●刘 竟 朱书梅 侯汉清

网络环境信息标引的测评与比较研究*

摘 要 网络环境下,文献信息具有数量多、增长快、新词层出不穷等特点。标引是对信息资源进行组织的有效手段和重要环节,标引的质量和效率直接影响信息组织的质量和速度。对受控标引、自由标引和自动标引三种标引方式进行了相符度、专指度、标引深度及通用词数的测试对比,得出自由标引优于受控标引,自动标引优于自由标引及受控标引的结论。图 1。表 5。参考文献 10。

关键词 网络环境 信息标引 受控标引 自由标引 自动标引 标引性能
分类号 G254

ABSTRACT In the networked environment, document information has the characteristics of mass quantity, rapid increase and large quantity of emerging new words. Indexing is an effective measure and important part of information resource organization, and the quality and efficiency of indexing directly affect quality and speed of information organization. The authors test and compare some indicators of controlled indexing, free indexing and automatic indexing, and get the result that free indexing is better than controlled indexing, and automatic indexing is better that free indexing and controlled indexing. 1 fig. 5 tabs. 10 refs.

KEY WORDS Networked environment. Information indexing. Controlled indexing. Free indexing. Automatic indexing. Indexing performance.
CLASS NUMBER G254

网络环境下的文献信息呈现两大特点。

(1)数量多,增长快。20 世纪末以来,网络信息急剧膨胀。根据中国互联网络信息中心对 2005 年中国互联网络信息资源数量的调查,截止到 2005 年 12 月 31 日,全国网页总数约有 24.0 亿个,一年内增长 17.5 亿个,年增长率高达 269%。著名的网络搜索引擎 google,其目录中收录了 80 亿多个网址。清华同方的《中国期刊全文数据库》是目前世界上最大的连续动态更新的中国期刊全文数据库,收录 1994 年至今约 7486 种期刊全文,至 2006 年 11 月,累积期刊全文文献 2161 万篇。通过对 2006 年 11 月 28 ~ 30 日 3 天该数据库公布的新增论文数量的统计,平均每天增长 23228 篇。由此可见,网络环境下,文献信息资源的数量非常庞大且增长速度惊人。

(2)新词层出不穷。据中国语言文字工作委员会做过的一个保守统计,中国自改革开放的 20 年来平均每年产生 800 多个新词语[1],每天都会有 2 ~ 3 个新词出现。随着计算机及互联网的发展,新词出现的速度更快。

目前国内大多数搜索引擎的检索是基于关键词的字面检索,存在诸如检准率低、搜索结果中存在大量的无用信息、用户花费时间长等问题。面对增长迅速的信息资源及不断涌现的新词,基于语义的标引工作变得更加重要,任务也更加繁重。

1 基于概念语义的标引方式

标引是指在文献信息的处理过程中,将文献的内容特征和外部特征分析转换成检索标识的过程。其目的在于使文献管理者能够有效地组织文献,并使文献的利用者能够迅速准确、全面地查找所需文献,实现概念检索[2]。

其中,从信息资源内容特征的角度进行的标引是标引的重要形式,包括分类标引和主题标引两种。主题标引是依据一定的主题词表或主题标引规则,赋予文献信息语词标识的过程。在本文中标引特指主题标引。图 1 是文献信息标引的示意。

从图 1 看出,自由标引以及受控标引都需要标引人员对标引源(如题名、文摘、全文)进行浏览阅读,然后采用概念分析的方法,概括、提炼和选择文献中具有标引价值的主题概念。所不同的是,自由标引是

* 本文受科技部"社会公益研究专项"资助(项目编号 2005DIB6J028)。

标引人员提取出主题概念后,不查看词表,而是按照一定的标引规则,自拟标引词;受控标引则需对提取的主题概念查表选词,进行概念转换,即将自然语言词转换为受控词。

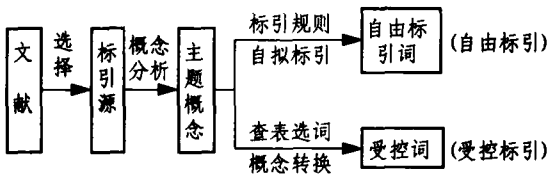

图1 标引示意

根据文献内容,依靠计算机系统全部或部分地自动给出标引符号的过程为自动标引,又称机器标引,分为自动抽词标引和自动赋词标引。自动抽词标引是指利用计算机直接从文献题名、文摘或正文中抽取关键词作为标引词,标引词直接来源于文献本身,不进行规范化控制;自动赋词标引,则是在自动抽词标引的基础上,依据自然语言词汇和叙词表中受控词汇的对应关系,将抽取出来的关键词自动转化成受控的语词。几十年来,国内外很多学者致力于自动标引的研究,出现了很多自动标引的系统和方法,按采用的理论来划分,主要有统计法、语言法和人工智能法3种类型[3]。其中基于词典的统计标引法是目前自动标引方法中比较成熟的一种方法。

由以上对3种标引方式的比较可知:受控标引的标引词由词表进行控制;自由标引的标引词通过标引规则控制;自动标引的标引词由机器进行控制。

在期刊论文网络数据库中,重庆维普《中文科技期刊数据库》和清华同方《中国期刊全文数据库》使用的是不依靠词表的自由标引;前者经数据库编辑人员加工审定,后者直接采用著者自标关键词。上海图书馆的《全国报刊索引数据库》主题标引基本按照《中国分类主题词表》并根据需要适当增补自由词,属受控标引。

2 标引性能定量测试与比较

为了比较受控标引、自由标引和自动标引3种标引方式的标引性能,我们从定量的角度对三种标引方式进行了测试比较。

2.1 测试数据

为了反映网络环境下文献信息的情况,我们采用了150条上海图书馆《全国报刊索引数据库》中的2005年财政类期刊论文标引数据,以及从互联网中下载的150篇财政类网页作为测试数据。为了进行对比实验,我们根据《中国分类主题词表》对150篇网页进行了受控标引;依据自由标引的标引规则和方法对上述300条测试数据进行了自由标引;期刊论文以题名、文摘为标引源,网页以标题、网页全文为标引源,使用中国农业遗产数字化研究中心实验室"基于知识库的自动标引与自动分类系统",对300篇期刊论文和网页进行自动标引,得到300篇文献信息的自动抽词标引结果,并在系统中使用同义词表对同义词进行了规范。

2.2 测试方案

文献主题标引质量可以通过多种因素进行衡量,比如标引的准确性、成本费用、标引速度等,但最核心最关键的因素包括准确性、专指度、网罗性、一致性以及适用性[4]。

除了以上衡量因素,仲云云、侯汉清等在测试网页自动标引方案的标引性能时[5],从手工标引的前3个词和全部词的词形分别与自动标引结果进行比较,设计了4种测评方案:①手工标引前3个词与自动标引全部词相比,仅限于词形完全相同的百分率;②手工标引前3个词与自动标引全部词相比,除词形完全相同外,还包括同义词、准同义词和组代关系词的百分率;③手工标引全部词与自动标引全部词相比,仅限于词形完全相同的百分率;④手工标引全部词与自动标引全部词相比,除词形完全相同外,还包括同义词、准同义词和组代关系词的百分率。其后,王兰成在测试自动标引系统性能时,提出了最大相似率和基本相似率的概念[6]。最大相似率是指自动标引结果中存在的手工标引结果词与手工标引结果全部词之比,仅限于词形完全相同的百分率;基本相似率是指自动标引结果的全部词或其同义词、等级关系(三级以内)词中存在的手工标引结果词,与手工标引全部词之比的百分率。最大相似率与仲云云等人的第三种测试方案相同;基本相似率与第四种测试方案类似。

在对比时,我们将标引词的相符分为4种情况:相同词、同义词或准同义词、等级关系词、组配关系词。同时考察待对比标引方式中未在对比标准中找到上述相符关系词的标引词数量。比如自由标引结果为"出口产值,固定资产,劳动力成本,资本效率",自动标引结果为"劳动力成本,资本效率,出口,资产,机器制造业,产值",若以自由标引词作为对比标准,则两种标引方式中,相同的标引词有2个,"出口产值"与"出口"、"产值"是组配词;"固定资产"与

"资产"为等级关系词;相符度比较结果为:相同词2个,同义词个数为0,等级关系词1个,组配词1个,新增词1个。

我们将"相同词"、"同义词或准同义词"、"等级关系词"以及"组配关系词"作为相符词,将四者之和所占对比标准全部词的比例作为"相符度"。用"相符度"作为衡量两种标引方式的接近程度。

2.3 测试指标

2.3.1 通用词

所谓通用词,是指那些没有独立检索意义的泛指词,如"意义"、"影响"、"对策"、"原则"等。在主题标引中,过多使用通用词,会降低标引质量,影响检索效率。

通过人工判断、统计,受控标引、自由标引及自动标引的总词数及通用词数见表1。

表1 3种标引方式各自的标引词总数及通用词数

	受控标引	自由标引	自动标引
标引词总数	1459	1100	1783
通用词数	162	10	178
比例(%)	11.1	0.91	9.98

表1表明,自由标引的通用词明显低于受控标引和自动标引。这主要是由于自由标引不受词表的限制,使用与文献主题概念最专指的词标引,限制通用词的使用。

2.3.2 相符度

由于通用词是一些没有独立检索意义的泛指词,因此在进行相符度的统计计算时,我们排除掉了标引结果中的通用词。进行相符度测算的受控标引词数为1297,自由标引词数为1090,自动标引词数为1605。

(1)受控标引与自由标引进行对比时,我们把受控标引作为对比标准,计算自由标引与受控标引的相符度,分析自由标引与受控标引的接近程度。自由标引与受控标引相符度比较见表2。

在自由标引与受控标引数据中,21.43%的标引词相同,11.64%的受控标引词在自由标引词中有同义词或准同义词,26.99%的受控标引词在自由标引结果中有其等级关系词,14.50%受控标引词可以由自由标引中的标引词组配得到。因此,自由标引与受控标引的相符度为74.56%。同时,自由标引结果中新增加了273个词,占自由标引词的25.05%,这一方面是由于人工标引时标引人员对文献主题概念分析存在差异,导致标引的不一致性;另一方面是由于受控标引需要查表选词,成本高,不可能对文献进行全面标引,而自由标引不受词表的限制,对新词的反映速度快,且标引成本较低。

表2 自由标引与受控标引相符度比较

	相符情况					新增
	相同词	同义词	等级词	组配词	总计	新增词
词量	278	151	350	188	967	273
比例(%)	21.43	11.64	26.99	14.50	74.56	25.05

(2)自由标引属于人工标引,加入了人的智力判断,与自动标引相比可信度较高。两者进行对比时,我们将自由标引的结果作为对比标准。自动标引与自由标引相符度比较见表3。

表3 自动标引与自由标引相符度比较

	相符情况					新增
	相同词	同义词	等级词	组配词	总计	新增词
词量	277	89	329	134	829	585
比例(%)	25.41	8.17	30.18	12.30	76.06	36.45

25.41%的自由标引词在自动标引词中有相同词;8.17%的自由标引词可以在自动标引词中找到同义词或准同义词,30.18%的自由标引词与自动标引词有等级关系,12.30%的自由标引词可以由自动标引中的词组配得到。将相同词、同义词、等级关系词以及组配关系词相加,可以得出自由标引词中有76.06%的主题概念在自动标引中被标引了出来,自动标引词与自由标引词的相符度为76.06%。同时新增加了585个词,占自动标引词的36.45%。为了解自动标引中的新增词对文献主题的表达能力,我们抽取了自动标引中的100个新增词,查看相应文献的题名、文摘信息。经过比照核查,100个新增词中,有63个词可以作为相应文献的标引词。由此可以判断,自动标引的新增词中有63%的词有标引价值,但在自由标引中没有被标引。这主要是因为自由标引属于手工标引,标引人员的智力负担重,标引成本高,导致某些概念未被分析和标引出来。

2.3.3 专指度

标引专指度是指赋予文献的检索标识与文献实际论述的主题概念的相符程度,通常很难准确测定。

如果标引词是标题词或叙词词串,则可计算它们在词表中的平均级别,即由几个词组成。以此类推,我们可以用标引词的先组度来近似计量专指度。先组度即词汇的先组程度,一般来说,某一词包含的单字越多,它包含的语义越复杂,先组度和专指性就越高[7]。因此我们通过计算词长,即每一标引词包含的单字数量,来测试标引词的专指度。某标引词的词长越长,专指度就越高。

受控标引词、自由标引词及自动标引词的专指度(词长)情况,分别见表4。

表4 3种标引方式的专指度(词长分布)

词长	<=2	3	4	5	6	7	8	9	10	>10	总计	最大词长	平均词长
受控标引	589	134	573	70	83	4	4	0	1	1	1459	11	3.28
自由标引	76	111	554	112	165	30	25	11	11	5	1100	15	4.49
自动标引	868	96	586	92	120	8	10	1	2	0	1783	10	3.20

三种标引方式中,自由标引的专指度最高,受控标引与自动标引接近,低于自由标引。原因是自由标引的标引词是标引人员自拟的、与主题概念最相关的自然语言词,不受词表的限制;而受控标引受词表的限制,造成主题的失真,自动标引也受到抽词词典的限制。

2.3.4 标引深度

标引深度,也称标引网罗度或标引穷举度,是指对一篇文献所给予的全部检索标识的数量。对于主题标引来说,指一篇文献所论述的各个主题概念被确认并转换为检索标识的完备程度,一般用每篇文献中标引词或/及词串的数量来表示[8]。它是根据对文献主题内容揭示的广度衡量标引质量的一个因素,较高的标引深度有助于提高检全率。

根据统计可知,受控标引的平均标引深度为4.86,自由标引的平均标引深度为3.67,自动标引的标引深度为5.94。自动标引的平均标引深度最高,受控标引次之,自由标引最低。

周全明指出,影响标引深度控制的因素有标引语言、经济因素等,一般情况下,标引时所选的标引语言的先组度越高,标引时所需的标引词就越少,反之就越多[9]。对于同属于人工标引的受控标引和自由标引,由于自由标引的先组度高于受控标引,因此标引深度小于受控标引。机器标引的成本低,因此标引深度可能比人工标引(含自由标引)方式高得多,且标引深度可非常方便地由自动标引系统来控制。

2.3.5 综合评定

3种标引方式各有优劣。为了考察它们的综合性能,我们根据以上的定量分析,依据标引性能各项指标排名对它们人工打分,性能指标最高的为3分,排名第二的为2分,最低的为1。3种标引方式各性能指标得分情况见表5。

表5 3种标引方式的性能指标比较

	受控标引	自由标引	自动标引
标引深度	2	1	3
标引一致性	2	1	3
标引专指度	2	3	1
标引速度	1	2	3
标引成本	1	2	3
标引员智力负担	1	2	3
语词更新速度	1	3	2
总分	10	14	18

由表5看出,若综合考虑标引质量、标引成本和标引速度等因素,则自由标引优于受控标引,自动标引优于自由标引和受控标引。

3 结语

由以上关于受控标引、自由标引和自动标引的测定和统计分析,我们可以看出,3种标引方式各有优劣。面对文献增长迅速且新词层出不穷的网络环境,我们认为:

(1)目前仍在讲行受控主题标引的文献信息部门,应逐渐由受控标引过渡到自由标引。前面分析表明,自由标引与受控标引的相符度为74.56%,因此,在标引质量方面,自由标引接近受控标引,但在标引速度、标引成本及标引员智力负担等方面,自由标引明显优于受控标引,且可通过后控制词表方便用户检索,提高检索效率。综合来看,受控标引与自由标引相比,自由标引更能适应网络环境下文献的信息标引。但受控标引仍有其生存空间,书目数据库和电子

政务文件进行受控标引仍非常必要。

(2)自动标引是发展趋势,应当尽早上马。手工标引费时费力,自动标引的效率大大优于手工标引。对比测试表明,自动标引与自由标引的相符度为76.06%,自动标引的质量接近自由标引。随着网络环境下信息资源的飞速增长,完全依靠手工方式对信息资源进行标引显然是不可能的。"由于文献数量和处理时间方面的要求,可以在一定程度上容忍比较粗放的标引处理"。因此在网络环境中,"应当对标引的要求进行相应的调整,降低对标引准确性的要求"[10]。文献信息部门在条件具备时,应当尽早采用自动标引。目前国内一些文献信息部门,如中国医科院信息研究所和中国电信集团上海黄页信息公司等均已转向机器标引;一直使用受控标引的上海图书馆《全国报刊索引》编辑部,从2000年开始与南京农业大学信息管理系合作开发了"《全国报刊索引》数据库自动标引与自动分类系统",该系统已经投入使用。

面对目前自动标引效果仍不能令人满意的现状,建议在自动标引系统投入使用的初期对机标结果进行人工判别(相当于标引终审),以确保标引质量。我们相信,随着自动标引及相关技术研究的深入,自动标引的质量将会不断提高,逐步达到令人满意的效果。

参考文献

1 邹纲,刘洋等. 面向Internet的中文新词语检测. 中文信息学报,2004(6)
2,4,8 马张华,侯汉清. 文献分类法主题法导论. 北京:北京图书馆出版社,1999
3 苏新宁. 汉语文献自动标引综析. 情报学报,1993(4)
5 仲云云,侯汉清,薛鹏军. 网页自动标引方案的优选及标引性能的测评. 情报科学,2002(10)
6 王兰成. 基于EMM中文抽词算法的XMARC主题信息挖掘. 情报学报,2005(1)
7 周小磊,侯汉清. 书目数据库与引文数据库标引质量的测评. 图书馆理论与实践,2003(1)
9 周全明. 六十年代以来我国标引深度研究综述. 情报学报,1994(6)
10 马张华. 论自动标引的实际应用. 图书情报工作,2003(2)

刘 竟 南京农业大学博士研究生。通讯地址:南京童卫路6号。邮编210095。
朱书梅 南京农业大学硕士研究生。通讯地址同上。
侯汉清 南京农业大学信息科学技术学院教授,博士生导师。通讯地址同上。

(来稿时间:2007-04-10)

(上接第59页)识到数字图像信息资源的各种优点,数字图像信息资源的用户接纳程度比较高;数字图像信息资源已经成为了人们学习、工作、生活中重要的信息资源,用户获取数字图像信息资源的途径、方式呈现多样化的特征,用户普遍具备了一定的数字图像信息处理能力。但由于数字图像信息资源本身的特点比较复杂,为了更好地促进数字图像信息资源的应用,不仅需要对应用图像信息处理技术的领域进行研究,还需要提高相关知识的普及率,加强数字图像信息资源开发与管理研究。

参考文献

1 朱学芳. 计算机图像信息资源管理研究. 现代图书情报技术,2004(12)
2 中国互联网络信息中心. 中国互联网络发展状况统计报告(2006年1月). [2007-04-10]. http://www.cnnic.net.cn/images/2006/download/2006011701.pdf
3 中国互联网络信息中心. 中国互联网络发展状况统计报告(2007年1月). [2006-04-10]. http://www.cnnic.net.cn/index/0E/00/11/index.htm
4 Shannon Williams. Information Literacy Instruction for Educatiors: Professional Knowledge for an Information Age. Library Journal. 2004, 129(20): 174
5 [美]安德鲁·利,罗伯特·阿特金森. 冷眼看数字鸿沟. 见:吴士余,曹荣湘. 解读数字鸿沟. 上海:上海三联书店,2003:69
6 朱学芳,智文广. 计算机图像处理导论. 北京:科学技术文献出版社,2003:1
7 心理学名人词典:R·利克特. [2007-04-10]. http://www.whpsy.com/person/1/Likert.R.

朱学芳 南京大学信息管理系教授,博士生导师。通讯地址:南京大学。邮编210093。
袁顺波 嘉兴学院教师。通讯地址:浙江嘉兴。邮编314001。
徐 强 南京大学信息管理系博士生。通讯地址:南京大学。邮编210093。

(来稿时间:2007-05-23)