词共现分析在构建概念空间中的应用研究综述*

冀玉静 李军莲 李 芳

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 利用共现分析构建概念空间，实现语义检索，是当前信息组织和检索领域研究热点之一。阐明概念空间的定义、意义、应用及构建方法，介绍词共现分析技术的内涵、应用前提、演进历程、步骤与方法，从多个角度系统综述词共现分析在构建概念空间、本体、揭示语义关系等方面的应用状况，为构建基于概念空间的信息检索可视化系统研究奠定基础。

〔关键词〕 词共现分析；概念空间；本体；语义关系

Application Research Review on Co – word Analysis in Structuring Concept Space *JI Yu – jing, LI Jun – lian, LI Fang, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Using co – word analysis to construct concept space and realize semantic retrieval is one of the research hotspots in information organization and retrieval fields at present. The paper clarifies the definition and significance, application and constructing methods of concept space, also introduces the connotation, application premise, evolution process, steps and methods of co – word analysis technology. From many aspects the paper systematically generalizes the application situation on construction of concept space, ontology and semantics relationship digging, in order to provide basis for constructing visual retrieval system.

〔**Keywords**〕 Co – word analysis; Concept space; Ontology; Semantic relationship

1 概念空间概述

1.1 概念空间的定义

概念空间（Concept Space，简称 CS）是 1983 年邓珞华在美国康奈尔大学教授 G. 索尔顿的信息检索系统代数模型基础上首次提出的[1]，发展至今尚没有明确清晰的定义，不同的专家、学者根据不同的研究目的给出了多种解释[2]。目前有关概念空间的定义大致可分为两种观点，一种观点认为概念空间是由概念及以概念为结点的语义关系网组成[3]。另一种观点认为概念空间是指以词共现为依据，由对象（Object）内的词相互链接构成的词语关联图[4]。综上所述，凡是涉及概念及其相关关系的集合可以认为是概念空间。目前流行的主要包括受控词表、知识库、领域本体等。

1.2 概念空间的意义及应用

概念空间是对知识概念及其相关关系的抽象和

〔修回日期〕 2010 – 04 – 13

〔作者简介〕 冀玉静，编辑，发表论文数篇。

〔基金项目〕 中国医学科学院医学信息研究所/图书馆基本科研业务费专项课题“基于词共现的可视化中文医学概念空间研究”（项目编号：08R0125）。

描述，不仅在语义检索、文献标引、帮助用户表达信息需求等领域广泛应用，而且在其他知识组织领域也显现出无限的潜力。特别是 Ontology 一类的高级概念空间的应用领域已扩张到知识整合、远程教育、知识管理与分类、电子商务等[2]。

1.3 概念空间的构建方法[2,4]

（1）利用现有的词典：一般是由各领域的专家，或者是在语言学家的帮助下通过完全人工或机器辅助的方式完成，缺点是耗时耗力；（2）自动生成词典：基于计算统计学思想，通过对大量的电子信息资源进行抽词、聚类，由此实现大规模地自动生成概念空间，但由于机器语言的局限性，必须经过大规模的增补和修改，这比新建一个词表更为困难和复杂；（3）利用现有词典，同时结合自动生成的技术，弥补前两者的不足。随着自然语言处理技术、词共现理论、信息可视化技术等的出现和发展，为自动生成概念词语及构建词语之间的关系提供了新选择，其中共词分析以其方法的简明性和分析结果的可靠性，在构建概念空间方面的应用取得了很好的效果。

2 词共现分析技术概述

2.1 词共现分析的内涵[5-6]

词共现分析方法，又叫共词分析、共现词分析，是指在一个足够大的文献集合中，两两统计一组词在同一篇文献中出现的次数（即共现频率），并利用包容系数、聚类分析等多种统计分析方法，以数值、图形等直观形式揭示出这些词间的关系，进而分析这些词所代表的概念结构和关系。它属于内容分析方法的一种，具体又包括共词聚类分析、共词关联分析等。

2.2 词共现分析的应用前提

运用词共现分析技术要基于以下假设：（1）作者所用的技术术语都是经过认真考虑和选择的；（2）同一篇文章使用不同的术语，就意味着它们之间存在一些关系要引起重视；（3）如果有足够多的不同作者都认可同一种关系，那么可以认为这种关系在所属学科领域具有一定意义；（4）用主题词进行共词分析的前提是标引人员所选的用来描述文章内容的主题词是值得信赖的。

2.3 起源和演进历程[5]

2.3.1 基于包容指数、邻近指数和相互包容指数 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。在包容指数和邻近指数基础上，Callon 等在 1986 年提出包容地图（Inclusion Map）和邻近地图（Proximity Map）的概念。

2.3.2 基于战略坐标的共词分析方法 1988 年 Law 等采用战略坐标（Strategic Diagram）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。它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。

2.3.3 基于数据库内容结构分析的共词分析方法（Database Tomography，DT） Kostoff 等 1995 年提出 DT 分析方法，用于大量数字化文本资源的分析。包含两个参数：频率分析（Frequency Analysis）和邻近分析（Proximity）。频率分析用于揭示数据库中较深入的主题；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。

3 词共现分析的步骤和方法

3.1 数据抽取[6-7]

3.1.1 全文直接抽取 即使用专门的软件工具从全文文本中直接抽取。优点在于可以避免标引偏差，解决由于词表更新不及时而无法监测特定领域新内容的问题。但是由于自然语言中存在大量同义词，如不进行合并同义词的预处理而直接统计，势必对结果造成很大干扰，影响结果的可靠性。

3.1.2 字段间接抽取 即从数据库的加工著录字段，如主题词、关键词等抽取。优点是这些字段的用词比较规范，较能反映文章的主题，无须合并同义词等预处理过程，但同时也存在标引用词的选择

偏差问题，会降低分析结果的客观性。但 1992 年 Law 和 Whittaker 的研究表明对数据库标引质量的担忧是不必要的。

3.2 根据共现频次构造共词矩阵[6-7]

由于词汇量巨大，不能比较所有词汇对的相似度，因此通常按照一定规则设定阈值，选择一定数量的高频词进行比较。目前对于高频词阈值（高频词的词频总和占所有词频总和的比值）的划定没有统一见解。主要有两种方法，一种是研究者凭经验在选词个数和词频之间取得平衡，另一种是结合齐普夫第二定律辅助判定高频词的阈值。选定后的 N 个高频词可以形成 N × N 的共词矩阵。由于词对频率是绝对值，所以难以反映词与词之间真正的相互依赖程度，因此有必要对其进行处理、分析，以反映两者联系的紧密程度。

3.3 数据处理和分析

3.3.1 分析词汇的关联度[7] 主要测度方法有 Dice 指数、余弦指数、Jaccard 指数，计算公式分别如下：

$$\text{Dice's Coefficient} = 2Cij/(Ci + Cj)$$

$$\text{Cosine Coefficient} = Cij/(Ci \times Cj)\ 1/2$$

$$\text{Jaccard Coefficient} = Cij/(Ci + Cj - Cij)$$

其中 Cij 是词 i 与词 j 共现的次数，Ci、Cj 分别是词 i 和词 j 在文本集中总共出现的次数。由于 Jaccard 指数能根据词的共现频率直接反应两个词之间的相似度并且消除高频词的消极影响，因此被广泛用做代表两词之间的标准化相关系数。

3.3.2 映射文本内容结构[5-7] 根据计算出的词汇关联度将词汇聚类，每个簇代表研究领域的一个子领域。然后再利用其它衡量指标分析这些簇，从而确定领域内部关联关系和核心领域，并可以比较不同时期的簇，以发现研究领域的发展变化。主要的衡量指标有包容指数（Inclusion Index，Iij）、邻近指数（Proximity Index，Pij）和等价指数（Equivalence Coefficient，Eij）3 种。包容指数：Iij = Cij/min（Ci，Cj），主要用于处理高频主题词，以确定特定时期的主要研究领域。邻近指数：Pij = （Cij/CiCj）× N，一般用于低频主题词，以确定发展迅速的次要研究领域。之后由于研究人员更关注主要研究领域的确定，于是又提出等价指数（又称相互包容指数），Eij = （Cij/Ci）×（Cij/Cj）= Cij^2/（Ci × Cj），Eij 值介于 0 ~ 1 之间。该指数是目前各种共词分析文献中用得较多的一种。

在以上几个公式中，Cij 代表词对 i 和 j 在文献集合中的共现频次，Ci 代表词 i 在文献集合中的出现频次；Cj 代表词 j 在文献集合中的出现频次；min（Ci，Cj）代表 Ci 和 Cj 两个频次取最小值，N 代表文献集合中文献的数量。

3.3.3 利用可视化技术展现共词分析的结果[4,7] 可视化将数据信息转换成几何的形态，有助研究人员发现潜在信息。多维尺度（Multidimensional Scaling）、Kohonen 自组织图和路径寻找网络（Pathfinder Networks）是主要的 3 种方法。可视化技术与共词分析相结合能够获得对文献内容关联、学科结构等的直观认识，更能体现共词分析优于其它文本分析方法的独特魅力。

4 词共现技术在构建概念空间中的应用

4.1 构建概念空间和 Ontology

词共现分析经过大约 30 年的发展完善，凭借其方法的简明性和分析结果的可靠性，在自然语言处理、文献计量学、科学计量学、信息检索、人工智能等众多领域的应用日益广泛。特别是在计算机技术的辅助下，词共现分析在揭示语义关系，构建概念空间和 Ontology，提高知识组织效率等方面显示出独特的优势。利用共现分析构建概念空间和 Ontology，以实现语义检索，是当前信息检索领域一大研究热点。国内外已经在这方面进行了诸多有益的尝试。Yi - Fang Brook Wu 2001 年提出了“Probability of Co - occurrence Analysis”（POCA）方法来自动构建概念空间[8]。通过综合利用共现分析和现有词表或分类学知识，Takeshi Morita 等 2005 年开发了 DODDLE - OWL 本体构建项目[9]。Ying Ding 在构建 IR 和 AI 本体时首先利用共现分析获得具有语义关系的关键词对，随后利用现有的领域词

表提供的 BT/NT 关系丰富词间层次关系[10]。美国国立医学图书馆（NLM）在其长期建立的一体化医学语言系统（UMLS）中开展了这方面研究[11]。国内这方面的研究并不多见，2006－2007 年期间张学福在其主持的“基于摘要信息的可视化检索研究”项目中尝试开展了“基于词共现的可视化概念空间研究”，初步认为基于词共现构建自然语词概念空间是可行的[4]。

4.2 揭示语义关系

Douglas L. T. Rohde 等 2004 年在“An Improved Model of Semantic Similarity Based on Lexical Co－occurrence”一文中提出了一种能够从大样本数据集中自动生成词汇含义的向量空间模型[12]。Neal Coulter 等人利用词共现分析映射软件工程研究领域的核心特征[13]。Cimino 等对主题词和副主题词的组配规则进行研究[14]。崔雷等尝试利用共词分析来抽取书目文献数据库中主题词/副主题词之间的语义关联规则，获得具体的药物与疾病之间的关系，并证实用这种方式抽取的关系高度可靠[15]。此外还有人利用关联规则算法对大型医学电子病历数据进行共词分析，构建了病人所接受的检查项目和最终诊断结果之间的关联规则。

5 结语

随着词共现分析在理论和方法上的不断完善，其应用领域也将逐步扩大和深入。该方法能克服传统的专家调查法（德尔菲法等）花费较大、操作复杂的弊端，准确、快捷地获得相关知识。并且由于可视化技术的进步，词共现分析方法将在构建概念空间和本体中发挥更大作用，实现更有效的知识组织和管理，从而为进一步开展高效的知识服务奠定坚实基础。

参考文献

1 邓珞华. 概念空间——定义、意义和局限〔J〕. 情报学报，2003，22（4）：393.

2 王国琴. 基于语义检索的概念空间研究〔D〕. 南京：南京理工大学，2004.

3 邓珞华. 图书情报教学〔M〕. 长春：东北师范大学出版社，1983.

4 张学福. 基于词共现的可视化概念空间研究〔J〕. 情报学报，2008，27（2）：205－211.

5 冯璐，冷伏海. 共词分析方法理论进展〔J〕. 中国图书馆学报，2006，32（162）：88－92.

6 钟伟金. 共词分析法研究（一）——共词分析的过程与方式〔J〕. 情报杂志，2008，（5）：70.

7 王曰芬，宋爽，苗露. 共现分析在知识服务中的应用研究〔J〕. 现代图书情报技术，2006，（4）：30.

8 Yi－fang Brook Wu. Automatic Concept Organization：organizing concepts from text through probability of co－occurrence analysis（POCA）〔D〕. Albany：State University of New York（PhD），2001.

9 Takeshi Morita，Yoshihiro Shigeta，Naoki Sugiura，et al. DODDLE－OWL：on－the－fly ontology construction with ontology quality management〔EB/OL〕.〔2009－11－25〕. http://www.ei.sanken.osaka－u.ac.jp/iswc2004/posters/PID－JURPMVUS－1090083983.pdf.

10 Ying Ding. IR and AI：using co－occurrence theory to generate lightweight ontologies〔EB/OL〕.〔2009－11－25〕. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.835&rep=rep1&type=pdf.

11 UMLS® Reference Manual〔EB/OL〕.〔2009－11－27〕. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch03#ch03.I33_Descriptions_of_

12 Douglas L. T. Rohde，Laura M. Gonnerman，David C. Plaut. An Improved Model of Semantic Similarity Based on Lexical Co－occurrence.〔EB/OL〕.〔2009－11－27〕. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.9401&rep=rep1&type=pdf.

13 Neal Coulter，Ira Monarch，Suresh Konda，et al. An Evolutionary Perspective of Software Engineering Research Through Co－word Analysis〔EB/OL〕.〔2009－11－27〕. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.3479&rep=rep1&type=pdf.

14 J. J. Cimino，G. O. Barnett. Automatic knowledge acquisition from MEDLINE〔J〕. Methods of Information in Medicine，1993，32（2）：120－130.

15 崔雷，李丹，冯博. 动用主题词/副主题词关联规则在医学文献检索系统中抽取知识的尝试〔J〕. 情报学报，2005，24（6）：657－662.