中文全文检索系统中实现主题词标引思路

吴 春 玉

(1. 南京大学信息管理系 南京 210093;2. 大庆石油学院 大庆 163318)

摘 要 介绍了在中文全文检索系统中实现主题词标引的思路及具体实现过程、各种词表的构造及更新方法和措施、基于主题词标引的优化检索功能等。

关键词 主题词标引 全文检索 关键词标引 主题词表

主题标引技术是科学技术纵深发展带来的产物,国外早在 19 世纪就用主题进行标引,目前,在科技文献检索系统中已全面实现了主题词标引和检索的局面。国内图书情报界则在 20 世纪初期才开始研究,但当时限于人力、物力、财力和技术设备等,以及受战争、政治运动的影响,还有汉语构词及中文书写方式的特殊性,使得汉语标引技术进展缓慢,仅仅停留在理论研究和小范围内的试验阶段。直到计算机日益普及和信息爆炸现象的强烈冲击下,汉语标引技术取得了突破性的进展。如中国学术期刊等全文检索系统,给人们查阅文献带来了极大的方便。然而长期以来计算机自动标引的研究和应用基本上停留在基于关键词的标引,包括中国学术期刊全文检索系统。关键词标引,一方面,由于标引时间短,能够及时反映新出现的专业术语,检索结果查准率高,因此备受青睐。另一方面,由于查全率低,常常因作者的用词习惯和汉语表达的复杂性、多样性造成了标引人员与检索人员之间无法统一思想,导致漏检和误检现象,在检索过程中需要全面考虑检索词,如同义词、近义词、反义词、相关词乃至上下位词等等,给检索者增加了负担,带来了很多不必要的麻烦和不可挽回的后果。那么能否在关键词标引的基础上扬长避短实现主题词标引呢?当大容量计算机的出现和全文检索技术发展已比较成熟的情况下,答案是肯定的。因此,可以说关键词标引的意义逐渐丧失,今后自动标引的研究应从关键词标引全面转向主题词标引的研究与实践当中。

目前中文全文检索系统自动标引采用词典切分法。词典切分法是一种先组式标引方法,检索时无须对字串的字间关系进行组配,检索速度快,但存在着词典的构造困难、更新滞后等不足。词典构造的完善与否直接影响到标引质量,影响检索结果,若在词典的构造和更新方面能够改进,词典切分法将更加完美。

1 主题词标引思路

所谓标引,是指给出信息特征的过程。主题词标引是指抽取信息中能够表达其核心内容的词或词组,并将这些词或词组转化为受控词的过程。这里所指的主题词是某一特定专业检索和标引用的规范词。

具体思路是利用汉语自动分词的研究成果,采用词典分词法将文献进行切分,通过词加权或词频统计法对切分后的词进行排序确定关键词,利用主题词表将关键词转化、合并、去重、重新排序后确定系统正式使用的主题词,并追加文献代号送入系统主题词字段中。在实现过程中,为了继续发扬关键词标引过程中能够及时反映新出现的专业术语,及时更新词表,把原文献给出的关键词一并加入到切分后的词汇集中,进行合并、去重、加权、排序后确定为关键词。在合并去重过程中,我们采用主题词表中的用代关系,将同义词合并、转化为规范词后排序的方法,这样可以避免关键词标引过程中出现的一词多义、多词一义,使得标引人员与检索人员之间出现理解分歧现象和加权过程中同义词分别加权导致文献标引的不准确现象,实现真正意义上的主题词标引。具体的流程如图 1 所示。

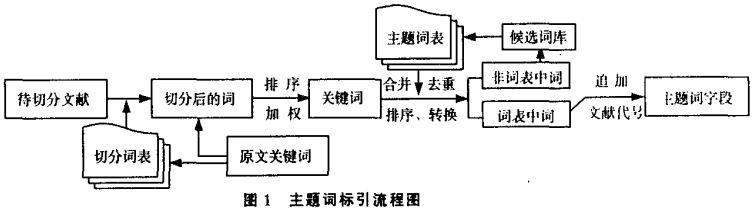

图 1 主题词标引流程图

2 主题词标引的实现

2.1 *构造词表* a.构造切分词表。利用词典法切分标引时需要事先构造词表,将普通词典导入系统中作为切分词的初始依据。为了提高标引的准确性和标引速度,在构造切分词表时尽量把泛滥的通用词和不能做名词和名词性词组的词汇不纳入词表中,即根据词性确定词表用词。b.构造主题词表。归并同义词和关键词转换为规范词需要用主题词表,主题词表包括词关系表和词族表。词关系表是词表中收录的所有有用代关系的词按字顺排列,并展示词的用代关系和族首词。利用用代关系,归并切分

作者简介:吴春玉,女,南京大学信息管理系博士研究生,副研究馆员,研究方向为知识管理、信息处理。

后的词并将关键词转换为规范词，切分后的词与词表进行对照，若是代词(非规范词)则将它转换为用词(规范词)，如图2所示，将电视显像管或监视管转换为显像管。词族表作为扩检和缩检的依据，词族表中的词均为规范词。词族表按族首词的字顺排列，在每一个族首词下按字顺排列其直接下位词，依次每一个词下面列出它的直接下位词，直到词表规定的级别为止，再在每一个词后面列出其相关词(即"C:"项)。通过超链接的方式将词族表中的词与词关系表中的规范词相互链接起来，如图所示将词关系表中的显像管与词族表中的电子管下的显像管链接。这样可以清晰地显示词的上下位关系和相关词，以便于扩检和缩检。主题词表如图2所示。

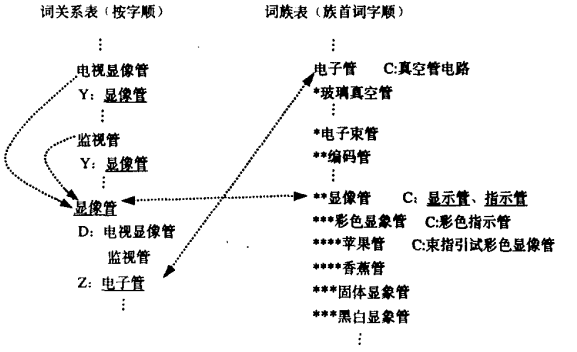

图2　主题词表片段

2.2　*主题词标引*

a.抽取关键词。用词典分词法对需要切分的文献进行自动分词处理，并计算词频和权值，然后对切分后的词进行排序，选出系统规定数量的词汇作为关键词。词频在自动标引中是一个很重要的信息，在一篇文章中出现次数多的词不一定比次数少的词重要。因此，根据实际情况适当考虑相对词频和绝对词频。另外，在一篇文献中词的位置可以说明它的重要程度。在标题、摘要、正文中词的重要程度依次为标题、摘要、正文，在正文中权值的大小按起始段、结尾段、中间段的顺序；在一个段落中是以起始句、末尾句、中间句的次序。因此，根据词的位置给予不同的权值。作者给出的关键词不一定科学、规范，但它一定反映文献的核心内容，因此，我们在确定主题词时把原文献中作者给出的关键词可直接纳入到切分后的词集中，并给予一定的权值，参与词频计算。

b.归并关键词。利用主题词表中的词间用代关系，对抽取的关键词进行合并同类项，即归并同义词、近义词转化为系统所使用的规范词。通过归并关键词能够准确统计词频，真实地反映文献的内容特征。

c.确定标引词。在归并、转换过程中一部分词是主题词表中存在的词，而有些是词表中不存在的新词。对词表中存在的词进行归并、转换，最后将规范词按词频和权值从大到小依次排序，选取系统规定数量的词作为标引词。而那些非词表中的词送入候选词库中，待专家审定。这一环节是全文自动标引过程中最为关键的一个环节，是区别于关键词标引的环节。这里的主题词表类似于手工检索工具中的主题词表，但在构造和工作原理上有所不同。

d.主题词标引。对确定的标引词追加文献代号送入主题词字段中。

3　基于主题词标引的优化检索功能

主题词标引的目的是提高查全率和查准率，减轻用户构造检索式的负担，缩短检索时间，便于扩检和缩检。

3.1　*完全逻辑或运算*　a.主题词字段检索。当用户向系统输入检索词时，系统首先利用主题词表将该词转换为规范词，然后利用转换后的规范词进行检索。如用户输入"监视管"，系统则把它转换为"显像管"在主题词字段上进行检索。b.全文检索。当用户向系统输入检索词并选择全文检索时，将输入的词转换为规范词，并把与该规范词有用代关系的词全部用逻辑或相连后进行检索。如用户输入"监视管"，系统最终用"监视管＋显像管＋电视显像管"在全文中进行检索。在检索结果相关度的计算过程中，同样把所有相关词的词频累计运算，这样可以避免用关键词标引和检索时因对相关词考虑不全面而出现的漏检现象，另外，不论作者的用词习惯如何，利用主题词表标引的结果减少了标引人员与检索人员之间的理解分歧，减轻了检索者的负担，检索者不需要考虑和掌握更多的相关词汇和复杂的词族关系网，使检索系统更加智能化、使用更加简单化，从而提高检索水平和检索效率。

3.2　*扩检和缩检*　扩检和缩检是检索系统应具备的功能。目前基于关键词标引的检索系统中的方法主要是通过字段的选择、运用布尔逻辑算符等来实现，而基于主题词标引的检索系统，则利用上下位词和相关词来调整检索结果实现扩检和缩检，从而增加了优化检索结果的功能。它主要利用词族表来实现。

很多中文全文检索系统的使用过程中发现检索结果不尽人意，要么没有，要么多达几百篇上千篇。尽管系统中设置了二次检索的功能，但大部分用户对词汇的上下位概念和字段的限制方面不十分熟悉，因此，给用户带来许多困难。为了能够把检索结果自动或半自动地进行扩检或缩检，系统中导入词族表。词族表可在直接引用手工检索工具中主题词表后的词族表的基础上增加相关词。在需要缩检(扩检)时根据用户的要求系统自动利用词族表中的直接下(上)位词和词关系表中的参见项来调整检索结果，直到用户满意为止。用户还可以根据个人的爱好选择自动或半自动缩检(扩检)方式。自动缩检或扩检的情况下，扩检时按直接上位词、相关词、上位词的上位词顺序选择词进行检索，而缩检时按直接下位词、相关次、下位词的下位词进行检索，下位词有两个或两个以上时分别列出检索结果，以便用户选择。半自动缩检或扩检时用户根据系统提供的主题词表选词进行检索。

4　词表的更新问题

系统中各种词表的构造和词表更新周期的长短会影响系统的最终效果。为了及时添加新内容，可采取以下措施：

a.在标引过程中将原文献的关键词添加到切分词表中进行去重处理，产生新的切分词表。这样可以及时补充新词汇、新术语，解决了过去用词典法分词时词表构造困难、更新滞后等问题，提高了标引的质量和检索效率。b.动员和调动各行各业的人员定期补充各领域的专业术语，层层向上级汇报，形成严格有序的科层制格局。c.推广科技文献写作格式和用词规范化。各出版编辑单位严格把好关，坚决杜绝"字面创新"现象。目前科技界出现所谓换汤不换药的"字面创新"现象较严重，新术语、新词汇泛滥，给文献的标引和检索带来了许多麻烦。科技文献　(下转第119页)

思想、编制模式甚至编制手段及编制技术等。如果这样即便是达到规范控制的目的,但也可能是自然语言最重要的优势——易用性受到严重影响,因此,笔者以为编制后控词表主要是借受控人工语言的理念,而非照搬《中国图书分类法》、《汉语主题词表》编制模式、编制手段及编制技术[9]。

2.2 超文本方式 超文本指的是一种电子文挡,其中的文字包含有可以链接到其它字段或者文挡的超文本链接,允许从当前阅读位置直接切换到超文本链接所指向的文字,通常使用超文本标记语言(Hyper Text Markup Language. 简称 HTML)书写。大多数网页都属于超文本[10]。事实上每一个网页都是一个文档,由于这个文档内含有 HTML 指令,所以这些文档又被称为 HTML 文档。超文本方式是将网上相关文本信息有机地编织在一起的信息组织形式,它是以节点为基本单位,节点间以链路相连,将文本信息组织为某种网状结构,使用户可以从任何一节点开始,根据网络中信息间的联系,从不同角度浏览和查询信息,超文本是一种非线性组织方式,能提供非顺序性的浏览功能,比传统的信息方式更为符合人们的联想思维方式[11]。超文本是一种全局性的信息结构,它将文档中不同部分用关键字连接起来,使信息得以用交互方式搜索。超文本这种灵活方便的信息组织方式如果用受控人工语言,比如主题分类语言,为不同专业、不同知识水平网络信息需求者提供按目录逐级搜索的网络搜索引擎,尽管用人工语言组织的搜索引擎其结构清晰,符合人的思维和使用习惯,但是建立这样的搜索引擎却需要大量的人力来搜索,并且主题分类具有很大的模糊性和主观性,用户也不一定知道所需信息属于哪个分类,且检索特定的主题所需的时间较长。如果用自然语言为用户提供关键词匹配的网络搜索引擎,那么用户就可在网络上快速、有效地获取网络信息资源。目前所有的网络信息检索工具都提供关键词检索。关键词匹配方法是最基本的方法,也是常用的一种方法。此方法的核心是关键字符的机械匹配,只要发现该网页中含有这个关键字,该引擎就将该网页作为查询结果返回给用户,还可以结合布尔逻辑运算提供更为复杂的查询方式。当然这种以关键词匹配的网络搜索引擎,由于参与匹配的是字符的外在形式,而不是它们所表达的概念,所以经常出现检索不全、查准率较低的状况[12]。

从以上种种信息组织方式来看,对整个网络资源整序过程,是一个不断自然语言化过程,自然因素逐步被采用,其最终目的是为了最大限度地使网络终端用户从各方面直接介入网络信息资源。情报检索计算机化,使人们对自然语言重新给予肯定。自然语言不依附于特定的数据库,采用自然语言标引和检索基本上可解决检索语言的兼容问题。总而言之,自然语言由于自身的优势使其更适合在网络环境下对网络信息资源标引和满足广大网络用户对网络信息资源的易用性便捷性的需求。

目前,想要在整个网络信息资源中,全面采用自然语言标引和检索,还只是一种理想和希望,因为到目前为止还有许多难题未解决,比如汉语的自动分词、电子文档中自然语言的应用等。总之它还不能在保持全部优势的前提下,充分吸取其它语种的优势,从而满足用户的全部检索需求。换言之,现在自然语言还不可以全面替代人工语言,但目前已经看到自然语言有着广泛的应用市场,只要自然语言不断优化改进检索手段,包括吸取检索语言控制技术如后控技术,并在检索技巧上优化,如在搜索引擎中注入辅助技术等,自然语言将成为网络信息资源存贮和检索的主流语言。

参 考 文 献:

1 张琪玉.情报检索语言.武汉:武汉大学出版社,1986
2 王松林.信息资源编目.北京:北京图书馆出版社,2003
3 胡朝棚,叶新明.网络时代情报语言的路向.情报理论与实践,2000;(4)
4 陈 晶.论网络环境下情报检索语言的发展.情报杂志,2002;(6)
5 戴维民.20世纪图学情报学.北京:北京图书馆出版社,2002
6 戴维民等.文献信息数据库建库技术.北京:北京图书馆出版社,2001
7 张琪玉.论后控制词表.图书情报工作,1994;(1)
8 [美]兰开斯特,F.W. 情报检索词汇控制.天津:同济大学出版社,1992
9 谭惠华.关于网络环境下情报检索语言的探究.河南图书馆学刊,2003;(4)
10 WWW.lblogchina.com /new /source /272html - 2004 - 04 - 02
11 洪 渤.信息网络环境下的信息组织.武汉大学学报(哲学社会科学版),1997;(2)
12 徐海燕.概念搜索引擎 CSE.情报杂志,2002;(1)

(责编:王京阳)

(上接第116页)的写作与文学作品的创作不同,文章的开头和结尾、起始段和结尾段、起始句和末尾句的写作格式和用词不宜多样化,需要进行统一和规范,以便在标引时按词位置进行加权运算,其结果能够反映文献的真实情况。d.对主题词表的构造我们可以借鉴国外大型联机检索系统的做法,不同专业构造不同的主题词表,在构建检索系统时,改变目前检索系统先按年代(并非按专业)再按专业划分文档的方法,可先按专业再按年代划分文档的方式,按专业选择文档进行检索,这样就可以解决综合性检索系统中构造主题词表的问题。在检索结果的显示方式也可以设计成按专业分门别类地显示,以利于用户最终判断取舍,这样在检索过程中用户还可以意外地发现和了解某一技术在其他领域的应用情况。

5 结 语

中文全文检索系统中实现主题词标引和主题词检索是众望所归、迫在眉睫的问题,也是摆在情报检索人员面前不可推卸的任务之一。为了使人类的知识充分发挥作用,为了使检索系统能够真实地反映社会知识资源,避免重复劳动带来人力、物力、才力和时间上的浪费,提高查全率、查准率,减轻用户构造检索式的负担,中文全文检索系统应尽早实现主题词标引。

参 考 文 献

1 苏新宁,杨建林,邓三鸿.信息技术及其应用.南京:南京大学出版社,2002
2 韩客松,王永成.中文全文标引的主题词标引和主题概念标引方法.情报学报,2001;(4)
3 文榕生.主题词标引问题探讨——兼与朱学先生商榷.山东图书馆季刊,2001;(3)
4 潘有能.一个自动分词分类系统的实现.情报学报,2002;(2)
5 张俭恭,陈定权.汉字全文检索系统的关键技术与实现.现代图书情报技术,2001;(2)

(责编:王京阳)

中文全文检索系统中实现主题词标引思路

作者： 吴春玉
作者单位： 南京大学信息管理系,南京,210093;大庆石油学院,大庆,163318
刊名： 情报杂志 PKU CSSCI
英文刊名： JOURNAL OF INFORMATION
年，卷(期)： 2005,24(1)
被引用次数： 4次

参考文献(5条)

1. 苏新宁;扬建林;邓三鸿 信息技术及其应用 2002
2. 韩客松;王永成 中文全文标引的主题词标引和主题概念标引方法[期刊论文]-情报学报 2001(04)
3. 文榕生 主题词标引问题探讨-兼与朱芋先生商榷[期刊论文]-山东图书馆季刊 2001(03)
4. 潘有能 一个自动分词分类系统的实现[期刊论文]-情报学报 2002(02)
5. 张俭恭;陈定权 汉字全文检索系统的关键技术与实现[期刊论文]-现代图书情报技术 2001(02)

本文读者也读过(4条)

1. 刘剑.王兰成.LIU Jian.WANG Lan-cheng 基于主题词表的数字档案馆概念搜索引擎的设计与实现[期刊论文]-山西档案2008(3)
2. 韩客松.王永成.Hah Kesong.Wang Yongcheng 中文全文标引的主题词标引和主题概念标引方法1[期刊论文]-情报学报2001,20(2)
3. 张改侠 主题标引中主题词的组配规则[期刊论文]-图书情报知识2003(3)
4. 刘华.Liu Hua 基于分类标注语料库的关键词标引知识自动获取[期刊论文]-图书情报工作2007,51(7)

引证文献(4条)

1. 李渤波 石油化工汉语主题词表的作用与修订[期刊论文]-当代石油石化 2007(6)
2. 熊回香.夏立新 汉语分词技术综述[期刊论文]-图书情报工作 2008(4)
3. 张朝霞 从PubMed看主题词与关键词的结合运用[期刊论文]-航空航天医药 2009(12)
4. 刘华 基于分类标注语料库的关键词标引知识自动获取[期刊论文]-图书情报工作 2007(7)