Journal of Information No. 11, 2003

基于关键词的网络信息资源检索

高 凡

(中国科学院文献情报中心 北京 100080)

摘 要 分析和归纳关键词语言的优缺点及网络检索中采用的一系列提高关键词检索效率的方法,探讨网络环境下关键词检索的发展前景和趋势。

关键词 关键词语言 搜索引擎 网络信息资源检索

关键词又称自由词,属自然语言范畴。关键词是反映文献主题概念,具有实际检索意义,主要从文献中直接选取,未经规范,用以标引和检索文献信息的词语。目前基于关键词的检索方法是几乎所有搜索引擎都采用的方法,因而成为网络检索的主要方法之一,其它的还有诸如以主题网关为代表的以元数据为基础的检索和深层网络资源(Deep-Web)检索等。虽然在情报检索早期,关键词语言因其在手工环境下难以控制它的词汇而备受冷落,但随着计算机的广泛应用尤其是网络的迅速发展,知识更新速度的加快和数字化信息的海量化增加,关键词语言再度受到人们的重视,成为搜索引擎的必备检索方法之一。

1 网络信息资源检索采用关键词语言的原因

1.1 *可采用自动标引,方便易行,成本低* 关键词是自然语言,不需要人工标引,因而搜索引擎软件可采用自动跟踪标引软件,如"机器人"〔Robot〕、"蜘蛛"(Spider)、"爬行者"(Web crawler)、"漫游者"(Web wanderer)、"蠕虫"〔Worm〕等,自动从网页中收集关键词,并建立索引数据库,提供关键词检索,节省了编制维护和标引作业的成本。面对这迅速膨胀的网络信息资源,要想完全用人工标引方式是不可能的,自动标引方式适应网络的发展,且建库速度快、效率高、成本低。因此,这种方式受到搜索引擎的青睐,并迅速为几乎所有搜索引擎采用。

1.2 *符合检索语言的文献保障原则和用户保障原则* 在情报检索中,文献保障原则指选词应有文献为依据,用户保障原则指拟选择的词在学术交流和信息检索中是否被人们经常使用。由于关键词直接来自网页和文献资源本身,专指度高,用户可任意检索,不受词表的控制,也不必对用户进行培训,就可较自由地表达主题概念和信息需求。检索方便、简单,而且对各学科的专业用户而言,使用他们自己本学科领域的自然语言检索更加简便易行。

1.3 *数据库更新速度快* 由于搜索引擎采用自动跟踪和标引软件,可以跟踪科学技术发展最新水平,及时增、删、改新词,时效性好,数据库更新速度快,某些搜索引擎几乎是随时更新。

1.4 *有可能达到较高的检准率* 关键词语言是完全专指的,它可以使用网站、网页的题名、摘要、全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,因而对那些确知名称的信息进行检索时,有可能达到较高的检准率。

2 关键词语言制约着网络信息检索的进一步发展

2.1 *关键词语言难以反映词间的相关关系* 在关键词之间存在着大量的同义现象、近义现象、一词多义和同形异义现象,而搜索引擎极少进行规范化处理,致使文献和检索提问中隐含的概念或需求往往难以表达出来,漏检率较高,甚至有时影响到检准率。特别是用单个关键词进行检索时,会检出一大堆无用的信息,有时达到无法容忍的地步。

2.2 *分散主题,影响查准率* 由于关键词选词没有限制,造成词库词量偏大且杂乱,反而会分散主题,影响查准率。搜索引擎对用户的最大困扰就是总能检索出一大堆无用的信息!

2.3 *建库成本低是以用户后期成本的付出为代价的* 搜索引擎对自动采集标引的网页不做筛选和处理,用户检出的资源丰富但质量参差不齐,甚至带来"信息垃圾",需要用户花大量的时间和精力去判别、选择自己所需的信息,因而加大了用户的负担。

2.4 *自动标引无法完全解决标引不一致的问题* 人们普遍认为,采用自动标引可以排除人工标引时由于人与人之间认识上的差异和同一个人在不同时间认识上的差异而造成的标引不一致,只要保持同样标引软件和抽词词典,则标引结果是不会有差异的。事实并非如此。设想一下,如果不同的著

者对同一内容或同一主题的文献采用了不同的表达方法，则标引结果就不会一致。这种情况是大量存在的，因而仅仅靠自动标引本身是无法消除标引不一致的问题的。

3 网络检索中改善关键词检索的方法

就搜索引擎的理论和方法而言，依据的主要是情报检索和数据库系统的理论和方法。因此，搜索引擎的研究引起了情报检索专家的重视。在计算机、情报以及其它相关领域专家的协同努力下，搜索引擎的检索功能在不断地发展完善中。关键词检索是搜索引擎基本的检索方法之一。但是采用简单的关键词检索方法容易造成检索结果过多，检全率和检准率都无法满足用户的需求。因此越来越多的搜索引擎都采用了强化关键词检索的措施，以提高关键词检索的效率。

3.1 *布尔检索* 利用布尔逻辑算符进行检索词的逻辑组配是情报检索系统和搜索引擎最常使用的一种方法。常用的布尔逻辑算符有"AND"(*)、"OR"(+)、"NOT"(-)三种，如输入"数据库 AND 管理"，检索结果同时包括"数据库"和"管理"两个词；输入"数据库 OR 管理"，检索结果中至少包括其中的一个词；输入"数据库 NOT 管理"，检索出的结果中包括"数据库"但是却不含有"管理"一词。因此，正确使用布尔逻辑算符既可以提高检准率(AND、NOT)，又可以提高检全率(OR)。应当注意的是，检全率和检准率两者之间具有互逆的关系，不可能在提高检准率的同时提高检全率。

3.2 *截词检索* 截词检索也是一种常用的检索技术，尤其是在西文检索工具中更是广泛使用。西方语言的构词灵活，在词干上加上不同性质的前缀和后缀，就可以派生出很多新的词汇。这些词之间的基本含义是一致的，如果不采取措施在检索式中列出一个词的所有派生形式，就容易出现漏检。截词检索就是防止漏检的有力措施，因此大部分搜索引擎都具有截词检索的能力。截词检索指的是用截断的词的一部分进行检索，并认为凡是满足词的这部分的所有字符串的记录均为检索命中的记录。截词检索有右截词(后端截词、前方一致)、左截词(前端截词、后方一致)、中间截词(前后方一致)和左右截词(中间一致)。但在搜索引擎中最常见的是右截词方法。使用截词检索可以提高检全率，因为截词检索具有字面成族的作用。

3.3 *词位置检索* 即要求检索词之间的位置满足某些条件，从而增强选词的灵活性，部分地解决布尔逻辑解决不了的词间关系问题，提高检索水平和筛选能力。如输入"数据库 ADJ 管理"，表示"数据库"在"管理"之后紧接着出现；输入"数据库 NEAR/n 管理"，表示"数据库"在"管理"附近 n 个词范围内出现；输入"数据库 W/n 管理"，表示"管理"出现在"数据库"之后 n 个词范围内，因此采取词位置检索可以提高检准率。

3.4 *限定检索* 在搜索引擎中采用了一些缩小或约束检索结果的方法，称之为限定检索。限定检索的方式有很多，如采用字段检索来限定检索词在数据库记录中出现的字段范围，可以是网站、网页或网页的层次、标题、内文、URL 等，还可以限定日期、语言、类型、范围、收费情况及是否是专家推荐等，一般而言，在搜索引擎中限定检索是以高级检索的形式出现的。通过该方式可以过滤一些不必要的信息资源，提高检准率，节省用户的时间和精力。

3.5 *加权检索* 加权检索是对检索词之间的组配关系从量上加以限制和表示的一种方法，它也是对布尔逻辑的改进。布尔检索不能列出每个检索结果的重要性等级，而加权检索通过判定检索词或字符串在满足检索逻辑后对文献命中与否的影响程度，根据权值的大小，即相关度的高低，依序输出检索结果。在实际使用中，并不是所有的搜索引擎都提供有加权检索功能，并且即使提供有加权检索，其加权方式、权值计算和检索结果的判定技术方法都是不一样的。

3.6 *其它方法* 在搜索引擎中，专家们还采取了其它的许多方法，旨在从不同的途径提高关键词检索的效率。如将组成词组或短语的若干词加双引号作为一个关键词进行检索的词组检索和短语检索，用自然语言语句表达检索要求的自然语言语句检索，以及在检索结果内的二次检索等方式。但是我们也应该看到，关键词语言本身所具有的优点，也恰好是它的缺点。尽管单纯的关键词检索简单易学易用，但是在采用了多种加强措施后，用户的负担变得越来越重，某些搜索引擎中的高级检索甚至变得比分类浏览还要复杂，关键词检索易用性的优点也在逐步丧失。如何在保障用户易用性的前提下，对关键词语言进行充分的分析、研究和改进，使之更适于网络检索的发展是首要的问题。

4 基于关键词的网络信息资源检索发展前景与趋势

从检索语言本身的发展前景看，无论如何，自然语言都代表着网络信息时代检索语言的发展方向。但是自然语言并不能完全取代人工语言，未来的发展趋势是人工语言和自然语言从互相结合到完全融合的过程。即专家们所预测的，是人工语言的自然语言化和自然语言的人工化。关键词语言是自然语言，从情报检索过程绝对不能没有控制这个基本原理看，在关键词语言中引入情报检索的控制原理是关键词语言的发展方向。目前，急需我们解决的问题是如何将控制原理应用到关键词语言；如何改善关键词检索难以反映词间相关关系的问题；如何利用最新的信息检索技术改善基于关键词的网络检索效率。

4.1 *人工标引与自动标引相结合* 将人工标引与自动标引结合是前控制的方法之一，在网络环境下，要对数量极其庞大的网络信息资源完全进行人工标引是难以想象的。搜索引擎应当在充分发挥自动搜集、著录和标引信息优势的同时，采用人工标引方式作为必要的补充。即在建库前筛选出一部分质量较高的信息资源进行人工标引，以专业或专题的形式提供特色服务。

4.2 *后控制词表* 后控制指的是在标引(输入)阶段使用自然语言，不对标引进行严格控制，而在检索(输出)阶段才对