词分析方法，第二代基于战略坐标的共词分析方法以及新一代基于数据库内容结构分析的共词分析方法。

3.1 基于包容指数和临近指数的共词分析方法

3.1.1 包容指数和临近指数

1979年和1981年，Serge Bauin等使用包容指数(inclusion index)和邻近指数(Proximity Index)，分别显示了水产研究的动态变化。包容指数和邻近指数主要用于测量款目之间关系的强度[3]。包容指数主要用来计算主题领域的层次，计算公式如下[4]：

$$Iij = Cij/\min(Ci, Cj) \quad [1]$$

其中，Cij代表关键词对Mi和Mj在文献集合中的数量；Ci代表关键词Mi在文献集合中的出现频次；Cj代表关键词Mj在文献集合中的出现频次；min(Ci,Cj)代表Ci和Cj两个频次的最小值。这个公式可以用来计算那些出现频次相对高的关键词。

当存在着一些中间关键词(mediator keywords)，而且这些关键词的相对出现频次比较低，但是仍然在这些非重要的关键词之间存在着一定的关系，于是用临近指数来计算潜在的领域，计算公式如下：

$$Pij = (Cij/CiCj) \times N \quad [2]$$

其中，Cij、Ci和Cj同公式[1]中表示的意思一样，N代表集合中文献的数量。

之后经过Turner、Whittaker、Law和Whittaker、Coulter、Coulter等人不断研究，Callon等提出等价系数(Equivalence Coefficient，简化为E)[5]，用来计算关键词之间的关联值。

$$Eij = (Cij/Ci) \times (Cij/Cj) = (Cij)2/(Ci \times Cj) \quad [3]$$

其中，Eij的值也是在0～1之间。由于Eij可以同时计算关键词i和j出现在对方集合的频次，因此Turner和他的同事称这个参数为相互包含的系数。

3.1.2 包容地图和临近地图

以上面3个指数为基础，把主题词或关键词聚类成组，并以网络地图的方式表现出来。通过比较不同时期的网络地图，就可以表现出科学的结构和动态变化。

Callon等在1986年提出了两个术语：包容地图(Inclusion Map)和临近地图(Proximity Map)。它们的创建都是在计算包容指数和临近指数基础上进行的[6]。

包容地图用于揭示领域内的中心主题，描述低频次关键词之间的关系，这个图涉及了更多某个主题的信息。建立过程是：在计算包容指数后，选择包容指数值最高的关联，这些关联的节点作为第一个聚类的起始点。其他的关联和相应的节点按照包容指数递减的顺序添加到地图中，直到达到阈值I_0。然后去掉所有这些包含在聚类中的节点，下一个地图再从剩下的关联中找最高的包容指数值。

临近地图用于揭示隐藏在中心主题之中的较小领域的关系，这个图更多地涉及了主题之间的关联。建立过程是：计算临近指数建立临近地图。如果阈值P_0达到足够低，关键词之间更多的临近关系将出现在地图中，同时，关键词中间值和热点主题也会在包容地图中出现。这样，就可以研究次要主题与热点主题之间的关系了。

还可以采用另外一个聚类的方法，是Callon(1991年)提出的[7]。在这个方法中，使用等价系数测量关键词之间的强度。使用阈值10来限制一个聚类的词语数量。首先选择E最高的关联。当一个聚类已经有10个单词的时候，下一个关联将被拒绝。这个第一个被拒绝的关联值被称为是"饱和阈值"(saturation threshold)。一个聚类产生后，另一个聚类开始了。一个新聚类的第一个关联的E值被称为是"最高限度阈值"(ceiling threshold)。在这些值的基础上，产生了3个不同的聚类：第一个是孤立的聚类(isolated clusters)，它的特点是与其他聚类的关联值为空或较低；第二个是次要聚类(secondary clusters)，它与其他聚类的外部关联值在最高限度阈值上，有足够的理由认为它们与外部的聚类之间的关系是正常的延伸；第三个是主要聚类(principal clusters)，其中一个或更多的聚类是有关联的，它们的关联值达到了饱和阈值。

之后，Coulter等(1996)把聚类的过程划分为两个阶段(two pass)[8]。在pass－1，和上面的建立包容地图过程类似，用E计算两个关键词之间的关联强度。在这一阶段，产生了一些描述符之间的关系，这些描述符被称为是内部节点，这些相应的关联被称为是内部关联。在pass－2，通过添加pass－2关联来扩展聚类。Pass－2中两个节点的关联必须包含在pass－1聚类中。Pass－2中的节点和关联都被称为是外部的。Pass－1建立用来确认集中研究的领域；pass－2可以确认不只与一个网络有关的描述符指出潜

在的关系。

3.2 基于战略坐标的共词分析方法

在早期,对采用共词分析方法产生的结果进行分析非常困难,一些专家开始怀疑共词分析结果的可行性。这种情况下,研究者提出建立战略坐标来分析结果[9],这在后来被称为是第二代的共词分析,而把前面提到的利用包容指数和临近指数进行的共词分析称为第一代共词分析。

3.2.1 战略坐标

战略坐标(stategic digram)是在建立主题词的共词矩阵和聚类的基础上,用可视化的形式来表示产生的结果。目前这个战略坐标已经被用于许多共词分析的研究中,例如 Turner 等(1988)、Courtial 和 Law (1989)、Turner 和 Rojouan(1991)、Callon 等(1991)、Coulter 等(1998)。

1988 年 Law 等提出了用"战略坐标"来描述某一研究领域内部联系情况和领域间相互影响情况。在战略坐标中,x 轴为向心度(Centrality),表示领域间相互影响的强度;Y 轴为密度(Density),表示某一领域内部联系强度。其中[10]:

向心度用来测量一个学科领域和其他学科领域的相互影响程度。一个学科领域与其他学科领域联系的数目和强度越大,这个学科领域在整个研究工作中就越趋于中心地位。对于特定的类别,向心度的计算可以通过该类别的所有主题词或关键词与其他类别的主题词之间链接的强度来进行。这些外部链接的总和、平方和的开平方等都可以作为该类别的向心度。

密度用来测量组成聚类的词语之间的关联强度,也就是聚类内部的强度。它很好地说明了维持一个聚类的能力以及在领域中发展的过程。某一类别密度的计算可以有多种方式,首先计算本类中每一对主题词或关键词之间的在同一篇文献中同时出现的次数,通过计算这些内部链接的平均值、中位数或者平方和,得出这个类别的密度。

以向心度和密度为参数绘制成的二维坐标即为战略坐标,它可以概括地表现一个领域或亚领域的结构。其典型结构是横轴表示向心度,纵轴表示密度,坐标的原点在两个轴的中位数或者平均数。这个地图将每一个二维空间的题目领域划分为4个象限,可以用来描述各主题的研究发展状况。

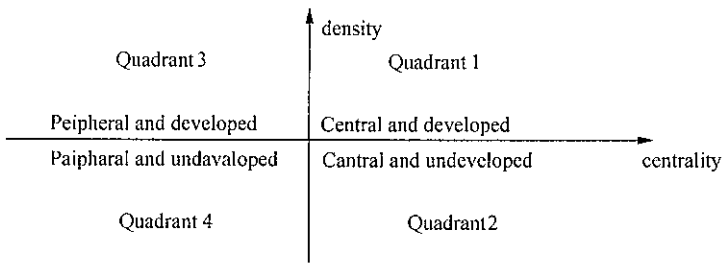

图1 二维空间的题目领域

在第一象限,主题领域内部关联,并且处于研究网络的中心。它们的密度和向心度都较高。密度高说明研究主题内部联系紧密,研究趋向成熟。向心度高说明这两个研究热点又与其余各热点有广泛的联系,也就是与其余研究密切相关。

在第二象限,主题领域比较集中,研究人员都有兴趣,但是结构不紧密,研究尚不成熟。这个领域的主题有进一步发展的空间,具有潜在的发展趋势。

在第三象限,主题领域内部链接紧密,这些领域的研究已经形成了一定的研究规模。有很多外围的社会组织加入研究,但是在整个研究网络中处于边缘。

在第四象限,研究主题密度和向心度都较低,是整个领域的边缘主题,内部结构比较松散,研究尚不成熟。

3.2.2 网络比较

以战略坐标为基础,可以对不同时期的研究网络进行比较和评价。对结果的分析和评价可以从两个方面考虑,即网络稳定性(thes tability of networks)和网络比较(network comparison)[11]。

(1)网络稳定性。

在战略坐标的基础上,可以分析聚类网络的稳定性,并预测未来的变化。这个问题在很多研究中都提出来了,在研究中使用的方法主要有两大类。

一个是直接基于战略坐标(例如 Callon 等 1991, Turner 和 Rojouan 1991)。比如,相对于第一象限而言,位于第二象限和第三象限的研究主题内容可能会发生巨大变化。在第二象限中,没有形成结构的主题需要提高它们的一致性。在第三象限中,主题的范围可能需要扩展,为了更好地表达它们在做什么,而它们变化的最终的目标就是向着第一象限的方向努力。

第二种方法是基于向心度和密度的比率(Courtial 等 1993, Turner 等 1994)。这个比率被认为是许多研究科学和技术的发展阶段有意义的指标。如果比率趋向 1,表明这个领域在研究网络中处于主流地位;如果比率远离 1,表明主题的支持率在下降甚至是可能在研究网络中消失。

(2)网络比较。

在共词分析中,几个聚类可以被同时构建。为了研究不同聚类在同一时间的差别或在不同时间的差别,研究者们提出要进行网络比较。Callon 等(1991)提出 3 个阶段的方法。他的方法的核心思想是先利用公式 T = (Wi + Wj)/Wij 比较两个给定的聚类,其中 Wi 是在聚类 Ci 中的词语数量,Wj 是在聚类 Cj 中的词语数量,Wij 是在聚类 Ci 和 Cj 中的词语数量。然后比较它们在战略坐标中的位置,最终建立一个聚类的生命周期曲线。

此外,Law 和 Whittaker 1992 提出参数影响指数(Influence index)和出处指数(Provenance index),用它们来说明在之后的时间内相似主题之间的关系[12]。

影响指数表明在一个聚类中某个主题内的词频与另一个聚类中任何给定的主题的关系,公式为:

$$Iij = (2 \times Mij + Lnij)/(2 \times Ni) \quad [4]$$

其中,Mij 是出现在主题 i 和后来的主题 j 中词语的数量;Lnij 是在主题 i 和与后来的主题 j,但是之后不属于其他主题中词语的数量;Ni 是在主题 i 中词语的数量。Iij 的值高表示前面的主题对后面主题影响大。

出处指数表明来自于任何给定的主题中的之前的那个聚类的第二个聚类中的词频。公式为:

$$Pij = (2 \times Mij + Lnij)/Nj \quad [5]$$

其中,Mij 是出现在主题 j 和先前的主题 i 中词语的数量;Lnij 是在主题 j 和与先前的主题 i,但是之后不属于其他主题中词语的数量;Nj 是在主题 j 中词语的数量。Pij 的值高表明第二个主题的出处主要来源于前一个类的一个主题。

3.3 基于数据库内容结构分析的共词分析方法

数据库内容结构分析(Database Tomography,简称 DT)是新一代共词分析方法,是由 Kostoff 等(1995)提出的,它是可以用于分析大量的数字化文本资源的系统[13]。在这种方法中出现了两个参数:频率分析(Frequency Analysis)和临近分析(Proximity Analysis)。频率分析用于揭示数据库中较深入的主题,而临近分析用于揭示这些主题之间的关系以及主题和子主题之间的关系。DT 分析方法通常分为 3 个步骤。第一步,确定文本分析的主题,计算文本中所有 1 个单词、2 个单词、3 个单词构成短语的出现频率。选择频率最高的技术内容短语作为全文数据库的深层分析主题。第二步,通过计算 50 个单词以内的短语与主题在一篇文本中共同出现的频率,构建词频字典来揭示该短语与主题之间的关系,引入 Numerical Indices 机制表明关系的强度,从而进一步确定主题和子主题之间的定性和定量化关系。第三步,为 Numerical Indices 设定阈值,筛选与聚类主题关系密切的短语。每个最后一步,追溯这些主题随时间变化的进展和它们之间的关系。

Kostoff 等利用 DT 研究了化学文献。同共词分析类似,DT 可以被用来确认主要的延伸领域以及这些延伸领域之间的关系。它可以提供整个研究网络一个全面的评价,允许进行对某个感兴趣的主题进行更详细研究。DT 还可以用于扩展最初的信息检索条件(Kostoff 等 1997)。Rotto 和 Morgan(1997)采用这个方法研究一个论文文摘中潜在的行业。

4 结束语

经过 20 多年的发展,共词分析方法从原理到使用都有了大幅度改进,这些改进已经不断在不同领域得以广泛应用。与其他的文本分析方法相比较,共词分析方法灵活,结果直观,有相当广阔的应用范围和前景。利用共词分析方法基本原理可以概述研究领域的研究热点,横向和纵向分析领域学科的发展过程、特点以及领域或学科之间的关系,反映某个专业的科学研究水平及其发展历史的动态和静态结构,基础研究和技术研究之间的关系,评价领域内研究成果投入和产出的关系,拓展信息检索领域以求帮助用户检索。但是这种方法也存在着一些弊端,比如方法的成立必须不考虑索引者的影响、词汇选择等一些人为因素的限制,这些问题如何改进都有待于在今后的研究中不断探索。

参考文献

1,3 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)

2,6,12 Qin He. Knowledge Discovery Through Co-Word Analysis. Library Trends, 48, 1999(1)

4 Ying Ding et al.. Bibliography of information retrieval research by using co-word analysis. Information Processing and Management, 2000(4)

5,7,11 Callon et al.. Co-Word Analysis For Basic And Technological Research. Scientmetrics, 1991 (22)

8 Coulter Neal et al.. An Evolutionary Perspective of Software Engineering Research Through Co-Word Analysis. Technical ReportCMU/SEI-95-TR-019ESC-TR-95-019

9 Law et al.. Policy and the Mapping of scientific change: A Co-Word Analysis of Research into Environment acidification. Scientometrics, 14(3-4), 1988

10 张晗,崔雷等. 生物信息学的共词分析研究. 情报学报,2003(5)

13 Kostoff Ronald N. et al.. Database Tomography for Information Retrieval. Journal of Information Science, 23(4)1997

冯 璐 中国科学院文献情报中心,中国科学院研究生院情报学专业2004级博士研究生。通信地址:北京市海淀区中关村北四环西路33号。邮编100080。

冷伏海 教授,中国科学院文献情报中心博士生导师、情报研究部副主任。通信地址同上。

(来稿时间:2005-04-13)

(接第75页)对于电子商务站点而言,有关客户购买行为的数据是非常重要的,这些数据可以用来识别客户关系生命周期的特定阶段,进而帮助站点制定相应的客户策略。

4.3 站点结构优化

目前,站点内容的组织方式都是从站点的角度来安排的。它往往与站点用户所期望的组织方式有所差异。站点结构优化的过程也就是消除差异的过程。这些差异可以通过分析点击流数据信息得到。比如,如果站点用户在所期望的位置找不到目标页面,往往会点击"后退"按钮或者直接点击新链接继续寻找。如果点击"后退"按钮,用户的"后退"点击流就为优化站点结构提供了一种思路,即用户点击"后退"按钮的页面所在位置就是用户针对这个目标页面的期望位置,此时站点设计可以考虑调整站点结构或者是在期望位置添加指向目标页面的链接。另外关联规则分析和序列模式分析会发现站点各个频道之间或者是站点一般页面之间的关联度,这样可以调整站点频道在页面中的位置,或者是在关联度较高的一般页面之间增加链接。

4.4 站点个性化

所谓站点个性化实质上就是为站点用户提供个性化的站点访问体验。由于传统的手工决策规则系统方法、基于内容的过滤代理系统方法、协作过滤系统方法的种种不足,点击流数据挖掘已经成为站点个性化主流方法[8]。比如可以采用聚类分析方法,在数据预处理的基础上实现基于站点使用和站点内容的交易事务聚类,然后导出站点的使用文档和内容文档,在此基础上结合当前用户会话形成基于站点使用和站点内容的个性化推荐集,最后在整合两种推荐集的基础上完成个性化推荐。关联规则分析方法和序列模式分析方法同样也可以用来完成站点的个性化推荐,其基本原理和聚类方法是一样的。

参考文献

1 马费成,李纲,查先进. 信息资源管理. 武汉:武汉大学出版社,2000

2 Peter Pirolli, Jame Pitkow, Ramana Rao. Silk from a Sow's Ear: Extracting Usable Structures from the Web. Proc. In ACM Conf. Human Factors in Computing Systems, 1996

3 D. D. Lewis, et al. Training algorithms for linear text classifiers. In Proceedings of the Ninetcenth International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996

4 R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and information Systems, 1999(1)

5,6 L. Catledge, J. Pitkow. Characterizing browsing behaviors on the World Wide Web. Computer Networks and ISDN Systems, 1995(6)

7 Jianhan, ZhuJun, Hong John G. Hughes. Using Markov models for web site link prediction. In Proceedings of the thirteenth ACM conference on Hypertext and hypermedia, 2002.

8 J. Srivastava et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000(2)

易 明 华中师范大学信息管理系讲师。通信地址:武汉。邮编430079。

邓卫华 华中农业大学经济管理学院讲师。通信地址:武汉。邮编430070。

(来稿时间:2005-07-07)