中国图书馆学报(双月刊)1999 年第 3 期
ZHONGGUO TUSHUGUANXUEBAO

●寇钧锋

论情报检索语言的自然语言化发展趋势

ABSTRACT As retrieval languages, both human-controlling languages and natural languages are complementary. At present, information retrieval languages are developing towards natural languages. In the future, there will be a new kind of languages with suitable control over natural languages. 5 refs.

KEY WORDS Information retrieval languages. Controlling languages. Natural languages.

CLASS NUMBER G254

一个理想的情报检索系统,人们希望它应该是一种"交互式的应答机",即人们提出什么样的问题,它就能做出相应的回答。它不但能接受任意的检索语词,还能理解检索者的检索意图。但是,就现实中的情报检索语言来说,无论是传统的分类语言,还是后来的主题语言,都属于受控语言。而作为情报检索语言的受控语言又存在着本身无法解决的缺陷。虽然说近年来在图书情报界兴起的自然语言也存在着自身的不足,但图情业内的有识之士已清楚地认识到,通过对自然语言进行适当的处理,却能比较好地解决受控语言存在的缺憾。这种相互补充的优势决定了未来的情报检索是一种自然语言的综合性情报检索语言系统。

1 受控语言和自然语言的相应概念对比

所谓受控语言,是指人们根据检索的需要,依据一定的规则对自然语言进行事先规范而形成的人工语言。受控语言通常指情报检索中的分类语言和主题语言。而利用分类表中的分类号和主题词表中的主题词或叙词表中的叙词作为情报检索的入口和控制检索的格式,称为受控语言检索。

我们通常所说的自然语言指人们日常说话、写文章和思想交流所用的各种语言。而情报检索中的自然语言是指用文献作者或文摘、提要的作者原来使用的语言,其中包括关键词、自由词和出现在文献题名、摘要或正文中的语词。目前,一种以相关排序和智能文本处理为特征的"自然语言处理"系统开始流行。自然语言检索,从技术上讲,就是将自然语言处理系统技术应用于情报信息检索系统的情报信息的组织、标引和输出。而从我们广大情报用户讲,就是把自我语言作为情报提问输入和对语接口的检索方式。

通过二者的概念对比可以看出,受控语言及其检索是经过事先规范化的人工处理而形成的;自然语言及其检索是直接将情报提问用词用于情报检索的。

2 自然语言检索的产生、发展和现状

正象规范语言当时作为情报检索语言时

一样, 规范语言是为克服自然语言的不足而产生的, 但它的出现并没有解决自然语言的一切问题, 反而带来了诸如专业性太强、使用不方便、维护和更新困难等许多致命性的不足之处。而自然语言本身的特点又是对规范语言先天不足的弥补, 使得人们转而研究自然语言, 并把它作为情报检索过程的语言保障。

而就自然语言本身来说, 它很早就被纳入了情报检索语言系统。只是由于长期以来检索手段落后(主要是手工检索), 使得规范语言在当时文献数量有限的情况下, 发挥出了方便快捷的作用, 使当时并没有显露出本身优势性的自然语言逐渐被"规范化"的分类语言和叙词语言所替代。

而当规范语言日益暴露出自身的弊端, 并且随着文献数量的急剧增加和电子计算机及其技术的更新换代, 并被用于情报检索过程时, 规范语言越来越不适应情报用户的检索需求。关键词这种自然语词和可以用"后控"方法改造的一部分不够规范的自然语言以其自身的优越特点和计算机等先进技术结合, 满足了情报用户的需求和愿望, 于是重新受到了重视。

特别是随着 20 世纪中叶以来, 计算机技术和通信技术的飞速发展, 加之图书情报工作者的不断努力开拓, 使得情报检索由原始的手工检索发展到联机检索的初步自动化阶段, 又由联机检索软件只能利用规范化叙词语言进行布尔逻辑检索的第一代发展到能利用自然语言进行"语境逻辑"检索的第二代。计算机容量的增大, 运行速度的加快, 检索能力扩充, 多用户共享的巨大变化, 使一度遭冷遇的自然语言显露出它的灵活、快捷、方便的特性, 从而使自然语言在情报检索方面又有了新的用武之地。虽然我国的图书情报检索自动化技术起步比较晚, 自然语言化的检索系统尚在探索阶段, 但国外已有了一定规模, 并显露出迅猛发展之势, 这将值得我们借鉴。英国的 INSPEC I 1967 年投入试运行, 1971 年即开始用自然语言试标引, 两年后的 1973 年便将其作为该系统的检索语言之一。DIALOG 系统和 ESAQUEST 系统的数据库中, 自然语言已占相当比例。据 1975 年对国外的部分数据库调查, 其中仅仅采用受控语言的数据库只有 22 个, 用自然语言的数据库有 18 个, 同时采用受控语言和自然语言的有 48 个。有人认为: 概念有限、结构复杂的叙词型情报检索语言已不能适应数量众多、要求各异的联机终端用户的检索要求, 情报检索又随之向自然语言方向发展。美国的俄亥俄州大学图书馆馆长, 华裔图书馆学家李华伟博士也预言: 未来的情报检索语言是以自然语言为主的发展方向。

3 自然语言和受控语言的优缺点对比及二者的互补性分析

受控语言的实质是表达文献情报特征的概念及其相关的概念标识系统, 它与自然语言的最大区别在于为了特定的需要对自然语言进行了人为的控制, 以便能唯一的表达事物。因此, 受控制语言通常具有以下优点: (1) 能简单明白又比较专指地表达文献以及检索课题的主题概念; (2) 容易将概念进行系统排列, 在检索时便于将标引用语与检索用语进行相符性比较; (3) 语词与概念一一对应, 能控制同意词、多义词和其它语义上的相关的词, 排除了多词一义和一词多义及词意含糊的现象。这样也可以使相关文献相对集中, 也能提高标引的一致性, 容易取得高查全率; (4) 能显示概念之间的相互关系, 如: 等级关系、上下位关系等。

受控语言为克服自然语言的不足而产生, 但在它的使用过程中却带了一些新问题, 从而构成了受控语言的局限性, 这就形成了受控语言自身无法解决的明显缺点: (1) 词汇管理费用高; 标引工作负担重、速度慢, 且成

本高。(2) 事物概念表达显示方面的局限性，使得文献主题概念转换成规范化索引词汇时易造成某些索引词的专指度降低，从而影响查准率。(3) 结构复杂性和易用性之间的矛盾难以克服，使得一般用户掌握这种特殊语言比较困难。就是对于专业情报工作人员来说，要掌握一门检索语言也要先进行培训和实际操作体会，还涉及许多微妙的经验细节，这些都得经过很长的时间。而对于一般的非专业情报用户来说那就更难了。我个人认为，受控语言的这个缺点决定了它在情报检索中的“市场”越来越窄小，从而决定了它的命运是被别的更先进的语言代替，或者是与其它语言相结合形成新的检索语言和检索系统。(4) 自然语言转换为规范语言时的情报失真不可避免。(5) 受控词表的编制、维护和更新难度大且成本费用高等。

而受控制语言的这些局限性反映在自然语言上却变成了其优点：(1) 自然语言符合客观需要，它可以不受限制地随时输入新词，因而可以跟踪学科发展，加速机检数据库的建设。(2) 自然语言相对于受控制语言来说，具有易用性，检索方便、简单。用户只要不脱离文献中的主要自然语言，便可以任意检索，既不受词表控制，也不需要培训，查询快。(3) 正是由于自然语言是文献作者的书面语言，用作情报检索能更好地体现文献保障原则。各学科的用户进行检索时一定会感到使用本学科领域的自然语言要比使用受控词表中的语词方便得多。(4) 自然语言的标引简便，易于实现自动化，标引速度快。(5) 自然语言是完全专指的，它可以使用文摘，索引或文献正文中出现的任何一个有实际意义的词进行检索，甚至可以指定检索的词在某一段落或某一句子中出现，因而有较好的检准率。(6) 遗漏率低。使用自然语词可能提供多条检索入口，从而可以避免由于检索入口少而造成的检索遗漏。(7) 统一性好。采用分类语言和叙词语言标引，依靠人工选择，标引人员的素质和理解、判断等方面的差异，往往造成归类和选词不同，而用自然语言标引在较小范围内采用“现成词”，即使多人标引文献，差异也不会太大。自然语言的这一特点最终决定了它能在广大潜在用户中受到欢迎。拥有广阔的市场。

虽然自然语言能较好弥补受控语言的许多不足，但其自身也有不足之处：(1) 不能反映概念词间的一一对应关系，也不能反映概念关系的隐含性，因而无法排除同义词、近义词、多义词等的词间含糊现象，从而影响查全率。(2) 由于选词没有严格限制，词量势必过多过杂，反而会分散主题，影响查准率，并且会过多地占用磁盘存贮空间。(3) 由于一个概念可以用几个不同的词汇来表达，使得相关文献不能相对集中，检索时容易漏检。

通过以上二者优缺点的比较，我们可以清楚地看到，自然语言和受控语言具有天然的互补性，这是二者能够结合发展为一种以自然语言形式为主的高级检索语言的先天优越条件。

4 对自然语言进行适当控制的方法探索

从上面的比较中可以看出，自然语言和受控语言各自的优缺点在很多方面往往表现为一种互逆关系。对于受控语言来说，它采用的是事先规范的方法，可称作为一种先控系统。而对自然语言来说，纯粹的自然语言检索系统在实际中几乎是不存在的，正如张琪玉先生在《情报语言基础》中指出的，“这种纯粹的自然语言检索如果说不是不可能的，也是低水平的。”因此，在实际操作中，通常要对自然语言采取一些辅助措施，以弥补其缺陷。采用的方法有：

(1) 事先控制法。就是在文献检索要求输入系统时进行控制，而在输出时不加限制。也就是说，检索者可以任意选择他们所需的

词汇, 然后通过一种入口词表把这些词"转变"成受控词。这种入口词表是出现于文献或提问中的自然语言表达方式的词表, 它可以提供一种同现有的受控词表系统相联接的自然语言接口方法。编制一个范围广泛的入口词表, 可能相对费用很大, 但它对改善受控语言的性能, 发挥自然语言的优势有着重要的作用。它能减少查全失误, 也能提高查准率, 还能减轻标引人员和检索者的负担。

(2) 事后控制法。就是在文献检索要求输入时不进行任何控制, 仅在输出时进行不严格的控制。这就是使用一种只供检索的叙词表(也可以称作自然语言叙词表), 也可称作后控制词表。它是对自然语言进行辅助的一种手段, 用来控制自然语言系统中的同义词或句法上相关的词。它具有自然语言所具有的优点, 亦能弥补受控语言在处理文献中新科学、新技术的主题时的不足, 以及由于主题概念转化所引起的专指度下降的弱点。利用这种后控制词表, 可以按高专指度用文献中的自然词进行检索, 又可以按便于族性的要求使用后控制词表中的词族进行检索。由此可以看出, 事后控制法兼有传统的受控语言和自然语言的长处, 不失为一种很有发展前途的新型检索方法。

(3) 设计一种混合词表。它是一种比较粗泛的控制词表, 可能只包括几百个词汇, 但全部是系统的上层结构。标引文献时, 使用一个或几个这种较粗泛的叙词, 同时也可以从文献题目或正文中抽取自然语词标引。这种方法把受控语言同自然语言结合了起来。自然语言词汇可以使检索有一定的专指度, 而粗泛的控制词方便了族性检索, 并可给出自然语言的上、下方。有限的控制词表同没有任何控制的自然语言结合使用, 将会提供强大的检索能力。我国现在已有了一部大规模、综合性的《汉语主题词表》, 可否在此基础上进一步搞混合检索, 还需图书情报界的同仁们做探索。

使自然语言通过适当的控制, 或者说使自然语言和受控语言兼容, 弥补自然语言的缺憾, 将会显露出自然语言在情报检索方面的更大魅力。

通过上面几个方面的探讨及国外的发展情况来看, 单纯的受控语言检索系统和自然语言检索系统都会由于其自身的弊端而使其发展受到限制。而在一个系统中, 自然语言与受控语言结合使用, 发挥二者的互补优势, 将成为未来情报检索的发展趋势。

参考文献

1 徐成兵. 情报检索中的受控语言和自然语言. 情报杂志, 1998, 17(1)
2 李发通. 谈自然语言检索的发展. 情报理论与实践, 1997, 20(5)
3 张玉麟, 文榕生. 论文献检索语言的发展趋势. 图书馆, 1995, (4)
4 徐燕萍. 论受控语言和自然语言的兼容. 情报业务研究, 1998, 5(1)
5 戴维民. 自然语言标引与检索的现状及趋势. 情报业务研究, 1991, 8(4)

寇钧锋　现在陕西师范大学图书馆工作. 通讯地址: 西安市, 邮编 710062。

(来稿时间: 1998-12-21. 编发者: 李万健)