中文全文标引的主题词标引和主题概念标引方法[1)]

韩客松 王永成
(上海交通大学,上海 200030)

摘要 中文全文标引正在越来越受到重视。本文主要研究了三个方面的问题,首先是全文主题词标引的加权问题,综合考虑了五个方面的因素;其次是介绍了一种用层次概念词典改进主题词标引质量的新方法;最后又提出了用三种不同的方法产生主题概念进行全文标引的的主题概念标引。受限范围内的实验结果显示本文的方法有一定的理论和实用价值。

关键词 主题词标引 主题概念标引 层次概念词典 自动标引

Methods of Keyword and Subject Concept Indexing to Chinese Full-text

Han Kesong and Wang Yongcheng
(*Shanghai Jiaotong University, Shanghai* 200030)

Abstract Research on full-text automatic indexing is a hotspot of today. This paper's research focuses on three points. Firstly, it discusses the method of keyword weighting, which taking altogether five factors into account; Then it introduces a new way to improve the precision of indexing by using a hierachical concept thesaurus; Finally, it presents three different techniques to produce subject concept, thus to implement concept indexing. The limited-area experiment shows that the methods introduced in this paper have both academic and practical value.

Keywords keyword indexing, subject concept indexing, hierachical concept thesaurus, automatic indexing.

1 引 言

近年来,全文数据库建设在我国取得了较快的发展,比较出名的有《中国学术期刊(光盘版)》,《中国大百科全书》图文数据光盘版,《人民日报》五十年(1945—1995)图文数据库系列

收稿日期:2000 年 4 月 14 日

作者简介:韩客松,男,博士研究生,1973 年生,研究方向为自然语言处理。王永成,男,教授,博士生导师,1939 年生,研究方向为网络智能信息处理。

1) 此项研究成果受到国家 863 计划资助(合同号:863 - 306 - ZD03 - 04 - 1)。

等。

但目前具有全文标引功能的系统还不多，如《中国学术期刊(光盘版)》也只是将期刊原文的关键词输入数据库作为主题词，而没有真正做到全文主题词标引，对文献本身没有带关键词的则连关键词也未给出[1]，究其原因有：

(1)全文标引如果不采用自动标引技术，不仅代价很大，而且速度较慢，影响建库速度。

(2)自动全文标引技术受自然语言处理各方面技术的影响，有待于进一步研究。

与以标题和摘要为主要标引源的二次文献标引相比，全文标引具有如下特点：

(1)可收集的信息多。标题、作者、摘要、引文等文献组成部分提供了丰富的信息。

(2)能够收集到的主题词的个数相对较多，绝对频次也比较高。

如果能将这些丰富的信息有效地利用，则标引结果应该会比二次文献标引具有更好的实用价值，更有效地支持用户的检索。

本文由三个主要部分组成，分别是：介绍全文标引的权重处理方法；提出一种利用层次概念词典提高主题词标引精度的新方法；提出用三种不同的方法产生主题概念进行主题概念标引。

2　全文主题词标引

对全文进行主题词标引的一般流程如下面图 1 所示：设计一种好的分词、抽词算法(包含同义词的转换等)，从原文的各个组成部分，借助于事先编辑好的通用词典和主题词词典(和处理的文献一致的主题词词典)，得到一系列关键词(据统计，一般来说，一篇 5000 汉字的科技文献大概能得到 200 个左右的关键词[2])。然后，对关键词的一些属性进行加权处理，按照权值的大小进行排序，再输出指定个数的主题词(手工标引一般取 3～8 个主题词，机器自动标引时可以适当多取几个)。

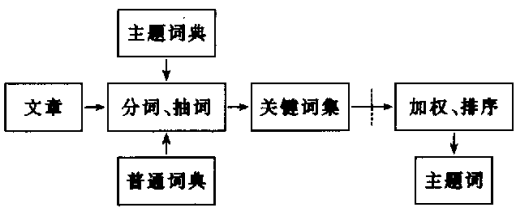


图 1　全文自动标引系统图解

由于标引源是全文，加权处理过程中可以获取的词的属性信息比较多。所以为获得对原文主题更准确的把握，需要一套合理的加权体系。在设计加权体系时，可以考虑的因子有：

(1)词频。作为统计意义上的主题词自动标引，词频是一个很重要的信息，因为，在一篇文章中出现次数较多的词一般比次数较少的词更重要些(除去泛滥的通用词)。根据实际情况可以选用绝对词频和相对词频。

(2)词的位置。词的位置是决定词的重要性的另外一个重要因素。词的位置大概有标题、摘要、正文等几种主要位置。细分下去，正文位置可以分为起始段、中间段和结尾段。再细分，词又可能出现在起始句、中间句、末尾句。

由于 95%以上的科技文献和大多数的其他文献的标题能很好地反映文章的主题，因此，一个词如果出现在标题中，则它的重要性比出现在摘要和正文中的词重要得多。据我们的经

验,标题中关键词比摘要中的大概重要3~5倍,比正文中的大概重要10~15倍。由于中国人写文章一般讲究"起,承,转,合",正文的首段一般简要介绍全文的内容,末尾段再总结一下主题,而在一个段落中,段首段尾句的情况也类似,所以,出现在首段、末段,段首句、段尾句的关键词也比中间段、中间句的重要一些。

(3)词性。由于主题词一般为名词性的,所以,名词最重要,名词兼类词(如"对比"之类的动名词)其次,其余的就相对不重要,这也是一个好的加权体系应考虑的。

(4)词本身的价值。即使是同一主题中的关键词,不同词本身的重要程度也不一样。例如,在生物学中,"脱氧核糖核酸"和"武夷山"都是关键词,但前者是在生物学中,尤其是分子生物学和遗传生物学中一个很重要的词,而后者只是在讲到生态分布时才会用到,所以,这两个词的本身的对主题的价值就不一样。但文献[3]统计后者的出现在农业期刊中也可以达到5.5%,因而这些词也不能被轻易忽略。

(5)词的长度。在长期研究中,我们发现,汉语中,一般来说,长的关键词往往具有较好的专指性,而短的关键词则往往具有较好的概括性。例如,"计算机","电子计算机","数字电子计算机"三个词的专指性依次增强,而概括性依次递减。

一个好的加权体系应该综合考虑上面这些因素。

就目前国内的研究现状来看,阻碍全文自动标引接近使用的最大的障碍是主题词典的建设。好的主题词典的标准是词条收录全,信息丰富,访问快捷,维护方便等,而做到这些需要大量枯燥的手工劳动,比如,上面第(4)点所需的信息没有长期艰苦的手工劳动是几乎无法得到的。

3 概念词典支持的主题词标引系统

传统的主题词标引,基本是在取得关键词的权重,排序后立即输出若干权重相对较高的词,这其实存在很大的缺陷。

我们以下面的例子来解释这一点。假如要求用三个主题词标引某一文章,现在系统已经从该文章中得到了如下关键词,并通过加权系统得到括号中所示的词在文中的权重:

情报(36) 情报检索(31) 情报技术(30) 多媒体(25) 声音输出(20)

按照通常的方法,可以很容易得到本文的三个标引词,依次为"情报"、"情报检索"、"情报技术"。但是,这样的标引结果显然不符合标引要求,它漏标了多媒体相关的主题内容,而重复标引了情报相关的主题。

现在假如我们有一部层次概念词典,则可以通过如下步骤产生主题词:

(1)可以通过聚类和归类,将主题词分为若干词义相近的子集。如,将上面五个关键词聚集成两个子集,Set1("情报"、"情报检索"、"情报技术"),Set2("多媒体"、"声音输出");

(2)再通过概念词典的层次结构构造如图2所示的层次概念树;

(3)为层次概念树中的各个非虚节点重新计算权重,方法是某一节点的权重为其子树各个节点的权重的和,然后再找出权重最大的若干节点为主题词。为实现主题词的分布尽可能均匀原则,首先从各棵树中各找一个主题词,如若还不够,再进行第二遍抽取。

上述例子中,第一个主题词应从(a)图中产生,由于"情报技术"一词还带了子节点"情报检索"。所以应算权重为30+31=61,大于"情报"一词,故应选定"情报技术"为该文的第一主题词。第二个主题词应从(b)图中产生,从"多媒体"和"声音输出"中选取权重大的"多媒体"为第

二个主题词。为使主题词分布尽可能均匀,第三个主题词再从(a)图中产生,这样选"情报(36)"作为第三主题词。

显然,采用本方法的标引结果比直接按权重输出主题词的方法有了明显的改进。我们将整个处理过程用下面的图 3 来表示。

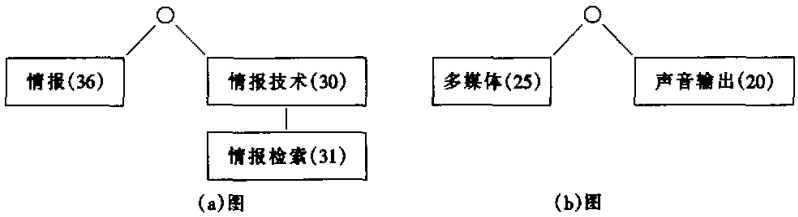
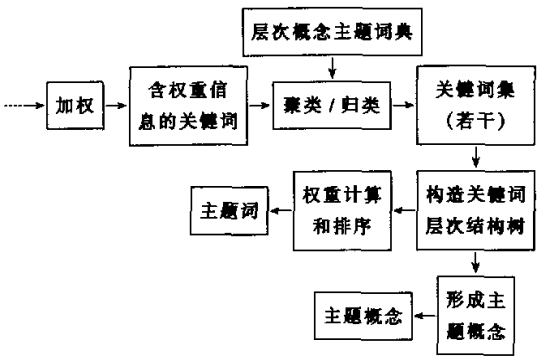


图 2　层次概念树



注:图中 ┄ 为图 1 中虚线的左侧部分处理

图 3　基于层次概念词典的主题词和主题概念标引系统

4　主题概念标引

用主题词标引文献固然不错,但是,主题词的不连贯性,导致检索者很难直接从主题词中较准确的揣摩出文章的主题,因而情报界已经提出了主题概念标引,它的好处是对文章主题的概括能力较强[4]。

我们在对生物学文献进行统计后,发现人工进行标引时,有 42.7% 的主题词是直接从原文中得出,有 47% 的从原文进行同义词转化而来的,剩下的 11.3% 是通过"拍脑袋"(Brain Storm)得到的[5]。这里的第一二部分词基本可以用上面的方法得到,但第三部分只有通过概念标引才能得到。

事实上,为保证所获得的主题词对文献都是足够重要的,我们添加了一个限制条件:设 $w(p)$为词 p 的权重,规定只有当 $\frac{w(p)}{\max\{w(p)\}} \geq T$ 时(T 为阈值),词 p 才有做主题词的资格。这一条件限制使我们得到的主题词不存在滥竽充数的情况,但有时标引的主题词个数偏少。因此,当获得的主题词较少时,加上主题概念将使标引效果得到加强。

目前,我们使用如下几种方法来得到概念词。

方法一:选取直接上位词作为主题概念

这种情况一般用在某主题词在概念层次中没有直接的同义词或准同义词,如:"Windows NT",选取其直接上位词"操作系统"。此时一般应同时将该词和概念词作为标引词。

方法二:通过聚类产生上位词作为主题概念

这种方法一般用在某个主题在层次概念词典中有若干直接同义词并且这些同义词在文章中也出现了,例如某文献中"小轿车"的权重较大,而且也提到了"卡车"、"小汽车"、"摩托车",则可以聚类为"机动车"。此时一般可以只标引概念词。

方法三:有两个(或以上)主题词合成生成主题概念

这种方法一般用在若干主题词在文章的标题或正文的某些分句中同时出现的情况。如果上面例子的文章标题为"多媒体情报技术的研究",则可由"多媒体"和"情报技术"两者相加结合生成新的概念"多媒体情报技术"作为标引的概念。此时一般将原词和概念词同时作为标引词。

很显然,能正确进行主题概念标引的一个最重要也是最困难的问题是获得一部好的层次概念词典。

5 结 论

本文首先主要讨论了全文主题词标引的加权问题,然后介绍了一种用概念层次词典改进主题词标引质量的新方法,最后又提出了用三种不同的方法产生主题概念进行全文标引的主题概念标引。尽管由于目前没有好的层次概念词典,尚不能进行大规模的真实文本测试,但受限范围的实验结果显示本文的方法有一定的理论和实用价值。

参 考 文 献

1 张政宝 . 对中文全文数据库标引和检索功能的探讨 . 情报学报,1997,16(增刊):87~91

2 王永成等 . 中文信息处理技术及其基础 . 上海:交通大学出版社,1991

3 董毅士 . 农业期刊学术论文关键词标引刍议 . 情报学报,1999,18(增刊):96

4 Chen H.,Lynch K.J.. *Automation Construction of Networks of Concepts Characterizing Document Database*. IEEE Transaction on Systems,Man and Cybernetics,1992,22(5):885~902.

5 Han Kesong,Wang Yongcheng,Wang Gang. *The theory and practice of automatic indexing system on biological documents*. International Conference on Machine Translation & Computer Languang Information Processing

(责任编辑 芮国章)

中文全文标引的主题词标引和主题概念标引方法1

作者： 韩客松， 王永成， Hah Kesong， Wang Yongcheng
作者单位： 上海交通大学,
刊名： 情报学报 ISTIC PKU CSSCI
英文刊名： JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATON
年，卷(期)： 2001,20(2)
被引用次数： 36次

参考文献(5条)

1. 张政宝 对中文全文数据库标引和检索功能的探讨 1997(z1)

2. 王永成 中文信息处理技术及其基础 1991

3. 董毅士 农业期刊学术论文关键词标引刍议 1999(z1)

4. Chen H;Lynch K J Automation Construction of Networks of Concepts Characterizing Document Database 1992(05)

5. Han Kesong;Wang Yongcheng;Wang Gang The theory and practice of automatic indexing system on biological documents

本文读者也读过(5条)

引证文献(36条)

1. 袁良平.汤建民 一份翻译研究期刊的学术脉络管窥——《上海翻译》(1986-2007)所刊论文标题词频统计个案研究[期刊论文]-外语研究 2009(1)

2. 刘海峰.王元元.丘国防 密度聚类模式下一种基于层次的自动文摘方法研究[期刊论文]-情报杂志 2007(3)

3. 常鹏.马辉 高效的短文本主题词抽取方法[期刊论文]-计算机工程与应用 2011(20)

4. 章成敏.许鑫.章成志 条件随机场标引模型的性能影响因素分析[期刊论文]-现代图书情报技术 2008(6)

5. 刘远超.王晓龙.徐志明.刘秉权 基于粗集理论的中文关键词短语构成规则挖掘[期刊论文]-电子学报 2007(2)

6. 吴春玉 中文全文检索系统中实现主题词标引思路[期刊论文]-情报杂志 2005(1)

7. 王海英 主题词间后显关系分析[期刊论文]-农业图书情报学刊 2005(8)

8. 吴春玉 中文全文检索系统主题词标引[期刊论文]-情报科学 2004(6)

9. 王泰森 一个基于本体论全文自动标引方案[期刊论文]-情报科学 2003(9)

10. 徐震 主题检索系统的优化技术研究[期刊论文]-情报理论与实践 2010(9)

11. 徐震 主题检索系统的优化技术研究[期刊论文]-现代情报 2006(10)

12. 杨亮.王永成 新型标引系统的构建[期刊论文]-计算机应用与软件 2004(5)

13. 逄焕利.周连喆.刘寒梅.计小宇 基于概念检索的中文搜索引擎[期刊论文]-吉林工学院学报(自然科学版)

2002(1)

14. 宋迪 基于概念提取的邮件自动回复技术研究[期刊论文]-微计算机信息 2008(3)

15. 温有奎. 温浩 关键词与创新点词句群分布分析[期刊论文]-情报学报 2007(1)

16. 叶肖惠. 王素芳. 杨华. 邵伟 Scorpion自动标引思想初探[期刊论文]-图书情报工作 2009(14)

17. 卢娇丽. 郑家恒 基于成对比较的关键词权重计算与主题词抽取[期刊论文]-山西大学学报（自然科学版） 2005(1)

18. 张清军. 朱才连 基于统计的中文文本主题自动提取研究[期刊论文]-四川大学学报(工程科学版) 2004(3)

19. 向桂林 学科分类知识库的构建及其在网络资源分类中的作用[期刊论文]-图书情报工作 2003(2)

20. 王宏生. 高岩 基于本体的信息过滤研究[期刊论文]-科技信息 2009(29)

21. 钟彬彬 中文关键词抽取技术的研究[学位论文]硕士 2005

22. 云飞 数字图书馆个性化推荐系统中信息过滤及其相关技术的研究与应用[学位论文]硕士 2006

23. 张铎予 试论基于文献的知识发现中资源标引的改进[期刊论文]-现代情报 2010(12)

24. 刘远超. 吴冲. 王晓龙 基于多知识源融合的关键词重要性评价研究[期刊论文]-哈尔滨工业大学学报 2007(7)

25. 王建会 中文信息处理中若干关键技术的研究[学位论文]博士 2004

26. 章成志. 张庆国. 师庆辉 基于主题聚类的主题数字图书馆构建[期刊论文]-中国图书馆学报 2008(6)

27. 刘盛博. 丁堃. 王贤文. 刘则渊 基于TF/IDF多因素改进算法的知识单元抽取研究[期刊论文]-情报学报 2011(10)

28. 张静 自动标引技术的回顾与展望[期刊论文]-现代情报 2009(4)

29. 肖慧珍 网络信息资源组织研究[学位论文]硕士 2001

30. 李纲. 戴强斌 基于词汇链的关键词自动标引方法[期刊论文]-图书情报知识 2011(3)

31. 章成志. 苏新宁 基于条件随机场的自动标引模型研究[期刊论文]-中国图书馆学报 2008(5)

32. 章成志. 梁勇 基于主题聚类的学科研究热点及其趋势监测方法[期刊论文]-情报学报 2010(2)

33. 杨晓懿 基于内容分析的信息安全过滤技术研究[学位论文]硕士 2005

34. 章成志 自动标引研究的回顾与展望[期刊论文]-现代图书情报技术 2007(11)

35. 邢玲 基于本体结构的网页信息自动标引技术[学位论文]硕士 2005

36. 张清军 基于位置的服务（LBS）中的文本挖掘研究[学位论文]博士 2005