中国图书馆学报(双月刊) 1997 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

●王知津

域分析与情报检索理论

ABSTRACT The author discusses domain analysis, a new method in information science focusing on the analysis of domain knowledge. Meanwhile, the author makes a comparative study of some theories and methods in domain analysis and information retrieval, such as statistical and probabilistic search, cognitive user modelling, expert agent system, hypertext system and citation retrieval technique.

KEY WORDS Domain analysis. Information science. Information retrieval.

CLASS NUMBER G354

1 引言

在对知识的理解方面, 80 年代占主导地位的是比较形式主义和类似计算机的方法。进入 90 年代后, 在知识理解领域开始出现了一种重要的新观点。这种观点强调知识的社会的、生态的和面向内容的本质, 它与 80 年代的主导观点相反。这个观点就是域分析(domain analysis)。

进入 90 年代后, 域分析观点在教育学、心理学、语言学和科学哲学等学科之间形成了一种跨学科趋势。由于上述学科都是与情报科学关系极为密切的相关学科, 所以, 作为一种新理论、新方法, 域分析也日益受到情报科学界的关注, 国外已发表了一些这方面的文章。这表明, 情报科学的域分析方法与相关学科中所发生的变化和趋势是一致的。

域分析方法认为, 为了透彻理解情报科学中"情报"的概念, 最好的办法就是研究作为思维群体或话语群体的知识域。知识组织、情报语言、情报交流、情报系统及其相关性判定都是这些群体工作客体的反映, 也是这些群体在社会中作用的反映。因而, 情报心理、情报知识、情报需求、情报行为等都应当从这一观点加以考察。

本文试图把域分析同当代情报科学中的一些理论与方法, 特别是情报检索的一些理论与方法进行比较, 从中探讨域分析方法在情报检索中的应用。

2 统计与概率检索

按相关程度排序文献, 按相似程度聚类文献, 以构造算法为目标, 检查词语在单篇文献和文献集合中的分布, 这是统计检索或概率检索的基本原理, 在这方面已经进行了大量的实验研究。要想发展这种检索方法, 可以采用计算机语言技术, 即自然语言处理, 也就是说, 在分析句法结构和词汇信息的统计文本分析的基础上, 对词语分布加以补充。然而, 在所有这样的研究中, 几乎根本没有考虑到像域和学科这样的一些概念, 而是把一个词语看成某种具体的东西, 独立于它所出现

的知识域。当然,我们知道这样一个事实:一个语法分析程序可能依赖于来自某一特定域词典的词汇信息,或者用同形异义词的辨义规则予以补充。但是,后一种方法所面临的是解释语言的实际问题,这是因为如果对某一特定文本的总体信息和目的没有透彻的理解,就不可能全面理解某一给定言词,如一个从句。

从语言学的角度来看,词语的含义是由社会语言劳动的划分决定的,也就是说,同一个词在不同的社会劳动部门(即不同的思维和话语群体)可能有不同含义。按照域分析结构,词的含义只能从它所出现的上下文中去理解。例如"黄金"这个词的含义只能通过解释该词所出现的话语而得到正确理解。"黄金"至少有如下三种含义:(1)化学含义,它是一种重金属,不溶于酸和碱;(2)经济含义,常规的经济测量和储备;(3)小说含义,与财富、幸福、国王、王后有关。还有哪些词与词表中"黄金"一词有关,完全取决于该词表是做什么用的。通过某一算法用该词检出的文献是否与某一提问相关,完全取决于该词是否有一个或其他含义。对于这些问题的解释,对于知识与含义的上下文、会话以及历史特征,统计与概率检索方法似乎是茫然的。

人们对统计与概率检索方法已越来越感到不满意,尽管在改进这一方法方面已经做了许多有益的工作,但可以预料,这一方法的检索性能不会有进一步的提高。

3 认知的用户模型化

关于认知的用户模型化的基本假设是:情报系统应当有一个用户模型(或他们的情报需求),同时,应当按照这一模型确定文献。当然,在一定情况下,这种模型化是可能的。而在某些情况下,可能阐明知识中的某些特定先决条件,并针对这些假定的先决条件来进行知识的排序。例如,中小学生不具备获得众多领域中科学知识的先决条件。从理论上说,建立一个供学生使用的情报检索系统是可能的,并且已经在像"Choice"这样的系统中实现了。不过,这个系统从来未宣布过它是在诸如认知模型化原理基础上建起来的。

丹麦的一个叫做"The Book House"的系统就是一个基于画像的供小说检索用的联机公共存取目录。该系统的界面以及文献的主题分析都是为了优化画像对组织和表示域组织的应用,而以深入研究用户爱好和心理为基础的[1]。

然而,情报检索的典型域是为科学和专业活动(如法律、医学)中问题的求解而确定相关文献的域。把一个专业性的医学情报系统建立在其用户的认识结构上,这种想法是危险的!为了使用一个医学情报系统,用户就必须获得有关医学、医学理论和医学术语等必要的知识。用户不是一成不变的,他们的知识和假设在变化着,而这些知识和假设都是在这个域中保持高度同步的作者与读者认知的组成部分。情报系统的主要任务是反映这个域,而不是个体用户。

情报科学研究中的精神论方法,主张通过研究用户的行为或想法,可以找出情报科学的一些隐含的原理、定律或规律,并用于情报系统的设计。这就是说,情报系统应当反映用户对知识和情报的主观感知,而不是有助于用户知识发展的客观现实。然而,问题在于,用户有关情报源和检索策略等的知识往往是不全面的和有缺欠的,因此,为情报系统设计提供依据的不应当是用户,而应当是情报专家。当然,为了帮助用户提高自己的行为质量,研究他们的不全面或有缺欠的行为也是有用的;不过,这是根据情报科学原理来对那些行为进行判断,与精神论方法研究情报用户行为有本质上的不同。由此可见,情报系统用户有自己的感知,而这种感知与当代科学感知又是不一致的,所以,他们不应成为情报系统设计的认知权威,认知主义方法关心

情报系统对用户友好, 主张情报科学原理与数据库内容无关, 或与其客体分类无关, 而与情报系统的用户界面有关。但是, 与情报系统用户的知识相比, 知识理论中的问题更加基本, 因为用户无法表达他们还没有想到的需求。可以得出结论: 精神论和认知主义是矛盾的; 应当承认, 认知的用户模型化研究已经不能超出其原型阶段。

4 专家中介系统

人工智能和专家系统是计算机科学中的子域和方法, 对情报科学的许多研究者也有启发, 因而也成为情报科学中的一种方法。但是, 一般的人工智能和专家系统程序过于乐观, 它们只是以对知识的某些相当原始的感知为基础。

心理学的认知革命是在 50 年代中后期开始的。那时, 人们把人脑看成计算机, 并研究了人脑的天生程序。人们的愿望是把人的推理变成一种演算, 因为推理无非就是命题的演算, 所以人们按照一些独立于内容的逻辑来进行推理研究。其结果是, 引导推理的是任务的内容, 而不是任务的形式结构, 内容是决定的。于是, 人们越来越倾向于把个体心理状态看成社会构成物, 而不再把人脑的机器和程序看成与社会文化背景隔离。与此相适应, 70 和 80 年代出现了依赖于内容的推理的种种迹象。但是在结构与内容之间, 在相关情报无关情报之间, 不能做出简单的唯一的划分。因此, 就一个域的推理而言, 相关结构需要特定域理论, 进而, 从某种意义上说, 如果人的推理是对特定环境(包括社会环境)的一种适应的话, 那么, 作为对重要现时环境结构适应的推理机制的生态学分析, 作为对重要现时环境结构适应的推理机制的进化论分析, 以及作为对重要过去环境结构适应的推理机制的进化论分析, 都是必不可少的。

从对人工智能运行的某些前提的上述分析中, 我们对以下情况就不难理解: 从 80 年代后期开始, 人工智能、专家系统、第五代计算机等的一般程序陷入极大的困难, 并且明显地萧条下去, 但没有停止。

除了人工智能的一般程序遇到的困难以外, 事实上, 情报检索似乎特别不适合这种系统。H. M. Brooks 得出的结论是: "由于种种原因, 对于某一专家系统应用来说, 情报检索似乎不是一个理想的问题域。这是一个既没有界定好又不是狭窄或相似的域。打算建立一个进行智能检索的专家系统, 现在看起来似乎是行不通的"[2]。D. R. Swanson 认为, 沿着这条道路走下去的情报科学已经开始寻找梦幻目标[3]。

在一般的人工智能程序和情报检索专家中介方法中, 也对域知识产生了某些兴趣, 如结合了某种特定用户域知识的情报检索系统 I^{3}R[4]。此外, L. M. Fu[5] 以及 J. M. Vlissides 和 M. A. Linton[6] 等人也讨论了如何使域知识模型化的问题。在 I^{3}R 系统中, 询问用户该域中的概念和关系, 并在用户知识的基础上, 逐步编出一个词表。可以认为, 这表示了构造某一特定学科域词表的一种渐进的实用的方法, 这一方法不必说明系统建立当时知识库中的全部概念关系, 尽管在与用户交互前指定某些关系是有用的, 而这些关系为该系统提供了至少是初步的域知识手段。

在域知识的人工智能方法中, 对知识或语言的哲学或社会学的兴趣很小。当然, 也有例外[7]。一般地说, 如果可能就包括某些域知识。关于如何最好地表示域知识的较宽理论现在还很缺少。最经常的是把单个用户或课本看作认知权威。

把一个用户或课本当作权威的倾向体现出计算机科学与图书馆和情报科学之间的一个区别。计算机科学主要关心找出可以实现自动化的问题。如为了编出实行医疗诊断和分析化验的一个基于知识的程序, 就要查阅医学方面的课本。重点在一致性上, 还必须忽

略课本之间的差别。即使要冒平凡的风险, 也必须要保证知识是稳定不变的。相反, 图书馆与情报科学关心的是向潜在用户交流最相关和最恰当的知识。从这一观点出发, 考虑课本的唯一性和每个课本的局限性是非常重要的。根本兴趣不在如何使某些多少有点平凡的工作实现自动化, 而是提供高度选择的和相关的知识, 这些知识要向用户提供对某一给定学科理论、课题和方法的尽可能全面的看法, 使用户能够得到通报并根据他们的需求加以选择。在这种情况下, 知识不应当是稳定不变和容易公式化的, 而应当是容易形成文献的和多侧面的。当数学成为计算机科学的一个最重要边缘学科时, 能够提供更一般知识的史学和哲学便成为图书馆与情报科学的一个最重要边缘学科。因此计算机科学和图书馆与情报科学在某一给定域的重点同起点有着很大的区别。

在现有的专家中介系统(如 I^{3}R)中, 域知识作为一个要素、作为许多其他专家(如用户模型构建者、提问模型构建者和浏览专家等)之间的一个专家而被包括进去。但问题是, 如果没有特别考虑到这些其他专家所起作用的域的话, 可以在多大程度上对他们进行设计?至今为止, 阐明域知识和情报科学中一般的情报检索知识之间的交互问题的研究还很少。然而, 专家中介系统的方法对这类问题是个启发, 这个启发本身就是一个有价值的成就。

5 超文本系统

情报检索的许多研究者对超文本产生了浓厚的兴趣。他们把超文本看成最富有成效的方法, 这一结论是在评价情报科学不同趋势后得出的。大多数人认为, 超文本的理论形成可以回溯到 1945 年, 当时 Vannevar Busch 提出MEMEX 机的想法, 其基础是文献或款目之间的关联节。作为一种情报检索方法, 这一想法一直被耽搁, 直到 80 年代它才成为主要趋势。

应当承认, 超文本是一个迷人的研究领域和一种有前途的技术。但它仅仅是一种技术, 而不能取代像域分析这样的理论方法。然而, 一种理论方法却可以阐明一种技术及其可能性。W. B. Rayward 对某些理论观点做了最重要的分析, 这一分析能把超文本纳入与本文的设想有密切关系的理论框架。与大多数人相反, 他认为超文本的想法甚至倒退到 1945 年以前, 可以追溯到 Paul Otlet, 这是一位文献工作进而也是情报科学奠基人。而 Paul Otlet 的某些基本想法被 Rayward 说成是"一种过时的范式: 19 世纪实证主义"[8]。Otlet 所关心的是客观知识, 而客观知识既包含在文献中, 又被文献所隐蔽。他的知识观点是实证主义的和过于简单化的, 分析和组织文献内容只是使某些过程固定不变。在他看来, 我们必须关心的那个文献内容问题都是事实, 而事实几乎无处不有。乍看起来, Otlet 所说的与对用现代批评理论理解超文本感兴趣的其他人所阐述的之间有着惊人的反差。但当以这种方式提出问题时, 这一差别使我们更加密切地注视: 现代超文本系统的某些描写(尤其是以其雄伟的理论的"宏文本"的表现形式)实际上在说些什么。尽管修辞华丽(这已深深地嵌入对某些超文本系统的描写中), 是不是还有被称之为 19 世纪实证主义"剩余物"的东西呢?

超文本中有关知识表示的另一个问题是知识的模块性, 它在超文本中是起关键作用的。插入新知识是否容易? 修改旧知识是否容易? 图书是否可以模块化? 这要视具体情况而定。如果从一本有许多相互参照的结构严谨的书中删去一章的话, 该书的其余部分实际上就变得不可理解了。这就好像从一个蜘蛛网上拉出一条线, 其结果是毁掉整个蜘蛛网。相反, 有些书的章节是独立的, 因而它们具有模块性。

当知识的实证主义观点由孤立的模块或组件构成时, 这种观点就与辨证法和科学范式理论形成鲜明对照。然而, 知识是一个文化统一体, 像万花筒式地不断改变着它的模型。新知识的出现修改了知识整体的结构, 因此知识没有固定的秩序, 模型每时每刻都是新的, 给人诗一样的想象。

域分析方法关心知识的实质、它可能的模块性以及话语中文本的自律性和清晰性。如上所述, 已经证明, 有关知识的理论(如实证主义)对情报系统(如超文本系统)的设计的确具有重要影响。当然, 仅仅维持在知识是模块的或不是模块的这一水平上还是不够的。为了在不同域或不同种类知识(如经验知识和理论知识)确定模块性的质量和数量, 还必须进行某些研究。对不同域中的超文本系统和不同类型的知识进行实验研究, 应当能够确定有关知识的相对模块性的比较具体的知识, 而这一切都是建立在设想研究者意识到这个问题并不只是遵循纯技术的思想路线基础上。

尽管事实上超文本数据库可允许新的机会, 但在这个新技术中仍有许多情报检索的人文科学问题。你仍然会遇到下面这类问题: 使用受控词汇还是非控词汇, 概念间是否有等级结构, 应当把什么优先权给予基于引文的关系和其它类型的关系(如叙词、自由文本关系等)。超文本是一种新技术。允许用新方法组织文献及其部分, 但它不是情报检索中人文科学问题(如不同主题存取点的情报价值)的理论方法。可见在某种程度上, 把超文本看作情报科学中新的理论方法是一种误解。总之, 超文本是一种技术, 对于修补情报科学中的人文科学问题来说, 它确是一片肥沃的土壤。

6 引文检索

启发情报检索中更多研究的一种关系是基于引文的检索技术。基于引文的检索利用文本之间的一种网络关系, 在不同文本中相似的作者或参考文献在这些关系中起到节点作用, 这些节点在检索过程中动态地互联到文本网络上, 情报科学的核心问题是提供某些理论背景, 从这一背景出发, 在所有这些可能的联系和关系之中作出选择, 这种可能性则是无限的。如果向用户提供一个系统, 而这个系统在没有给出主要联系的选择的情况下就有过多的可能性, 那么, 就会使用户的负担过重, 这个系统就是低效的了。能够提供这种理论框架的正是情报科学的域方法。

7 结论

以上我们对照情报科学中其它当代理论、范式和方法, 对域方法进行了讨论。情报检索研究中的传统方法没有注意到各种知识域里的区别, 也没有注意到诸如学科、研究群体、话语群体、历史文化观点等概念似乎起到各种作用。就这些概念和问题而言, 这些理论是空洞的, 它们含蓄地把这一点看成是无关的或无形的。这种对知识、知识发展和研究群体等概念的缺乏跟一定的知识观点有关。把知识理解成独立的模块, 这是一种实证主义或理性主义, 而其反面, 则是更加生态的、整体的理论基础上的认识论观点。

情报科学的发展受到其它相关学科的影响, 如认识科学、语言学、心理学、教育学、计算机科学、社会学和哲学等。但是, 这些影响都是间接的。例如, 计算机科学和人工智能中的重要趋势改变了情报科学同哲学的关系。也就是说, 当像解释学这样比较软的哲学被计算机科学这样比较硬的科学所接受时, 在情报科学中就产生了对解释学的越来越大的兴趣。情报科学的许多研究者正在关注 解释学, 以求从中得到答案。毫无疑问, 解释学可以为情报科学问题的解决提供新的有价值的看法。问题是解释学是否充分。就情报科学中的问题而言, 也应当对科学(下转第 37 页)

紧密的期刊群。

(3) 第三子系统由《图书馆建设》、《图书馆理论与实践》、《大学图书馆学报》、《中国图书馆学报》、《图书馆》和《四川图书馆学报》构成。

(4) 第四子系统由《图书馆》为中点, 和《中国图书馆学报》、《图书馆杂志》、《四川图书馆学报》、《图书馆建设》、《图书馆学研究》组成一个互引网。

(5)《图书馆杂志》与《中国图书馆学报》、《大学图书馆学报》、《图书馆》和《图书馆建设》相互联接, 构成第五个子系统。

除此之外, 从图中可以看出,《图书馆工作与研究》和《图书馆理论与实践》是两个悬挂点, 也即它们和其它期刊联系不够紧密。

这是从一个仅用 9 种期刊、5 年互引量的互引系统所得到的模型和分析结果, 即使这样它仍具有一定的普遍意义, 它给我们指出了在大样本条件下分析的方法及可能性, 对于剖析互引系统的结构和对互引现象的进一步分析奠定了良好的基础, 给出了一个有效的算法及其理论保证。

参考文献

1, 4, 5 汪应洛主编. 系统工程理论与实践. 北京: 高等教育出版社, 1992

2 党亚茹. 著者自引系统的解释结构模型. 情报学报, 1995, 14(2): 129～133

3 翟凤岐, 张芝兰. 我国图书馆学情报学期刊引文进展分析. 中国图书馆学报, 1996, 22(2): 61～67

6 党亚茹. 引文网络系统的结构模型化. 图书情报工作, 1996, (4): 58～61

党亚茹　新疆大学数学系资料室副研究馆员。通讯地址: 乌鲁木齐市, 邮编 830046。

(来稿时间: 1997. 5. 5。编发者: 翟凤岐)

(上接第 22 页)

现实主义、社会构成主义等在情报科学中应用的可能性进行研究。总之, 这些哲学上的差别构成了域分析研究的核心问题。

参考文献

1 Hjorland, B. and Albrechtsen, H. Toward a new horizon in information science: domain-analysis. Journal of The American Society for Information Science, 1995, 46(6), 400～425.

2 Brooks, H. M. Expert systems and intelligent information retrieval. Information proeessing and management, 1987, 23, 367～382.

3 Swanson, D. R. Historical note: Information retrieval and the futnre of an illusion. Journal of the American Society for Information Scienee, 1988, 39, 92～98.

4 Croft, W. B. & Thompson, R. H. I^{3}R: A new approach to the design of doeument retrieval systems, Journal of the American Society ofr Information Science, 1987, 38, 389～404.

5 Fu, L. M. Knowledge-baood connectionism for revising domain theories. IEEE Transa: tions on Systems Man and Cybernelics, 1993, 23(1), 173～182.

6 Ramoni, M. et al. An epistemologieal framework for medical knowledge-based systems. IEEE Transactions on Systems Man and Cyberne-tico, 1992, 22(6), 1361～1375.

7, 8 Rayward, W. B. Visions of Xanadu: paul Otlet (1868～1944) and hypertext. Journal of the American Society for Information Science, 1994, 45, 235～250.

王知津　教授。通讯地址: 天津南开大学信息资源管理系。邮编: 30071。

(来稿时间: 1997. 3. 24。编发者: 刘喜申)