中国图书馆学报(双月刊)2004 年第 4 期
ZHONGGUO TUSHUGUANXUEBAO

●王知津 张国华

知识组织概念模型及相关问题

摘 要 知识组织是一个复杂的、系统的智力过程,共同范围、网罗性、专指性、查全率、查准率、一致性、相关性是这个过程中的主要概念和要素,可以把知识组织中的主要概念及其相互关系结合成一个简单的知识组织概念模型。图 3。参考文献 22。

关键词 知识组织 概念模型 查全率 查准率
分类号 G250

ABSTRACT Knowledge organization is a complex and systematic intelligential process, with common area, wide coverage, specificity, exact search hit rate, etc., as its major concepts and factors. In this paper, the authors propose a simple conceptual model of knowledge organization, which integrates the main concepts and their relationships. 3 figs. 22 refs.

KEY WORDS Knowledge organization. Conceptual model. Wide coverage. Exact search hit rate.
CLASS NUMBER G250

1 引言

知识组织(KO)是在知识分类的基础上发展起来的。近年来,我国学者对知识组织的研究也在逐步深入。在这种形势下,构建一个清晰的知识组织模型不仅是迫切的,而且也必将有利于知识组织研究的延伸和扩展。

在情报学中,通过检查专指性、详尽性、相关性及其他以主题分析为中心的因素,使信息组织和检索的有效策略有所提高。通过将研究过程与知识组织的概念和技术联系起来,研究者、教育者、学生以及从事跨学科(包括图书馆学情报学)研究的评论者将在有效数据准备方面,受益于一种有价值的新方法。

本文提出一个简单的知识组织概念模型,并对该模型中所涉及的问题、主要元素及其相互关系进行分析和探讨,旨在构建知识组织模型,并将它用于研究数据的实践。

2 知识组织概念模型

当对知识组织的主要概念及其相互关系进行广泛解释时,这些概念及其相互关系可为准备分析数据的智力工作提供一个模型。以下部分将把这些概念和关系发展成一个简单的知识组织模型(见图 1),以便比较容易地把它们应用于对研究数据的组织。这个模型在应用于组织研究数据的过程中是非传统的,但其基本概念与把对知识组织的理解应用于建立和评价信息检索(IR)系统时是相同的。

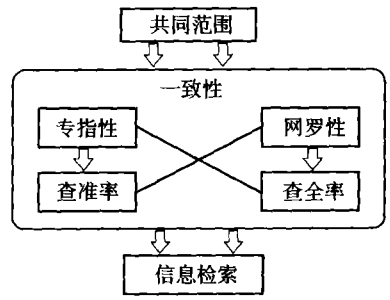

图 1 知识组织概念模型

知识组织模型的焦点是从大量信息中检索出相关的信息。在将此模型运用到研究数据时,相关性是通过所收集和编码的信息是否有助于回答所研究的问题和研究中的假设来确定的。这个标准与用于检索提问或用户需求的相关性概念相似,只是在组织研究数据的情况下,专门是指所研究问题。与确定从信息检索系统中检索出文献的相关性的过程类似,有关研究数据的相关性是一个多层面的概念。可以把有关研究问题的特定数据的相关性看成与论题相关性类似。L. Schamber 等人对相关性的不同流派作了总结,S. Mizzaro 讨论了相关性的详细历史,B. Hjorland 阐述了相关性和近似性之间的关系[1~3]。

相关性类似过于单纯的近似性(Aboutness)的观点,即文献(数据)是关于什么的看法可以直接反映

在标引(或编码)中[4]。然而,对近似性的更广泛的看法可能包括预测用户的需求,而一个对相关性的面向用户的看法,除了要考虑论题的简单意义,还要考虑效用和环境因素。同样,数据编码与研究问题的关系必须考虑范围广泛的潜在因素。

当把 KO 模型用于组织信息的时候,KO 模型通过标引从受控词汇的选择流经它的运用,以便为检索创建一个逻辑上组织完备的数据库。当把 KO 模型用于组织研究数据的时候,此模型从待编码的变量或主题,流经这些数据的运用,创建为分析而组织完备的数据实体。

2.1 共同范围

创造一个受控词表的第一要素是确定它的范围(如它将涵盖什么),即要有一个共同范围(coextensiveness),然后决定怎样把此范围划分成类目。理想的是,让这些类目与所涵盖领域的各个方面同样广泛,以便促进相关检索[5~6]。例如,表示动物的词汇表可能遵循基于某种生理特征的传统的动物学分类体系,即某个骨头的有无、食物和繁殖方式(见图 2)。在这个分类顺序中,猫被分入猫科动物类,猫科动物被分入食肉动物类,食肉动物被分入哺乳动物类,哺乳动物被分入脊椎动物类。如果检索者的问题是有关食肉动物和猫科动物方面的,那么,这些类目的定义是有用的。然而,如果检索者想了解有关"宠物"的知识,这个分类顺序就是无效的。这种情况下,基于动物与人类有关的功能的组织就可能对它们有用(见图 3)。在这样的分类顺序中,动物可以分成家养动物和野生动物的类目。家养动物又可以分成宠物、家畜和使役动物等等。在"宠物"这个类中,包括精选的猫科动物、犬科动物、鸟、鱼,甚至连蚂蚁也都可能包括进去。这些动物可能遍布动物学分类的组织结构中。其中,可能设有与"宠物"这个概念范围相同的类目。

在对研究数据进行组织时,变量和主题代码的定义必须与研究问题或研究假设所表示或隐含的概念具有相同的范围。像"图书馆"这个概念,在一项研究中可能被定义为物理空间,例如,当地公共图书馆馆藏的用户需求分析。而在另一项研究中,可能与构建话语有关,例如,图书馆是一个向所有公民提供信息的民主机构。根据特定研究的目标不同,所选择的类目与其他研究有很大的不同,但与手边的研究问题具有相同的范围。控制词表或编码系统及其运用的其他两个因素是专指性和详尽性。

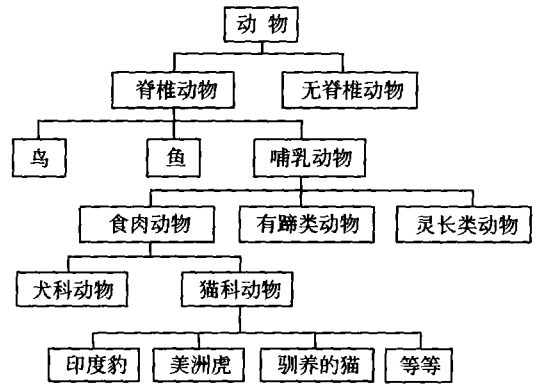

图 2　动物学分类(节选)

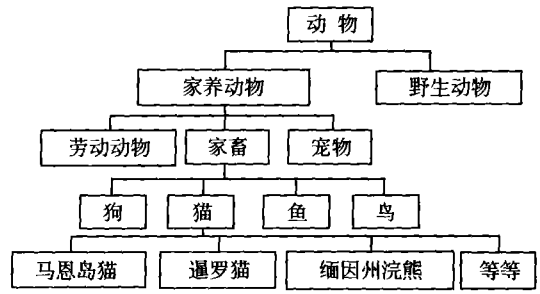

图 3　与人类相关的动物

2.2 专指性和查准率

专指性至少可以从两个方面加以考虑:既可以作为词汇的一个特征,又可以作为词表应用(标引过程)的一个因素。在第一种情况下,专指性是词表内的相关细节——所定义的等级结构的数量。如图 3 所示,只停留在区分家养动物和野生动物的分类系统的专指性很低。各个级别的专指性随着等级的加深而增加,最后一级能区分特定动物的特定品种(如暹罗猫),这样的专指性就相当高了。

当范围相同时,专指性水平应能满足这个组织系统的用户需要。就动物而言,一个兽医索引的专指性必须比一个社会学索引的专指性更高。当处理研究数据时,研究问题和(或)研究假设决定了对研究者指定的类目和编码进行组织所需要的专指性水平。一旦确定了专指性水平,那么,就必须运用能够发挥专指性的全部优点,即考虑专指性的第二种方法。应用的专指性是标准的标引实践,每个项目必须在可获得的专指性最高水平上进行编码。再来看图 3,这个类表将指定把暹罗猫标引成"暹罗猫"而不是"猫"、"宠猫"或"家养动物"。同样,一个划分不同类型"图书馆"的类表可能包括"学术图书馆"、"公共

图书馆"、"专门图书馆"等等，并且在标引过程中运用到最专指的水平。

查准率是检验某一系统检索相关信息有效程度的一个标准的测度方法。它是指在一次检索中所检出的相关信息的数量与不相关信息的比较。如果查准率高，则检到的所有信息是相关的且没有或几乎没有非相关信息。对于研究数据而言，这表明，为特定变量收集到的或根据特定属性编码的所有数据，实际上的确具备这个属性。这意味着，类目之间要相互排斥，并且数据已经被分配到这些类目中。为了实现这个最终目标，查准率通过专指性来提高。C. W. Cleverdon 明确了专指性和查准率之间的关系[7]。80 年代进行的一项"图书馆资源委员会联机公共检索目录"(CTR OPAC)研究支持了 Cleverdon 的结果[8~9]。

高专指性意味着在一个非常精确的水平上标引(或分类或编码)数据，即利用类目的细分，对所选词汇进行精心处理。相反，低专指性容易导致高查全率[10]。

例如，如果研究者对有关"猫"的数据进行编码，他们可能仅仅指定代码"猫"，或者如果需要高专指性水平的话，就指定为个别品种。然而，问题就来了，因为这个实践的前提假设是这些类目之间相互排斥。如果为特定品种分类或编码，则每只猫必须属于一个品种。杂交的或没有明确品种名的猫将在类目间消失，即使这些猫是数据集合中的绝大多数。获取查准率并不像乍看起来那么非常明确。

2.3　网罗性和查全率

网罗性也是编制和应用词表中的一个问题。网罗性被定义为表示的宽度，即被标引或被编码的元素数量。与专指性一样，网罗性也是词汇及其应用的一个特征。在词表或编码系统的层面上，网罗性要考虑所包括的不同方面或侧面。例如，图 2 和图 3 的结构可能由以下各个面结合而成：表示功能的面(如消化和繁殖)，表示环境的面(如水族馆、牛马房和房子)，表示与人类接触类型的面(如食物和交往)等等。再来看看有关图书馆类型的类目，网罗性将会出现有关分组中每个类目本质的问题。例如，应当把公共学术机构图书馆与私人学术机构图书馆划分开吗？应当把拥有独立的儿童设施的公共图书馆与那些没有儿童设施的图书馆划分开吗？在一个特定的研究项目中，必须对研究问题或研究假设的本质做出选择。

在实际应用中，网罗性与可标引事物的水平有关，也就是说，在把一个特定主题表示在类表中之前，载有信息的文献必须涵盖给定主题到什么程度[11]。如果一本书有一半是关于某一主题的，那么应当把这个主题表示出来吗？如果相关的只有四分之一，那又怎么样呢？如果仅仅有一章(或一章中的一节)相关，那又怎么样呢？如果该内容仅局限于一个段落，那又怎么样呢？必须根据与研究问题的描述水平相适应的原则做出选择。

一个相关问题如下：对特定文献应当表示出多少个概念？在这里，网罗性就牵涉到专指性。如果一个有关宠物的讨论包括猫和狗，在划分这些概念时好像既专指，又泛指。然而，如果讨论包括猫、狗、鱼、鸟、老鼠、兔子和无峰骆驼，它就不能包括有关每种动物的足够信息，以证明高专指性的表示。这种情况下，"宠物"可能是一种更合适的选择。再回到前面提到的图书馆研究(如公共图书馆藏书的用户需求分析)，"图书馆"这个概念可能用高专指性的藏书分析的类目来表示，其中包括"图书"、"录音带"、"软件"、"录像带"等等。研究者必须根据类表的网罗性和专指性做出决定，例如，如果一本儿童书除了包含故事之外还配有磁带，那么这篇文献既可分到"图书"的类目下，又可分到"录音带"的类目下吗？在第二个研究中(即图书馆作为一个民主机构的话语)，其类目可以不做出指向图书馆物理属性的参照，而可能用与图书馆概念化目标有关的类目来表示，其中包括"社区设施"、"信息自由"、"普遍获取"等等。

网罗性与查全率密切相关，查全率是一个从所有可获得的相关信息中检出了多少相关信息或数据的测度。最大查全率意味着，检索到了有关某一主题或变量的每个最新例子。然而，要想获得高查全率，就不可能只检索到相关信息而不包含非相关信息。这就是为什么查全率和查准率总是表现为一种彼此相反的关系，这个问题我们将在后面讨论。这个因素将会对后面讨论的分析构建数据和编码产生影响。提高查全率的一个途径是高网罗性的运用。与专指性和查准率之间的关系一样，Cleverdon 也定义了详尽性与查全率之间的关系[12]。这并且得到了 CLR OPAC 和其他研究者的证实[13]。Sparck Jones 把标引中的详尽性水平同检索式中所需要的详尽性水平反向地联系起来，并且得出了查准率和查全率的类似结果[14]。

如果网罗性高，就可以使用更多的编码，也就可

以检出和分析更多的数据。每个最新主题将被识别和编码。每当编码另一个元素时,这些数据就更可能被检索出来。从而,每次检索或分析的数据也会更多,可能包括更大数量的相关信息和非相关信息。

3 知识组织概念模型中的问题

相关性、查准率、查全率、专指性和网罗性的概念,好像能够产生出完美的类目和编码,我们称之为"数据彩虹",但是,标引员都知道,在应用这些来组织概念时,存在几个潜在的问题。在追求理想化的数据组织中,有两个问题显得特别复杂:(1)查准率和查全率之间的关系;(2)一致性问题。

3.1 查准率与查全率

第一个棘手问题是查准率和查全率之间的反变关系(在图1中用相互交叉表示)。通常把查准率和查全率之间的关系看成是反变的,至少是从Cranfield研究就开始了[15],并且D.R.Swanson通过重复工作证实了Cleverdon的数据[16]。M.Buckland和F.Gey又强调了这一关系[17]。但是,R.Fugmann做出的论证是令人信服的:这种反变关系是一种倾向,而不是一个定律[18]。

高网罗性必将导致较低的查准率,因为增加越来越多的编码,将会导致在检出相关数据的同时,也连带出越来越多的非相关数据。相反,高专指性必将导致较低的查全率。由于高专指性使用更狭窄的类目,它的类目比低专指性产生的数据更少。

从理论上讲,查准率和查全率都可能有一个理想的水平,但实际上很少能够这样。当制定类目和编码时,研究者必须决定对于数据分析过程哪个倾向最为重要。例如,在某些信息检索研究中,"图书馆"这个一般类目是表示一个地方的合适标志,在这个地方,人们可以找到能够解决其信息需求的信息,此外,这样的地方还有"朋友"、"学校"、"家人"、"电视"及其他信息源类目。然而在其他研究中,研究问题可能需要一个更高层次的专指性,因而需要人们找到信息的不同种类图书馆的多元编码(如"学术图书馆"、"公共图书馆"、"专门图书馆"、"学校图书馆")。如果研究要求这个更专指的途径,那么,研究者分析数据的工作就更为困难;定量类目必须包括所有类型的图书馆,或者,定性编码必须反映出数据收集过程中的每一种类型的图书馆。在任何一种情况下,为多元分析编码进行的检索比为单个词"图书馆"进行的检索耗费更多的时间。

3.2 一致性

标引员之间的一致性是知识组织模型的第二个棘手问题,这指的是要解决词和概念应用中的不一致问题。在创建或应用词表系统时,如果一篇文献的几个标引员使用不同级别的专指性或网罗性,其最终结果将使这篇文献成为检索者难以或者不可能检出来的文献,也就是说,查全率和查准率都低。在编制研究数据时,如果不同的分类员或不同的编码员(或者相同的分类员或编码员在不同时间)使用不同级别的专指性和网罗性,那么,这个分析就可能产生容易使人误解的结果。

例如,在一项有关图书馆利用的定量调查中,调查者不一致的调查提问可能会产生不一致的结果:一个调查者可能问:"你一个月至少到图书馆几次?"而另一个调查者却问:"你一个月至少到公共图书馆几次?"在这里,使用了"图书馆"的不同水平的专指性,这两个提问的答案相差很大,产生了让人误解的结果。

查全率和查准率都会受到影响,因为两者都取决于类目的准确性,不管是强调专指性,还是强调网罗性。这种情况下,使用标准的调查提问将会提高研究数据的一致性水平。在定性研究中,通常使用编码员之间或内部的可靠性来保证运用主题编码的一致性。

不一致性把噪音引入分类和编码过程,又可能产生非相关的结果。遗憾的是,获得一致性是相当困难的。就证明存在着大量不一致的研究而言,长期以来,有关标引的文献是很充分的,即使在经验丰富的使用熟悉系统的专业人员中也是一样的。作为一种定论,标引缺少一致性已经强调了几十年,并且记载在70年代的许多研究者的文章中。B.M.Preschel、L.E.Leonard K.Markey评述了早期的研究,并发现了类似的结果。最近,有些研究者,如A.Bertrand和J.Cellier继续测定改变一致性水平所涉及的变量[19],而像L.Y.Collantes,M.Iivonen,M.Iivonen和K.Kivima ki等研究者已经探讨一致性的更多分支,把Zipf分布与同较高的详尽性和专指性的更大不一致性联系起来,对这一方法有了更好的认识[22]。

最能提高一致性的三个元素是:帮助使用词表或编码系统的文字资料、低专指性和低网罗性。显然,解决一致性问题将产生其他缺陷。用低专指性和低网罗性来获得一致性将会导致低查全率和低查

准率。此外,进行权衡是必要的,并且必须纳入当前的研究问题。如果族性词"图书馆"能够满足研究目的,那么,在把它作为一个研究编码或类目时,就能够保证研究者取得更加一致的结果。但是,如果研究需要较高水平的专指性或网罗性的话,那么,为了获得尽可能接近的一致性和进行检查(如编码员之间的可靠性),就需要更加努力,以便提高该项目的精确性。在文本分析中,也必须做出同样的区别。当维持为避免不恰当广义化的一致性解释时,在研究问题的恰当专指性和网罗性水平上来解释个别语句。另外,在文本研究中,允许足够的自由来解释创造性发现,这一点很重要。

4　结论

知识组织是一个复杂的、系统的智力过程,而共同范围、网罗性、专指性、查全率、查准率、一致性、相关性是这个过程中的主要概念和要素,可以把知识组织中的主要概念及其相互关系结合成一个简单的知识组织概念模型,这个模型可以用于解释知识组织中经常遇到的一些问题。专指性和网罗性都是编制和应用词表中的重要问题,它们既可以作为词汇的一个特征,又可以作为词表应用(标引过程)的一个因素。查准率是通过专指性来提高的,查全率是通过网罗性来提高的。

在知识组织模型的应用中,有两个问题特别复杂和棘手:一个是查准率和查全率之间的关系;另一个是标引员自己和标引员之间的一致性问题。查准率和查全率之间通常表现为反变关系,而提高一致性的途径是降低专指性和网罗性,但势必又降低查准率和查全率,解决办法是进行权衡。

参考文献

1 Schamber, L., Eisenberg. M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 26, 755－776.

2 Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48, 810－832.

3 Hjorland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content, and relevance. *Journal of the American Society for Information Science*, 52, 774－778.

4 Albrechtsen, H. (1990). Subject analysis and indexing: From automated indexing to domain analysis. *The Indexer*, 18, 219－224.

5,11 Milstead, J. L. (1984). Subject access systems: Alternatives in design. Orlando, FL: Academic Press.

6,9, 10, 15 转引自 Given, L. M. & 01son, H. A. (2003). Knowledge organization in research: Aconceptual model for organizing data. *Library & Information Science Research*, 25, 157－176.

7,12 Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation*, 28, 195－201.

8 Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research*, 6, 155－177.

13 Boyce, B. R., & McLain, J. P. (1989). Entry point depth and online search using a controlled vocabulary. *Journal of the American Society for Information Science*, 40, 273－276.

14 Sparck Jones, K. (1973). Does indexing exhaustivity matter? *Journal of the American Society for Information Science*, 24, 313－316.

16 Swanson, D. R. (1965). The evidence underlying the Cranfield results. *Library Quarterly*, 35, 1－20.

17 Buckland, M., & Gey, F. (1994). The relationship between precision and recall. *Journal of the American Society for Information Science*, 45, 12－19.

18 Fugmann, R. (1994). Galileo and the inverse precision/recall relationship: medieval attitudes in modern information science. *Knowledge Organization*, 21, 153－154.

19 Bertrand, A., & Cellier, J. (1995). Psychological approach to indexing: Effects of the operator's expertise upon indexing behaviour. *Journal of Information Science*, 21, 459－472.

20 Collantes, L. Y. (1995). Degree of agreement in naming objects and concepts for information retrieval. *Journal of American Society for Information Science*, 46, 116－132.

21 Iivonen, M. (1990). Interindexer consistency and the indexing environment. *International Forum for Information and Documentation*, 15, 16－21.

22 Iivonen, M., & Kivimäki, K. (1998). Common entities and missing properties: Similarities and differences in the indexing of concepts. *Knowledge Organization*, 25, 90－102.

王知津　南开大学国际商学院图书馆学系教授。通讯地址:天津。邮编300071。

张国华　南开大学国际商学院教师。通讯地址同上。

(来稿时间:2004-01-12)