中国图书馆学报(双月刊) 1997 年第 6 期
ZHONGGUO TUSHUGUANXUEBAO

●张 进

情报检索系统: 从布尔逻辑到向量空间

ABSTRACT By analysing basic structures, applicability, functions and potentials of Boolean logic based information retrieval system and vector space based information retrieval system, the author thinks that the latter is an advanced stage and will be dominant in the future development of computer information retrieval system. 3 refs

KEY WORDS Information retrieval system. Vector space based retrieval system. Boolean logic based retrieval system.

CLASS NUMBER G254.4

在计算机情报检索界, 谈及计算机情报检索系统, 人们往往会想到以布尔逻辑为基础的检索系统, 这与以布尔逻辑为基础的情报检索系统普及性好、实用性强、商业化程度高是密切相关的。例如, 闻名遐迩的DIALOG 联机系统、ESA 系统、STN 系统以及国内有影响的联机系统均是属布尔逻辑类型的。

这里讨论的以布尔逻辑为基础的情报检索系统包括我们所熟悉的脱机情报检索系统和联机情报检索系统。这两种系统均以布尔逻辑作为提问逻辑构成的理论基础。由于联机情报检索系统在处理的灵活性、检索速度以及与用户交互性上比脱机系统占优势, 因此成为目前情报检索系统中的主流, 可以说它已处在发展的高峰期。

然而, 以布尔逻辑为基础的联机情报检索系统的地位正在发生动摇, 它的地位正面临着来自以向量空间为基础的情报检索系统的严峻挑战。在美国, 以反映情报检索界新技术、新思想、新理论而闻名于世的TREC 实验计划中, 绝大多数的情报检索系统是以向量空间为基础的, 而不是以布尔逻辑为基础。由于TREC 计划所包括的系统往往代表着计算机情报检索未来发展的趋势, 左右情报检索未来的走向, 因此, 情报检索界有人戏称以布尔逻辑为基础的情报检索系统已是强弩之末, 它的强盛时期即将结束。对计算机检索界未来发展显现出来的这一重大现象, 人们不禁要问: 究竟是什么内在的原因导致以布尔逻辑为基础的情报检索系统的衰败以及以向量空间为基础情报检索系统的兴起呢?

1 以布尔逻辑为基础的情报检索系统分析

布尔逻辑为基础的情报检索系统之所以有其辉煌的兴盛发展过程, 这是有一定的历史背景的。首先, 以布尔逻辑为理论基础的提问逻辑, 在表达用户情报需求时有其独到之处: 简洁、精确、实用。更为重要的是计算机硬件发展水平的限制, 在七八十年代, 计算机硬件的处理速度以及计算机的存贮容量难以胜任以向量空间为基础的计算机情报检索系统对它们的要求, 联机系统尽管比脱机在速度和存贮空间上要求高, 但它们的总体要求在当时的环境下是可以被满足的, 这就为它们的发展提供

了物质条件。另一方面,从检索算法的复杂性上看,基于布尔逻辑的情报系统相对基于向量空间的情报检索系统要小得多。

总之,以布尔逻辑为基础的情报检索系统的优点是数据结构相对简单、存贮空间以及处理速度要求相对低、检索算法相对容易、表达情报需求方式实用。但它们也存在着许多不足。

(1) 各文献信息之间没有存在相互联系的复杂机制。脱机系统的顺排文档的组织方法,把一个文献信息的所有内容集中管理成一个记录,各文献信息之间是相互独立的、相互隔离的。联机检索系统的倒排文档虽然将各相同的标引词集中管理,起到了把相同词为基础的文献联系起来的作用,但是,在倒排文档中的每个入口词之间缺乏联系。因此,我们说在联机情报检索系统中,文献信息之间的联系是点式的、局部性的、各点之间缺乏多层次的联系。因而说这种联系的程度是较低的。

(2) 难以向用户提供判别文献与文献以及文献与提问之间精确的相似性测度方法。相似性测度的价值不仅仅在于能应用于用户最终需求的判别,同时它也是进行其他复杂运算的基础,例如聚类分析、自动分类等。

(3) 布尔逻辑为基础的提问逻辑在检索算法的实现过程中,对待处理文献的结果要么是命中,要么是不命中,很难确定位于两者之间的一个"灰色"地带,位于这个"灰色"地带的文献可能只在某种程度上满足用户情报提问。由于缺乏对"灰色"地带文献的描述以及处理机制,导致在命中结果的输出上难以有效地根据命中文献与提问的相关程度向用户提供一个合理的依相关程度(满意程度)排序的结果表。当然,基于布尔逻辑的检索系统在进行改进以后,也可以同样向用户提供这样一种有利于用户做相关的判断的排序表,例如加权法,但一般来讲,由于布尔逻辑为基础的检索系统不是以相似性决定文献的取舍,所以采用的排序方法很难全面、深刻地反映它们实际的相关程度。

(4) 基于布尔逻辑的情报检索系统为用户提供的检索手段较为单一。用户的情报需求往往是多样化的,不同类型的情报需求往往需要采用不同的检索手段,这样才可取得满意的效果。基于布尔逻辑的检索系统,除了全文检索系统所采用的相邻度检索技术以外,关键词以及关键词间的逻辑组配是它所开出的主要"处方"。

(5) 在描述用户情报总体需求时,缺乏一种全方位的、全视角的描述方法。我们知道,能否全面精确地描述用户提问是情报检索系统能否最终满足用户提问、提供满意结果的第一步。无可争辩,以布尔逻辑为理论基础的提问逻辑在描述用户需求方面是一种有效的手段,但是这种方法很难合理有效地反映多个提问式之间对提问产生的合效应。并且这种合效应也不能简单地将有关的提问式进行逻辑加或者逻辑乘而获得。而这种连续地、全方位地反映用户需求的方法对于从深层次反映用户需求是至关重要的。例如,一个用户多次检索所使用的检索式之间,往往存在着一定内在联系,因为用户从事的研究领域、学术背景等因素往往与他的提问存在着某种联系,这些因素或多或少地会影响用户提问的检索结果。所以,一个好的检索系统应该能从多方面反映用户的真正需要。

(6) 在脱机检索系统的顺排文档以及联机检索系统的倒排文档之上,难以开发出新的检索手段。以布尔逻辑为基础的检索系统已经过了几十年的发展和完善,可挖掘的潜力已基本挖掘,同时,由于在文献信息表达方式上存在着的固有缺陷,也阻止了它在开发新的检索手段以及综合利用上的深入发展。

2 基于向量空间检索系统的分析

在详细讨论基于向量空间的检索系统以

前, 有必要对这一概念做简单的介绍。我们可以首先把它定义为一个矩阵。这个矩阵的行分别对应于不同的关键词(标引词); 这个矩阵的列分别对应于不同的文献。每一个文献如果它的内容与某个行所对应关键词存在着某种语义上的联系, 则在相应行列上对应单元里赋上与语义相关程度一致的值。(当然, 单元的赋值方法依据不同的算法会有所不同, 但基本思路是一致的。)有人又称这个矩阵为文献—关键词矩阵。

2.1　推动基于向量空间情报检索系统兴起的因素

(1) 计算机硬件支撑环境的改善。计算机处理速度迅速提高, 目前, 100 兆/秒的计算机已相当普及。计算机内外存容量也以令人吃惊的速度增长, 新的超大容量存贮介质光盘已走向成熟, 这一切为基于向量空间的检索系统推向实用提供了良好的外部环境。

(2) 计算机软件技术的日趋成熟。例如, 数据压缩技术为大型稀疏矩阵的合理存贮提供了软件上的保障。前面讨论的文献—关键词矩阵, 是基于向量空间检索系统中文献的计算机内部表达形式, 它的特点是行距和列距均很小, 同时, 它也是一个稀疏矩阵, 即在矩阵中一行中绝大多数内部元素是零元素。对于这样一个大型的稀疏矩阵, 不使用数据压缩技术, 是难以使系统走向实用的。

(3) 满足用户需求多样化以及检索手段多样化的需求。传统的基于布尔逻辑的检索系统在满足日益增长的用户需求多样化以及检索手段多样化的需求时, 多少显得有些力不从心, 因而人们自然要寻求新的系统、新的理论以取代现有系统以及理论。基于向量空间的检索系统在智能化检索系统理论及技术尚不成熟的今天, 当然被人们作为挑选的替代目标。

2.2　基于向量空间的情报检索系统的特点

(1) 文献场的建立奠定了各种操作的基础。前面谈到的文献—关键词矩阵, 是一个便于读者理解的计算机内部存贮模型。为了加深对基于向量空间检索系统本质的理解, 有必要换一个理解角度来修正这个矩阵: 将这个矩阵中的每一行对应的关键词转换为一个新的向量空间体系中的一个新维数, 这个新建立的向量空间的维数等于文献—关键词矩阵的行数。显然, 在这个高维空间中任何一篇文献均可以根据它被标引的关键词具体情况找到一个点, 一个唯一位置点。这样, 我们就建立了一个文献场, 使任何一个提问、任何一个文献均可以在这个高维向量空间中找到它们的位置并进行相应操作。

(2) 高维向量空间中的信息点反映了组成该节点关键词之间的合效应。与文献相关的任何一个关键词以及对应权值均会影响该文献在空间中的位置。文献在向量空间中的位置是各种相关因素的合效应, 这充分体现了各因素相互影响、相关联系的特点, 摆脱了各关键词(标引词)之间相互独立、互不影响的不利因素。

(3) 高维向量空间中节点的位置及方向有其深刻的语义含义, 它们是确定各不同节点相似性测度的基础。在文献场的环境条件下, 两个节点相对于原点的方向是有其深刻的情报检索含义的: 如果它们方向相同, 说明它们在标引词的使用上是相同的, 有时赋给的权值可能不同, 却也能满足一定的分配比例, 反映在内容上说明节点对应的文献极为相关; 如果两个节点在场中的距离很近, 也同样可以说明它们无论是在使用的标引词上或是对标引词所赋的值上很接近, 两点在内容上很相关。因此, 我们可以以距离和方向为两个基础要素, 建立某种相似性测度的判别机制。这样一种相似性测度的判别机制不仅仅可以确定文献与文献或是文献与提问是否相关, 同时也可以精确地确定它们之间的相关程度。这些量化的相关值可以体现某些介于相关与不相关的"灰色区域", 也为检索结果按相关程度排序铺平了道路。

(4) 在高维向量空间中实现了用户检索

手段的多样化。在向量空间环境下,用户可以根据自己的需求特点选择一组可供使用的检索手段。例如,以用角度方向为基础的 Cosine 检索模型;以距离为判别基础的 Eulidean 检索模型;其他还有 Conjunction 检索模型、Disjunction 检索模型、Ellipse 检索模型、Cassini 检索模型等。每一种检索模型在满足不同情报需求方面均有自己区别于其他检索模型的特点。

(5) 向量空间为基础的检索系统具有很大的开拓空间有待认识和利用。利用已建立的向量空间可以对位于空间中的文献群进行聚类分析、分布性能分析、关键词的分辨值分析等。利用这个向量空间可以开发出新的检索模型。更为重要的是还可以利用这个向量空间,建立全新的情报检索可视化工具,向用户提供一个更为方便、更为虚拟、更为有效的情报检索环境。

(6) 其他方面的优点还包括可以通过调节标引词对应权值的大小来反映该标引词与被标引文献的相关程度。权值的来源既可以通过对文献全文自动统计分析后自动确定,也可以由标引人员人为地根据主题分析结果确定。在用户提问表达方面,由于没有必要将提问词按逻辑规则构成结构化的提问式,因此方便了用户情报需求的表达,利用用户用自然语句的形式向系统陈述提问。更为重要的是在向量空间的环境下,十分有利于分析和综合利用相关提问的合效应,这为多层次、多视角、全方位地反映用户提问真正需求奠定了可靠的基础。

上面我们对基于向量空间的情报检索系统的结构、优势以及潜在的价值做了较为详细的分析与讨论,这决非意味着此类型系统是完美无缺的,恰恰相反,它在许多方面与基于布尔逻辑的检索系统的优缺点是互补的。例如,它占用存贮空间大、要求机器运算速度高、检索的复杂性大,在提问式的表达上对逻辑非的处理以及词间层次关系表达上远不如基于布尔逻辑的检索系统。同时,尽管基于向量空间的检索系统在综合、全面反映文献内容上提高了一大步,但这种文献内容的表达方式忽视了关键词与关键词之间的语法信息和深层的语义信息。也许在未来,当人们在标引文献时,要揭示文献内部更深层内容特征时,基于向量的检索系统也会显得力不从心,届时,人们将会探索更好的理论及系统。

总之,通过对两大类型系统进行全方位的分析,不仅发现它们之间有很大差异,同时也存在着相互联系。仔细研究前面定义过的基于向量空间的检索系统的文献—关键词矩阵存贮结构,我们会发现一个有趣的现象:如果我们仅考虑这个矩阵的行方面的因素(每一行对应一个标引词),它们实际上就是联机检索系统中的一个倒排文档结构,每一行是倒排文档的一个入口词;如果我们仅考虑这个矩阵的列方面的因素(每列对应是一个文献),它们实际上就是脱机检索系统中的顺排文档结构,准确地讲,还不能说它与顺排文档结构完全一致,它只不过反映了顺排文档最基本的特征,即按一个文献内容组织一个基本逻辑单元。在检索功能上,基于向量空间检索系统中的 Conjunction 以及 Disjunction 检索模型实际上就是布尔逻辑中的逻辑乘和逻辑加检索,从这个角度上讲,它已可以实现布尔逻辑的主要检索功能。

任何事物的发展都有一个从低级到高级的过程,情报检索系统的发展也同样如此。基于向量空间的检索系统在系统结构、功能以及发展潜力上均优于基于布尔逻辑的检索系统,可以说它是基于布尔逻辑检索系统的高级形式,在未来的计算机情报检索系统发展过程中它将占据主导地位。这一点应引起同行们足够的注意和重视。

参考文献

1 张进 计算机信息检索软件设计原理 武汉:武汉大学出版社,1994 (下转第 42 页)

动适应。

分类语言是先组式检索语言。传统观点认为它的编制、设类只能被动地适应科学发展,即使是刚刚出版的分类法,在某些方面也会不适应或落后于某些发展迅速的学科或新兴学科。我们认为这样认识有失偏颇。

应再次明确的是文献分类法是用以进行文献归类整理,而不是其他分类用途;判断其成功与否的标准只能通过对文献归类的实践检验。

分类法的类目设置、类名名称等都具有相对的稳定性。它是根据较成熟的学科(主要依据是否有学术专著、相当数量的专家、相应的研究机构与学术团体等标准)情况设置、命名,而不能将昙花一现的、虚假的学科包容进来。

新学科有一定的渊源,分类法的上位类具有一定的概括性,规范、准确的类名在一定时期暂时容纳最初出现的新学科文献。

类目设置可以预测。任何事物都有发生、发展、成熟、衰亡过程,学科的发展也不例外。只要我们对学科文献的发展趋势进行充分、全面的预测,对类目的设置和变动就会有更多的战略眼光。例如通过国际大型检索工具报导的各类文献统计分析,通过报—刊—书这种文献逐渐成熟顺序的时差[12]等了解,使设类由盲从转变为有信心的前控。

取等级列举式与分面组配式分类法各自所长,优势互补,使分类法有较强的灵活性、适应性。

采用计算机编制、修订分类法,缩短分类法从修订到应用的时差。

总而言之,要对学科、文献、分类法的发展历史、现状、发展趋势不断调查、研究、预测,使它们有机结合,达到类目控制的目的。

参考文献

1 文榕生,张玉麟. 我国分类法的发展趋势. 图书馆,1996,(1)
2,3,5 文榕生. 论分类语言的标准化与类目控制. 图书馆界,1996,(3)
4 文榕生. 回溯中文图书源数据库建设. 图书馆界,1995,(4)
6 文榕生,冀丽芳. 复分新识. 晋图学刊,1996,(3)
7,8 王惠敏. 浅谈《科图法》(第三版)的成功与不足. 津图学刊,1997,(1)
9 陈晓华. "卡拉"柯处"OK"及其他:图书分类漫谈题外话. 图书馆论丛,1996,(4)
10 陈能华. 论分类法与主题法的本质分歧. 江苏图书馆学报,1996,(1)
11 朱丽. 我国分类检索语言计算机化的回顾与前瞻. 图书馆,1997,(1)
12 胡明,谢忠. .《中图法》出版物系统的研究. 图书馆界,1996,(3)

文榕生 中国科学院文献情报中心工作。通讯地址:北京海淀区科学院南路 8 号,邮编 100080。

(来稿时间:1997. 3. 31。编发者:赵薇)

(上接第 78 页)

2 Charles T. Meaolow. Text Information Retrieval System. London: Academic Press, 1992
3 Ian H. Witten, Managing Gigabytes, Van Nostrand Reinhold. New York, 1994

张 进 武大图书情报学院教授。主要从事计算机情报检索系统的教学与研究。论文曾在国内外刊物上发表。通讯地址:武汉市。邮码 430072

(来稿时间:1997. 3. 31。编发者:翟凤岐)