DOI :10.13530/j.cnki.jlis.1999.05.003

●王知津

知识空间:知识组织的概念基础

摘 要 文章从情报组织和知识组织的具体化出发,提出了"知识空间"这样一个知识组织的新的概念基础。探讨了可在多维空间中表示的知识领域的"概念向量"、基于对这些概念理解的个人"状态向量"、包括知识子空间的检索系统中可能存在情报项的"表示向量"、用户扩检或缩检子空间的"检索体积",以及这一新概念意义。参考文献 11。

关键词 知识空间 知识组织 概念向量 状态向量 表示向量 检索体积
分类号 G350

ABSTRACT Starting with the materialisation of information organisation and knowledge organisation, the author proposes "knowledge space", a new conceptual basis for knowledge organisation. Then, the atuthor studies "concept vectors" in knowledge field which can be represented in multidimensional spaces, personal "status vectors" based on these concepts, "representation vectors" of information items which may exist in retrieval systems including knowledge subspaces, "retrieval volume" of subspaces of users expanding or narrowing retrieval subspaces and implications of the new concept. 11 refs.

KEY WORDS Knowledge space. Knowledge organisation. Concept vector. Status vector. Representation vector. Retrieval volume.
CLASS NUMBER G350

1 引言

自远古以来,人们一直在对所积累的知识进行分类和组织。这些工作占用了人们的许多精力,一代又一代的图书馆学家、情报学家甚至一生都在致力于这项工作。

早在 1953 年,A.Rapaport 在一篇题目叫"什么是情报"的文章中指出,"生活主要依赖于有序过程,竭力避免混乱,而在无生命世界里总是表现为无序。但是,增加任何事物的有序性就意味着可以用较少的情报(较少的努力)来描述它。这一过程是否是知识以及科学本身的实质? ……当科学家把大量似乎无关数据构建成完整理论时,他们都投身于减少一部分世界的熵并使之用较少的努力理解的过程。"他接着说:"Korzybski 等人坚持结构是知识的唯一内容,……结构或者大量组织的测度是必需的。"[1]

20 年后,Bronowski 认为:"物理学的一个目标是精确描述物质世界。20 世纪物理学的一大贡献是证明这个目标是达不到的,……绝对知识是不存在的,……所有情报都是不完美的。我们必须谦虚地对待它。"他引用了 Max Born 的话,"我现在确信,理论物理学是真正的哲学",其含义是:"世界不是物体的固定不变的排列,因为它不能完全脱离我们对它的感知。它在我们的眼皮底下转变,与我们相互作用,它所产生的知识也由我们来解释。不需要判断行为的情报交流方法是不存在的。"他把"不确定性原则"重新命名为"容限原则",并希望永远牢记,所有知识都是有限的。他最后说:"人类生存的所有知识和所有情报只能在容限原则的作用下进行交流。"[2]

C.Cherry 设想,科学是"在类似拼板玩具的碎片中创造的,把碎片加进这个玩具,碎片与其相邻部分相关,但同完整图形没有明

显关系。"[3]他提出,从根本上重新安排科学、技术以及一切知识领域的结构。

2 现有的分类与标引方法

现在的问题是,文献工作者、图书馆员和情报学家组织知识和情报的各种方法是否跟得上知识的增长以及我们对其构成的改变?

60 年代,B·C·Vickery 按控制程序的升序列举了 7 种分类和标引方法[4],并对标识方法、展示形式或专门术语进行了专门研究。这 7 种方法如下:

① 从题名或正文中选词,但排除常用词;

② 从正文中选词,排除常用词,考虑词形变化;

③ 从正文中选词,排除常用词,考虑词形变化,也考虑族性关系;

④ 从正文中选词,考虑标引词之间的句法关系;

⑤ 上述方法中的任何一种,再补充正文中未使用的词;

⑥ 从固定的规范表或分类表赋予索引款目;

⑦ 从代表主题某些观点和方面的规范表或分类表赋予索引款目。

3 分类和标引系统的不充分性

上面所列的第③~⑦中的一些表都是对情报学的重要贡献。但是,这些表的编制和维护也受到一些批评,因为几乎没有提供:

① 用较少的情报(较少的努力)描述事物的有序结构(这是 Rapoport 提出的)。往往简单地增加规范表中叙词的数量,而不是设计表的结构。当把结构设计出来后,这些表大概就有了固定的局限。

② 适应未来发展方向的知识领域结构的彻底重组机制,这个方向现在还不可预见。(借用 Cherry 的话)

③ 与这些结构相互作用的方法,这个方法考虑到对实际叙词的不完全认识以及判断行为(回答 Bronowski 的要求)。

在情报项的表达和情报检索的描述中,分类和标引方法几乎完全依赖于关键词和短语的使用。当然,在标引和检索过程中,必须用词进行交流,但似乎缺少一种概念框架,该框架应能提供一个恰当的有序结构,有助于结构和用户之间以及用户和用户之间的情报交流。

在考察词本身并找出其基础框架方面,已经进行了许多尝试。首先想到的是图书馆分类法,但它们固有的局限性使其完全不适合我们面前的任务。

S·R·Ranganathan 曾说过:"图书馆分类法等价于单维中的多维连续统一体的表示法。图书馆分类法容易受到机械的、维数的和经费的限制,它有不可消除的局限性,这些限制通过特定主题之间的有用顺序对它施加影响。"[5]

Lauren Doyle 关于语义图的开创性论文以及欧洲原子能核文献系统的箭头图(1964 年第一版,后来的版本中叫 INIS 术语图)曾经尝试提出的标引词汇结构的例子[6]。然而,George Miller 认为这些努力是组织情报资源所需的"显著改进",它们并未真正提供"空间标识"[7]。按照 Miller 的观点,我们必须提供"对存贮情报的空间进行组织,这种组织与情报本身的结构更为兼容",并且"帮助用户探索巨大的空间形象"。

Miller 对情报检索中人机交互的理论化倾向进行了批评:"用户有一个概念系统,情报存贮也有一个概念系统。当用户发现他必须用自己的系统填补缺口时,他就形成了关于它的问题。由于这个内容正在丢失,所以,该问题只能指出这个缺口在用户概念系统中的一般位置,为了辨认'指出的位置',被询问的系统应当有类似的概念组织。作为一种交谈,交互的性能主要由概念标引的充分性决

定，因此，无论怎样描述，可以迅速而准确地辨认和填补用户缺口。可以设想，用户不必学习分类系统(如同他们现在必须学会图书馆分类系统那样)，因为可以让计算机系统把用户的问题(不管怎样措辞)转换成它自己的标引系统。"Miller 还说："我怀疑，没有空间维数的检索系统是否能与描述用户思考的概念充分匹配。"[8]

Mitroff 和 Turoff[9]以及他们之前的 C·W·Churchman[10]都强调，我们对现实的基本印象的差别会引起不一致和不同评价。用 Vickery 的 7 种方法中的任一种设计的查询系统，似乎不适应查询的基本原理方面的差别。

需要新方法和新结构的领域存在着严重缺口，因为要组织和表达词或词组，我们正是要用这些词和词组来标引和分类文献以及用户提问。

4 知识结构化的多维空间模型

科学家发现，使物体实际状态具体化(如按多维向量空间的原子)是有用的，可用数字公式或其它标识法(如标引词)对这种状态进行描述。但是，构造多维空间将会大大提高使这些状态及其相互关系具体化的能力。

知识多维空间具有如下特征：

(1) 把概念有用地表达成类似多维空间中向量的对象，概念的意图是把所用的词同标识概念分开。

向量具有量值和方向并遵循准确的数学规划。这里所说的对象包含不确定性，不象向量那样准确。但是，在帮助情报学家使概念之间的关系以及组成情报检索的各种过程具体化方面，这种构造的作用很大，而且非常有用。

如果总体上无关，从概念是互相垂直的意义上说，可以把概念看成是有方向的；如果关系密切，就具有空间方向大致相同的特点。关系密切的概念的方向不会精确地相同。它们之间的角度越小，概念的关系就越紧密。知识空间中的相关领域具有共同的确定的概念向量，但各个领域又是概念向量的不同组合。

若知识空间的原点表示知识的零状态，则可把概念延伸具体化为向量长度的扩展。概念距离的测量不能在欧几里得几何空间中的笛卡尔数轴距离的意义上严格定义。唯一的解释是，一个点沿着这个而不是另一个概念向量进一步伸展，但我们不能说进一步多少。准确定义概念的量值似乎是不可能的。概念向量横跨所有的知识空间。由于许多概念之间有关系，所以，概念向量的数量比知识空间维数(垂直方向)的数量多。必须要作的是，确定一个基本概念向量集合，每个向量与所有其它向量都是相互垂直的，以便通过最少的概念向量横跨所有的知识空间。当新的基本概念形成时，知识空间的维数也会发生变化。

(2) 一个人的知识状态可以在多维空间用状态向量来表示，状态向量带有分量，分量是关于被那个人"理解"的那些基本概念。

可把学会新概念看成增加知识空间的状态向量(或使其循环)，以便该人发展他刚刚学会的概念的分向量。这相当于把原子状态的变化描述成它在多维空间的状态向量方向上的变化。也可把状态向量看成对用户 SDI 需求档的表达。

(3) 在这个多维知识空间中，可以用带相应概念向量的分量来表示图书、期刊论文、胶片及其它情报项。

一旦建立起相互垂直的基本概念向量集合，对情报项的分类行为就完全由分向量组成。该分向量在基本概念向量集合上确定其表示向量。这里，Zadeh 的模糊集合可以发挥很大的作用，如果愿意有多个这种分向量的原始二进制标识的话[11]。

5 多维知识空间向量举例

对于求解字顺规范词表表达概念混乱问

题,对于求解树型结构表达概念和知识领域之间相互关系的局限性,对于求解现有分类和标引方法的固有不灵活性,多维知识空间的概念结构对我们应当是有帮助的。为了更好地理解多维知识空间的实质,我们下面列举一些例子,其中大部分是二维或三维的。

(1) 多维空间向量的概念

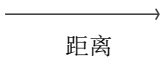

图1 基本概念向量

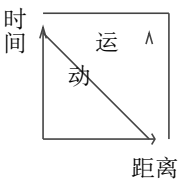

图3 相关概念向量:运动

基本概念向量的"长度"可在不同区域表

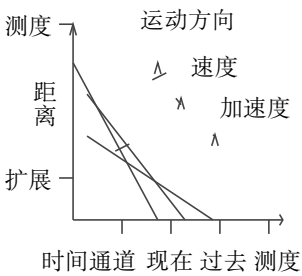

图5 概念向量的长度

(2) 表示某人理解概念向量的状态向量

可以认为,原始人的知识空间具有这种概念向量(如空间和时间),可以把 Einstein 和 Newton 的状态向量看成具有这两个概念向量的分量。但是,容易看出,概念向量到原点的距离,Newton 和 Einstein 的比原始人的远。一个人的知识状态可用带有他所理解的概念向量分量的状态向量来表示。意味深

图 1 表示一维,展示了一个基本概念向量:距离(或空间,或长度)。图 2 表示二维,展示两个基本概念向量:空间和时间(垂直向量)。图 3 表示带相关概念向量的二维:运动(含时间和距离基本概念向量的分量)。图 4 表示三维,展示三个基本概念:距离、时间、质量。

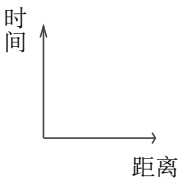

图2 基本概念向量(垂直)

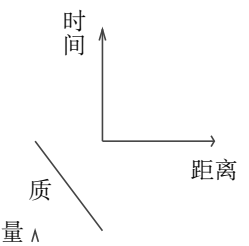

图4 三个基本概念向量

示(或"测度")和定位,如图 5 所示。

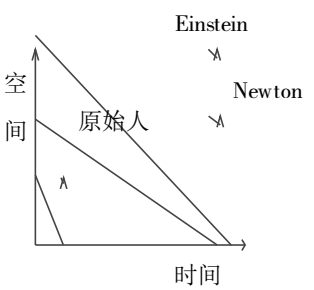

图6 状态向量和概念向量

长的是,Einstein 对时间的理解比 Newton 要深,事实上,可以这样说,他进一步扩展或发展了时间的概念,因而增加了概念向量的长度。但是,很难说他的概念是否远两倍,或者比概念向量远 3 倍(图 6)。我们所能做的就是确定 Einstein 离时间概念向量比 Newton 远的点(图 6)。

为了适应有关思想的聚类,大多数看法

不得不定位在知识空间，而偏离最新看法。例如，时间延伸是个与 Einsteind 时间概念有关的新概念，应当定位在表示时间的概念向量最新端的附近区域。

(3) 记录知识和知识领域中情报资源的表示向量

可以把多维知识空间的子空间看成包含某一领域已知(以前知道)记录知识的区域或体积，对于从子空间有效地检索情报来说，这是有用的框架(图 7)。

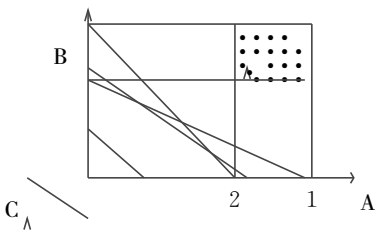

图7　多维空间的子空间

在检索系统中，广义词或狭义词有利于扩检或缩检，这与假设的多维知识空间中的检索体积的大小有关。显然，检索体积越大，检出的情报项数量越大。在新构成的属性一定的情况下，可以看出，检索者能控制的维数的子空间是容易确定的，形成一个由参照轴上的坐标(或基本概念向量)所规定的检索体积，这个坐标标明他的理解程度。如果用户要缩小检索体积，或者在三维空间中移动，那么，通过定位一个新的对象集合，该系统就应当能够适应这个新安排，通过替换他的原始检索需求档中的一个或多个概念向量，对该集合进行定位。由于相关概念向量的方向可能类似，所以，这个过程的实现并不困难。

如果不为检索系统中的情报项调整表示向量的原始位置的话，这些轻微的移动和转动可能导致要确定的另一个检索空间，因为基本概念向量在知识空间中是被"独立导出"的。来自规范表(叙词表)的叙词将被表示向量取代。情报项的标引记录不再只同标引时

6　检索体积

可以这样看待检索提问：首先选择一个合适的子空间；然后限定在这个子空间中的一个检索体积(Search Volume)，从而，进一步缩小这个检索的范围。该过程可以恢复所有情报项，它们带有这个检索体积的表示向量，如图 8 所示。

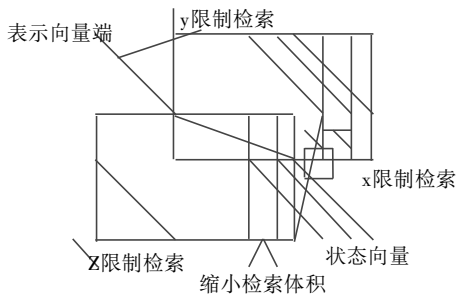

图8　检索体积

所用的词连接。词汇用法的转移可以同较早的标引记录连接，因为表示向量可能位于知识空间中近似相同的子空间中。

7　讨论与总结

针对分类和标引，可作如下补充："在一个带基本概念向量的分量的多维知识空间，赋予表示向量。"某一个领域知识的这种框架或结构具有灵活性，即随着该领域的发展而重新安排、可能的转动以及可用较少的情报(较少的努力)来描述的扩展。可以把这一项加在 Vickery 的 7 项之后，算是第 8 项。这种方法可能成为分类和标引新方法的概念基础。

这一方法也允许图示正在查找情报的人的知识状态的状态向量。这些状态向量带有基本概念向量的分量，它表示一个人对这些概念的"理解程度"。该方法允许对被检的检索系统中所用的实际表示向量的认识不完全。状态向量和表示向量可位于同一或邻近

区域，因为它们都带有该领域概念向量的分量，而且，指定的"检索体积"标志待检的子空间。当交互发生时，用户可检查出所进行检索的最优限制。

上述表达多维空间知识的概念基础具有如下意义。

① 描述概念之间的关系

描述知识空间中概念之间关系的结构更为灵活和有用。它只需用较少的情报(用户部分)来描述和构造他的检索体积。开始时，可用三个基本向量确定一个维数可控的子空间。

当重新确定某一领域时，对该知识领域结构的重新安排，可在多维知识空间中进行，把相关的表示向量通过一次转移与一个新的概念向量集合联系起来。

由于这个多维知识空间的概念基础允许某个人的状态向量、某一领域中情报项的表示向量(正如某一检索系统表示的那样)以及作为代表某一知识领域的知识空间中"对象"的实际概念向量之间的不完全匹配，用户和检索系统之间的交互将会增加。

知识空间的这个概念基础把组配标引的想法延伸到它的界外，并有望进入其它维度，留下一维和二维，扩展到 n 维的检索体积。

这个概念基础把广义和狭义延伸到等级树型结构以外，进入伞形面，在这个面上可以确定情报点，区分密集聚类的有关概念，参照轴引向该面的其它部分或"内容"。也可确定与其它知识领域的界面，并图示这些关系的结构。

② 较好地理解情报传递

知识空间的使用有助于我们对情报传递的理解。在有关交流和传递过程的大多数图解中，往往忽略了接受者状态的变化。在本文提出的构成中，可以随该系统的每次使用，对该人的知识状态进行决策和记录。

在我们所知道的大多数检索和查询系统中，无法真正区分概念向量、状态向量和表示向量，这可能是改进检索系统设计的阻碍因素。情报专家面临的工作是确定基本概念向量的最佳集合，以便描述知识领域的范式。这一方向的起点是一个总体上不相关概念的表，它随着基本集合的一致而减少，该集合代表某一知识领域中的范式。这将扩大或取代过去发现基本类和子类、基本部类以及基本划分的工作。这应当有助于较好地表达情报项和用户检索。

参考文献

1 Rapaport, A. What is information. ETC. 1953 Summer;10:259

2 Bronowski, J. The principle of tolerance. Atlantic. 1973 December;232(6):60～66

3 Cherry, C. The spreading word of science. Times Literary Supplement, 1974 March 22:301～302

4 转引自 Bourne, C.P. Methods of Information Handling. Wiley;1963,13～20

5 Ranganathan, S.R. Philosophy of Library Classification. Munksgaard, 1951,94～95

6 Doyle, Lauren B. Indexing and Abstracting by Association. American Documentation. 1972, 23:378～390

7, 8 Miller, Georgy A. Philosophy and Information. American Documentation. 1968, 19:286～289

9 Mitroff, Ian I.; Turoff, Murray. The ways behind the Hows. IEEE Spectrum, 1973, March, 62～70

10 Churchman, C.W. The Design of Inquiring Systems. Basic Books, 1971

11 Zadeh, L.A. Fuzzy Sets. Information and Control. 1965,8:338～353

王知津 南开大学信息资源管理系。地址：天津。邮编 300071。

(来稿时间：1999-04-15。编发者：李万健)