

●王曰芬 宋爽 卢宁 朱烨

共现分析在文本知识挖掘中的应用研究*

摘 要 共现分析是将各种信息载体中的共现信息定量化的分析方法,是支撑知识挖掘和知识服务的重要手段和工具。能够体现信息的内容特征和外表特征共同构成了文本知识关联揭示和知识挖掘的基础。探索了共现分析方法在基于空间分析、时间分布和内外关联映射的文本知识挖掘中的应用。图8。参考文献13。

关键词 文本挖掘 共现分析 空间分布 时间分布 内外关联映射

分类号 G254

ABSTRACT Co-occurrence analysis is a quantitative analysis method of co-occurrence in various information media, and is an important tool to support knowledge services and knowledge mining. The content characteristics and outer characteristics representing information comprise the basis for the revelation of text knowledge correlation and knowledge mining. The authors also discuss the application of co-occurrence analysis in text knowledge mining based on space distribution, time distribution and integrated mapping of outer and inner information. 8 figs. 13 refs.

KEY WORDS Text mining. Co-occurrence analysis. Space distribution. Time distribution. Integrated mapping of outer and inner information.

CLASS NUMBER G254

共现分析是将各种信息载体中的共现信息定量化的分析方法^[1],以心理学的邻近联系法则和知识结构及映射原则为方法论基础。通过共现分析,人们可以发现研究对象之间的亲疏关系,挖掘隐含的或潜在的有用的知识,并揭示研究对象所代表的学科或主体的结构与变化。在计算机技术的辅助下,共现分析在构建概念空间和 Ontology 实现语义检索、改进知识组织中文本分类效果、分析文献中知识内容关联、挖掘知识价值等方面彰显出独特的功能^[2],正在成为支撑知识挖掘和知识服务的重要手段和工具。

在知识表达中,能够体现信息的内容特征和外部特征不仅具有语义内涵而且是相互关联的,这些内容特征与外部特征共同构成了文本知识关联揭示和知识挖掘的基础。本文通过对文本知识的内容特征和外表特征进行不同组合的共现分析研究,探索共现分析方法在基于空间分布、时间分布和内外关联映射的文本知识挖掘中的应用。

1 在基于空间分布的文本知识挖掘中的应用研究

基于空间分布的文本挖掘研究是指将一定时间范围内的文本知识转换至二维或三维网络空间中进行展示,以节点代表知识单元,以空间距离反映知识的关联,进而挖掘隐含的或潜在的信息和知识,以揭示研究对象所代表领域的整体结构、研究热点及侧重点。归纳起来,共有两种实现思路可实现基于空间分布的文本挖掘。

1.1 通过由部分到整体构造的文本知识关联网络挖掘知识

通过渐进式的共现分析,构造由部分到整体的文本知识关联网络,以挖掘基于空间分布的知识内容。该思路的实现方式分为两个步骤:首先从相似度最高的词汇开始,逐个连接相似度高于阈值的其他词汇形成子网络,形成对文本知识中某个部分的细致描摹;然后,将其他属于某个子网络的节点对子网络进一步扩展,得到子网络之间的关联关系,构成上层网络,以反映文本整体的知识关联关系^[3]。

Florida Atlantic 大学的 Neal Coulter 对软件工程领域的研究就是基于上述思路^[4]。作者对大量出版物用标准分类表 CCS(Computer Classification System)进行标引后,选出至少有一个标引词属于“软件工程”大类下的文献作为分析对象。对分析对象计算相似度($S(c_i, c_j, c_y) = \frac{C_{ij}^2}{C_i C_j}$)后,从所有可供选择的标引词中找到 S 值最高的词,开始生成 Pass-1 网络。该研究的具体做法是:

从 S 值最高的连接出发,深度优先地找到其他可连接的节点,直到不存在相关节点或 Pass-1 网络的节点和连接数达到最大值。将这些已连接的节点从下一步可供选择的 Pass-1 标引词表中删去。从标引词表中剩余词汇中再次选出 S 值最高的链接直至标引词表中不存在共现程度足够生成网络的标引词,此时所有的 Pass-1 网络全部生成。图 1 表示的

* 本文系 2006 年度国家社会科学基金项目“图书情报领域知识服务的体系架构和服务能力评价研究”(项目编号 06BTQ027)的研究成果。

是从 User Interfaces 出发生成的基于空间分布的文本知识关
联子网络,总共生成了 11 个类似的子网络。

图中节点代表共现的标引词,节点间的连线代表两节点
之间有较强的共现关系,连线上的数字记录两端节点实际共
现的次数。可以认为,处在网络中心的节点是该子网络中的
中心概念,与中心概念相互关联的其他词汇反映了网络的研究
内容,节点之间的连接关系体现了研究内容之间的关联关
系。这样,以文本资源中词汇的共现特征为依据,文本知识

中的某个组成部分得以在二维空间中得到细致全面的描摹。

将在 Pass-1 中出现的所有标引词定义为可供选择的标
引词,从第一个 Pass-1 网络开始,以 S 值降序的顺序生成与
Pass-1 网络节点相连、在可供选择的标引词表内属于其他
Pass-1 网络的标引词连接,直到不存在达到共现最小次数的
节点或已达到节点和链接总数最大值。对所有的 Pass-1 网
络执行该步骤直到生成所有的 Pass-2 网络,如图 2 所示。

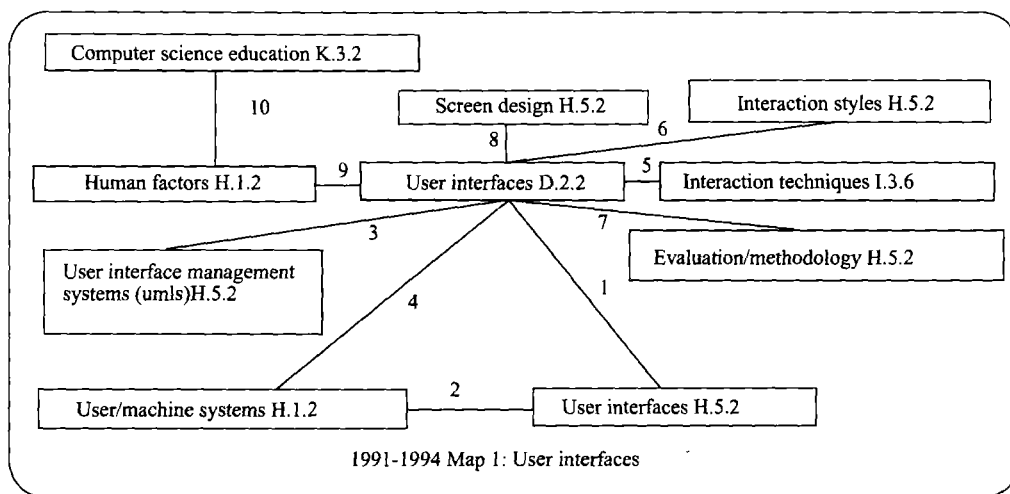


图 1 pass-1 文本关联网络实例

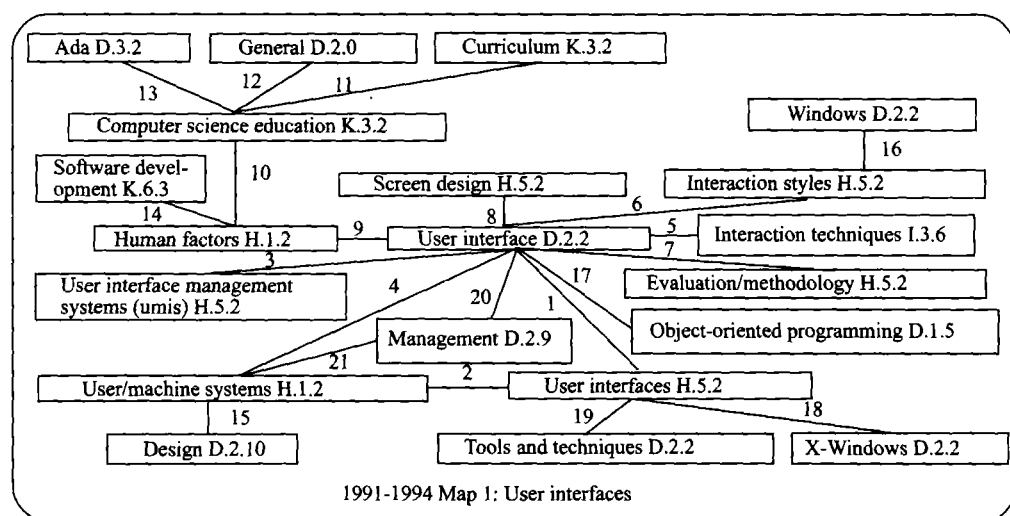


图 2 Pass-2 文本关联网络实例

对分析出的 11 个 Pass-1 网络和 11 个 Pass-2 网络进行
整合,总结出原本孤立的子网络之间的潜在关联。如果网络
A 有内部节点与网络 B 的 Pass-2 节点有连接,并且这些连接
的 S 值超过网络 B 的 Pass-1 图中的最小的 S 值,则从 A 网络

出发构建一条到 B 的连接,最终建立起以 11 个子网络为节
点,反映整个文本空间关联的上层网络,结果如图 3 所示。

通过上述方法,可以使得研究人员能够跟踪共现后词汇
之间的空间分布关系,分析过程简明可控,分析结果可信且

容易理解。

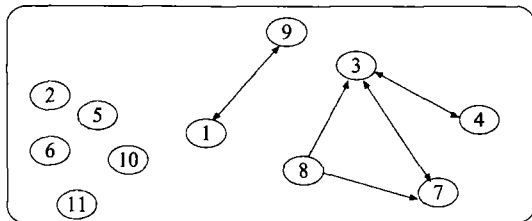


图3 对1991-1994软件工程文本资源构建的上层网络

1.2 通过由整体到部分构造的文本知识关联网络挖掘知识

与前述不同,该研究思路不拘泥于词和词之间实际存在的空间共现关系,而是将所有词汇作为一个整体考虑。计算

得到 $n \times n$ 的共现矩阵后,每个词由该词对应的 n 维向量代表。这样,比较两个词汇间的相似程度,也就是对这两个词汇与共现矩阵中其他所有词汇项的共现情况的近似程度进行比较。由于该方法涉及到多变量之间错综复杂的关系,多元统计分析方法成为研究过程中的重要工具。

南洋科技大学的 Ying Ding 等人在 2001 年对信息检索领域的核心问题进行研究时,应用的就是这种方法^[5]。他们根据文本出现的词汇情况,选择了 240 个有代表性的词汇作为分析对象。首先利用的层次聚类方法中的 Ward's 方法将词汇划分为 5 类,以代表 5 大主要研究问题,并选出每类中的高频词作为类名代表该类;然后使用多维标度技术在二维空间中展现信息检索领域中这 5 大问题之间的关联。分析结果如图 4 所示。

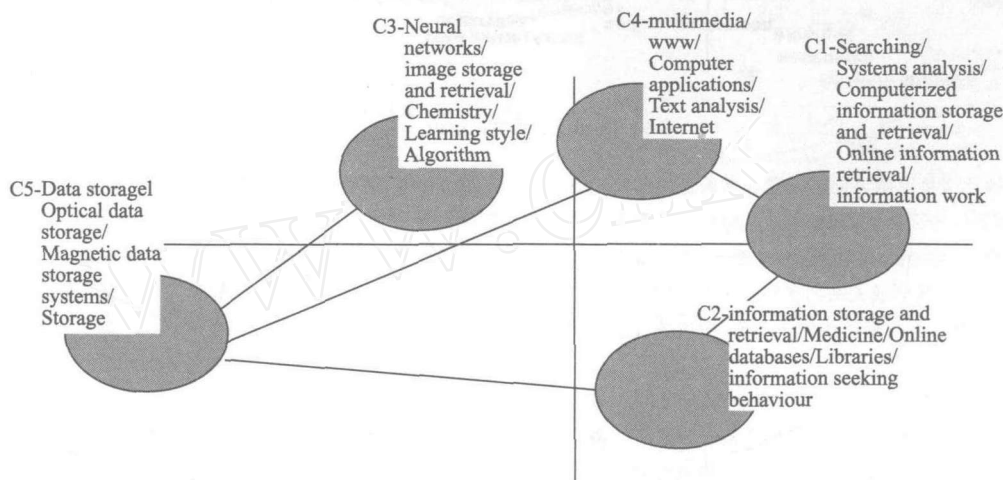


图4 1987-1997年信息检索领域的整体知识关联网络

图4中两个节点之间空间距离紧密说明与其他节点相比,该节点对有着更密切的关联。该网络图清晰地反映了整个信息检索领域中的知识关联关系,从文献资源中挖掘出了更深层次的潜在信息。

Ying Ding 等为进一步细化信息检索领域的知识构成,深入理解各类内部的知识结构,再次利用多元标度技术对这 5 个大类分别进行分析,其中对 Cluster 1 的分析结果如图 5 所示。

该思路由于具有实现方法易理解且可以借助现有软件进行分析等优点,成为基于空间分布的文本挖掘研究的主流方法。Drexel 大学的 Theodore Allan Morris 基于该方法对医药情报学的学科结构进行了考察^[6], Royal School of Library and Information Science 的 Irene Wormell 等利用该方法,对福利理论的发展和在现代福利国家中福利的实践工作进行了研究,挖掘“福利”这一概念研究历史、研究现状和研究趋势的认识^[7]。并且出现了以此思路为核心,利用共现分析方

法考察网站关联结构的研究^[8]。这些都极大地体现和扩展了基于空间分布的文本挖掘研究的范围和意义。

2 在基于时间分布的文本知识挖掘中的应用研究

基于时间分布的文本挖掘研究是通过比较同一时间段内网络图中的不同子网络或比较不同时间段的知识关联网络,获取对知识组成的分化和渗透、知识领域发展及消亡趋势的认识,是对基于空间分析研究结果的分析和引申。目前,这类研究的手段有静态预测和动态跟踪两种。

2.1 静态预测

静态预测是指以固定时间段的基于空间分布的文本挖掘研究结果为对象,选用有探测意义的分析指数对主要知识类别进行分析,根据分析的结果展望未来的发展趋势。其中最为常用的分析指数包括密度(density)、向心度(centrality)和向心度与密度的比值(c/d)。

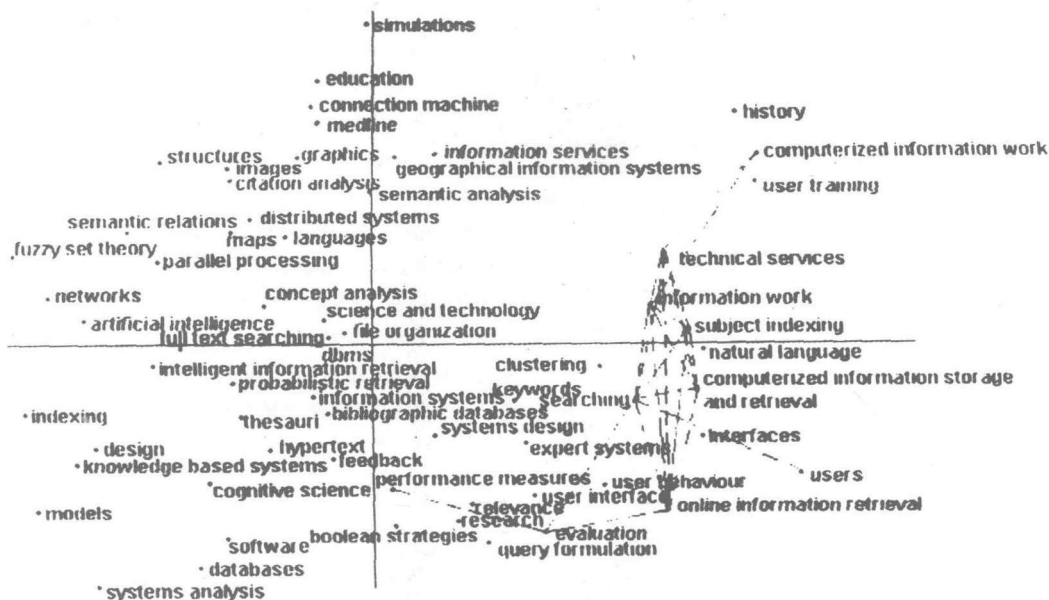


图5 Cluster 1的细化结构

密度主要用来量度使字词聚合成一类的这种联系的强度,也就是该类的内部强度,它表示该类维持自己和发展自己的能力。某一类别的密度计算可以有多种方式,如首先计算本类中每一对主题词或关键词在同一个文本中同时出现的次数,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。

向心度主要用来量度一个知识领域和其他知识领域相互影响的程度。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,向心度可以通过该类别的所有标引词或关键词与其他类别的标引词之间链接的强度进行计算。这些外部链接的总和、平方和的开平方甚至是前6个外部链接的和,都可以作为该类别的向心度。

向心度与密度的比值表明了科学和技术的发展阶段。如果比值接近1,表明这个领域在研究网络系统中处于主流地位;如果比值远离1,表明这个主题有可能在网络系统中消失。

以向心度和密度为参数绘制成二维坐标,可以概括地表现一个领域或子领域的结构。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。这个图将每一个二维空间的题目领域划分为4个象限。一般而言,在第一象限的主题领域内部联系紧密并处于研究网络的中心地位。第二象限的主题领域结构比较松散。这些领域的工作有进一步发展的空间,在整个研究网络中有较大的潜在重要性。第三象限的主题领域内部链接紧密,题目明确,并且有研究机构在对它进行正规研究,但是在整个研究网络中处于边缘。第四象限的主题领域在整体工作研究

中处于边缘地位,重要性较小。战略坐标各个象限的特点如图6所示^[9]。

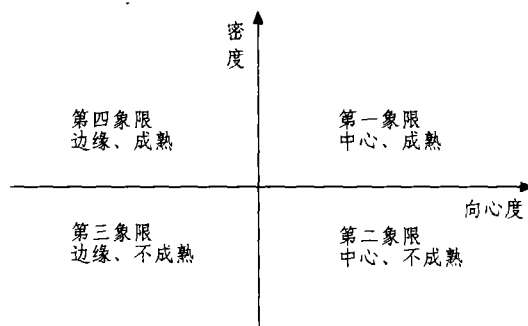


图6 战略坐标及其含义

研究表明:在第二、三象限的主题随着时间的推移而变化的程度要比在第一象限的主题变化程度大;第二象限中主题往往会经历一个内部结构重组的过程以提高耦合程度,增强内部关联;第三象限中的主题范围则有可能扩大,以吸收网络中其他领域的知识。促使出现这些变化的原因可以理解为是由于科研人员都有使自己的研究主题中心化的意愿,或者通过扩展研究范围,或者通过对研究领域概念的深入阐述来实现这种改变^[10]。

2.2 动态跟踪

跟踪不同时间段网络组成和关联的变化能够直观和准确地获取知识的产生、发展、演化和消亡趋势的认识。然而,实际操作中很难通过直观观察找到前后两幅网络图中代表相同知识组成的类别,因此也就无从理解知识组成变化的特点。针对这一困境,Coulter等提出了采用相似指数 SI (simi-

larity index)以比较不同时期的类别相似程度。相似指数 SI 的定义如下:

$$SI = \frac{2c_{ij}}{c_i + c_j}, \text{其中 } c_{ij} \text{ 是类别 } i \text{ 与类别 } j \text{ 中共有的词汇数}$$

目, c_i, c_j 是类别 i 和类别 j 中各自总共包含的词汇总数。

这样,找到相似系数较高的类别组,比较不同时间段类中词汇的变化可以理解该类研究内容和方向的变化,比较在战略坐标中所处位置的变化能够判断该类进一步的发展趋势。

此外,还可以找到研究领域中的“生命周期曲线”。如果分 3 个时间段对知识网络分析,发现第一个时间段内的类 A 与第二个时间段内的类 B 高度相似,而类 B 又与第三个时间段的类 C 高度相似,则称这个相似类集为一个系列。显然,网络越稳定,就存在越多的系列类。这些系列类的存在,可提供充分的信息以理解随着时间的发展各类演变的情况^[11]。

3 在基于内外关联映射的文本知识挖掘中的应用研究

文本的知识关联应该包含两个方面:文本内容关联和文本外部特征关联。两者相互关联,相互影响,统一于一个整体。但是长期以来对文本挖掘的研究却往往采用二者取其一的思路进行分析,或分析文本内容构成,或分析作者或刊物外部的分布规律。这种人为的割裂必然会导致许多有意义信息的损失。将共现分析方法应用于基于内外关联映射的文本挖掘研究,则能综合文本内容特征和外部特征,形成对研究目标更为准确的认知。

基于内外关联映射的文本挖掘研究的思路是根据文本的内容关联关系,映射文本内容对应的作者、研究机构、期刊等外部特征的空间分布和变化规律。与以往的外部特征关联分析相比,该研究突破了实际意义上的合作关系,能够挖掘到作者、研究机构、期刊等因内容的关联带来的潜在语义联系。

在 1995 ~ 1998 年的第 4 届 European Framework Programme 项目中, the Department of Technology Management of ARC Seibersdorf research GmbH 应用共现分析方法和可视化技术将知识以二维图的形式展现。研究的目的在于确定特定研究主题在整个学科空间中的位置,以及找到各个主题领域内的主要作者或机构以反映他们之间的间接合作关系^[12]。这是一个典型的基于内外关联映射的文本挖掘研究案例。在构建机构的关联网络时,该项目不再单纯依据机构在文献或研究项目中的实际合作构建关联,而是依据文献或研究项目中出现的相同文本内容形成关联关系。也就是说,如果两个研究机构各自发表的文献使用了相同的关键词,则这两个研究机构之间就产生了一定联系。这样,即使研究机构之间并不存在实际的合作关系,但只要研究内容相同,这些机构也能聚类在一起。图 7 为以该思路为依据形成的部

分机构网络,观察该图可见,研究替代能源(AF)的研究机构集中在图的下半部分,而研究 fuel cells(FC)的机构则主要分布在图的上半部分较广的区域内。同时研究两个主题的中心研究机构也在图中显示了出来。图中圆的大小与研究机构在所有出版物中使用的关键词数对应。

以同样的方式可以将作者的网络关联展现出来。作者网络关联的典型形式如图 8 所示,其中存在孤立作者(IA)和作者团队(Teams),以及连接不同团队的“连接点”作者(Link)。

得到研究机构(作者)的潜在关联网络之后,需要选用合适的指数对网络结构进行分析,获取对网络更深入、更准确的认知,找到最核心的研究机构(作者)和有相同研究内容的研究机构(作者)群。

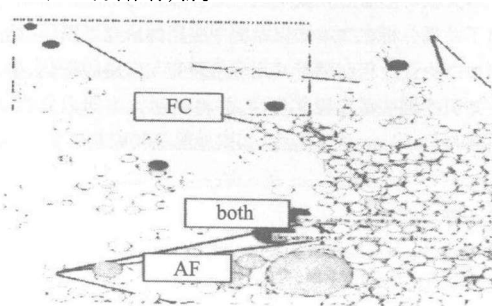


图 7 燃料领域的主要研究机构网络

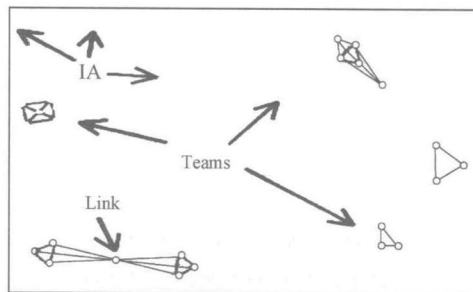


图 8 合作作者网络的典型形式

在 1998 ~ 2002 年的第 5 届 European Framework Programme 项目中, the Department of Technology Management of ARC Seibersdorf research GmbH 进一步深入探讨研究机构的关联网络,找到网络中的核心机构并分析整个网络的合作紧密情况。主要利用作者关联紧密度和组间关联集中指数对各网络节点进行比较^[13]。

作者关联紧密度能够反映作者在网络中出现的频繁程度。有较高作者关联密度的节点与许多节点有联关系并且在网络中占据主导位置。因此,有高作者关联密度的节点通常处在从一个节点到另一个节点的最短路径上。可以将节点 n_i 的作者关联密度定义为:

$$C_B(n_i) = \sum_{j < k} g_{jk}(n_i) / g_{jk}$$

其中, g_{jk} 代表路径总数。

这样, 作者关联密度指数值为0。取值越大, 该节点在所在网络中的“交流性”越好。

组间关联集中指数是假设组内所有节点具有相同的作者关联紧密度时该组节点的最小整体关联值, 将其定义为:

$$C_B = \frac{2 \sum_{i=1}^g [C_B(n^*) - C_B(n_i)]}{(g-1) \wedge 2(g-2)}$$

其中, $C_B(n_i)$ 代表节点 i 的关联紧密度, $C_B(n^*)$ 是整个节点集中最大作者关联密度值。

4 总结

本文在对已有研究成果进行综合考察与分析的基础上, 提出了共现分析在文本知识挖掘中应用的研究思路, 并结合实例加以论证。相信随着共现分析研究与应用的深化, 该方法在知识挖掘中将发挥其作用, 并将推动文本知识分类、概念空间与 Ontology、语义检索、知识地图等领域的研究。

参考文献

- 1 Kostoff R. N. . Database Tomography: Multidisciplinary Research Thrusts from Co-word Analysis. Proceedings: Portland International Conference on Management of Engineering and Technology, 1991. 2005-08-10
- 2 王曰芬, 宋爽, 苗露. 共现分析在知识服务中的应用研究. 现代图书情报技术, 2006(4)
- 3, 10 He Qin. Knowledge Discovery Through Co-Word Analysis. [2006-03-29]. http://www.findarticles.com/p/articles/mi_m1387/is_1_48/ai_57046530
- 4 Coulter N. , Monarch I. , Konda S. , Carr, M. . An Evolutionary Perspective of Software Engineering Research Through Co-Word Analysis. [2006-04-20]. <http://www.sei.cmu.edu/publications/documents/95.reports/95.tr.019.html>

- 5 Ding Y. , Chowdhury G. G. , Foo S. . Bibliometric Cartography of Information Retrieval Research by Using Co-word Analysis. Information Processing and Management, 2001, vol. 37 : 817 ~ 842
- 6 Morris T. A. . Structural Relationships within Medical Informatics: a Classification/indexing Co-occurrence Analysis , Drexel University(PhD) , 2001
- 7 Wormell I. . Bibliometric analysis of the Welfare Topic. Scientometrics, 2000, 48 (2)
- 8 Polanco X. . Clustering and Mapping Web Sites For Displaying Implicit Associations and Visualizing Networks. [2006-04-20]. http://www.math.upatras.gr/~mboudour/articles/web_clustering&mapping.pdf
- 9 崔雷. 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 , 22(4)
- 11 He Qin. Component study of co-word analysis. University of Illinois at Urbana-Champaign(PhD) , 2001
- 12 Widhalm C. , Gigler U. , Kopcsa A. , Schiebel E. . Co-occurrence and Knowledge Mapping to Identify Hot Topics and Key Players in the Field of Mobility and Transport. [2006-04-20]. http://www.semantic-web.at/file_upload/root_tmpphpA4EeZH.pdf
- 13 H · rlesberger M. , Dachs B. . Behaviour of the Trans-border Co-operation within the European Framework-Programme. [2006-03-29]. http://www.eicstes.org/EICSTES_PDF/PAPERS/Behaviour.pdf

王曰芬 南京理工大学经济管理学院副教授。通信地址: 江苏南京。邮编 210094。

宋爽 卢宁 朱辉 南京理工大学经济管理学院硕士研究生。通信地址同上。(来稿时间: 2006-07-06)

(接第 58 页) 利用研制的词汇编辑器、框架元素编辑器、框架关系编辑器和计算机辅助语义标注器等工具, 一定程度上可减少工作量, 提高工作效率, 最终实现基本概念的结构化描述, 以机器可读形式对语义知识编码, 为信息科学领域本体论的深入研究提供了参考作用, 同时为今后自动语义标注、问答系统、信息抽取、信息检索的应用奠定了基础。

参考文献

- 1 Ruppenhofer Josef, Michael Ellsworth. FrameNet: Theory and Practice. [2005-11-18]. <http://framenet.icsi.berkeley.edu>
- 2 Fillmore, Charles J. Frame semantics. In Linguistics in the

Morning Calm, Seoul, Hanshin Publishing Co. 1982: 111 ~ 137

- 3 Fillmore, Charles J. Frames and the semantics of understanding. In Quadern di Semantica, 1985, Vol. 6. 2
- 4 冯志伟. 现代术语学引论. 北京: 语文出版社, 1997

贾君枝 山西大学管理学院副教授, 博士。通信地址: 太原。邮编 030006。

陈幼华 上海大学图书馆科技情报所。通信地址: 上海。邮编 200436。(来稿时间: 2006-06-20)