中文文献主题的自动标引

王永成　顾晓明

摘要

本文介绍了１９９２年我校开发的中文文献主题自动标引系统CSAIS 2 的主要考虑、基本算法及其基本结构。

主题词：自动标引、中文文献、主题、人工标引

<—>

大量地开发中文信息库已成为时代的必需。但中文信息的标引非常困难，已成为开发中文信息库之瓶颈［１］。

困难之一在于信息的标引目前还仃留在手工阶段。即使对西文来讲，手工标引亦有如下的缺点一时难以克服：

一、工作量大。据文献［2,3］报道：因为其工作繁难，国外的信息检索系统中有７５％的运行费用要用于人工标引。

二、效率低。标引员要正确标引一篇文献，往往要一个半小时以上。

三、一致性差。美国的 Cleverton 曾做过一些试验，他指出：

１、二组人为同一主题编出的叙词表，其中词的同一率仅达６０％。

２、二位有经验的标引员用同一叙词表对同一篇文献进行标引，其标引词的同一率仅有３０％左右。

３、二个在同一库中用同一检索系统检索同一问题的人，检索出的结果的同一率仅有40%。

４、二位科研人员根据同一提问判断一组指定文献的相关性，其同一率不会超过６０％。因此，必须发展使用电脑来进行自动标引。早在１９５６年，美国的H.P.Luhn就开始了文献自动标引的试验；６０年代初，美国的G.Salton教授更在自动标引方面取得了世界瞩目的成就；但是，我们中国的中文文献的自动标引工作仅仅开始于１０多年前。

除西文自动标引的困难之外，中文自动标引还要克服中文词的界定的困难。因为迄今为止，全世界都还没有一个为大家所公认的、科学的、经得起推敲的词的严格的定义，更不用说通用的词切分标准了。

自从1980年我们在中文自动分词方面取得初步进展之后，我们就开始了向中文文献自动标引方面的进军。

1985年，我们首先与上海图书馆合作，成功地开发了《中文自动抽词、自动编制索引与自动检索系统》；

1987年，我们成功地开发了中文文献自动赋予分类号的模型系统 AC；

1988年，我们成功地开发了中文文献的自动文摘系统 SJTUCAA；

1990年，我们成功地开发了中文文献关键词的自动标引系统 CKAIS；

1991年开发了无词表中文文献主题自动标引系统 CSAIS 1；

1992年我们又开发了建立在自动分词基础上的中文文献主题自动标引系统 CSAIS 2。本文将对我们1992年开发的系统的基本算法及系统的概况作一个简明的介绍。

<二>

自动标引主题词的困难很多。

其困难首先在于分析和提炼文献的主题需要很高的智能并且它还与标引人的主观立场观点密切相关。毛泽东说＂红楼梦是一本阶级斗争的好教材＂，但俞平伯则并不是这样地认为。我们相信：这二人对＂红楼梦＂的主题标引的同一率一定相差甚大。

困难还在于：同一主题的用词并不一定相同。为了交流与检索的方便，早期的标引要求任何文献一定要用规范化的词进行标引，这就要求标引工作一定要借助于人们事先精心编制的主题词表来进行，也就是要求人们用所谓的＂受控词＂（受词表控制的词）来进行标引。用词表以后，不仅有用词规范划一的好处，而且因词表中已明确指出了词与词之间的＂属＂、＂分＂、＂参＂、＂代＂、＂用＂的关系，因此，在检索时还可借助于词表进行扩检以提高查全率或缩检以提高查准率。

但是，在过去的实践中，它也暴露出以下几个缺点：

1、词表的不完备性影响了标引的质量。

2、词表的更新赶不上时代发展的步伐。

3、庞大的词表，使标引速度大大地下降。

4、标引员与检索员都必需熟悉词表才能工作，使用非常不便。

因此，近年来，人们已不再强调要用受控词进行标引，而可允许人们用自已最方便最熟悉的词进行标引，在检索时再借助于词表进行同义词替代以保证其查全率、用上位词进行扩检、用下位词进行缩检。

另外，由于电脑的语义分析能力还很低下，目前的主题标引，我们还只能仃留在对文中能反映主题的关键词的自动抽取与筛选上。所以，本文的题目不用＂中文文献主题词的自动标引＂，而是＂中文文献主题的自动标引＂，它实质上乃是能反映主题的文中关键词（又常简称为＂主题关键词＂）的标引。

<三>

自动标引文献主题的关键在于如何抽出文献主题。如前所述，归根结蒂，这需要进行语义理解，而且这又与标引人员的立场观点及原有知识密切相关。目前的电脑发展水平还不能完成这样的史命。参照国外的做法和我们长期的实践经验，我们提出了从下述几个方面来考虑从文献中抽出部分词来作为文献的类主题词。

1、只考虑实词。

任何文献中都有不少＂介词＂、＂连词＂、＂助词＂等虚词。这些词在一般情况下，都不必作为类主题词来考虑。对实词，也可根据其可标引性，而给予不同的价值。

2、重视高频实词。

国内外的研究表明：高频实词，特别是文内相对高频实词，往往与文献主题或风格密切相关。

在考虑词的文内相对频率时，我们不仅要考虑某词的完整出现（我们称之为＂全词＂），而且还应把实词的组成部件（简称为词部件）及其代词的出现频率加以统计。在我们的CSAIS 2 中，我们曾规定：

全词在文内的相对频率＝全词在文中出现的相对频率

＋全词中的词部件在文中出现的相对的相当频率之和

而全词中词部件在文中出现的相对的相当频率之和＝

（词部件在文中出现的次数＊词部件的替代度）／词在文中出现的相对频率

其中，词部件的替代度＝组成该词部件的诸词部件的标引价值之和

/组成全词的诸词部件的标引价值之和

代词的指代关系的确定有时非常地困难，特别是处理隐含代词时会迂到更大的困难，但在大多数情况下，把它们的出现也考虑进去将能更精确地反映出一个词的出现频率。

国内外的研究还表明：因为标引的目的在于检索，因此，我们不仅要重视标引词是否能反映主题，还更要考虑标引词的区分文献特性的能力，为此，我们还应考虑实词在文献库中的相对频率。

3、特别看重特征词。

所谓特征词就是包含在文中某些特征部位或特殊句中的实词。

最受人们看重的特征词就是文献标题词。因为文献的标题往往与文献的主题密切相关。国内有人抽样统计了国内中文期刊自然科学论文标题与论文正文主题的符合率，其基本相符的竟高达９９％以上。这就说明：文献的主标题、副标题、乃至小标题中的关键词在文献主题的自动标引中具有极其重要的作用，应予特别的重视。当然，由于其事实上的差别，不同层次的标题应给予不同级别的权值。

ISI的《新刊目次》磁盘版以及G.Salton 等人还特别重视以引文中的标题关键词等信息作为主题参考信息，也有相当的效益。

我校已借助于文献标题等信息，成功地开发的中文文献自动编制文摘系统，这就证明了文献标题对提取文献主题的重要性。

在[1]中，实际上还指出出现＂本文讨论了＂、＂综上所述＂等等所谓的＂予置关键词＂的＂主题提示句＂往往是很好的文摘候选句。因为其中往往高度地概括了文献主题，因此，不言而喻，在这类句子中出现的实词，也应特别被看重。

另外，美国 P.E.Baxendale 的抽样统计表明[4]：反映主题的所谓＂论题句＂，其中８５％出现在段首，7%出现在段尾。因此，段首与段尾的实词，也应适当地加大其标引的权值。

特别有意思的是：文献中用括号括起来的部分，如：ISDN(综合业务数据网)；用破折号引出来的部分，如＂数据的自动识别输入－－条码技术＂；用＂所谓＂所引出的部分，

如"所谓的予置关键词"，其中的实词往往也应当给予特别地加权。

海外早就有人将文献分为三类：

1、规范文。它具有功能词（虚词）介于48%-56%之间，句长平均为18-26个词的特征。

2、浮夸文。它具有功能词超过56%的特征。

3、过精文。它具有的功能词少于48%而且35%以上的实词只出现一次的特征。特别是综合性的科技文献，有时关键词出现次数并不多，但我们的研究表明：很多文献往往以段为单位，每段一个主题，因此，我们可以不只看重词在文中的频率，而且可以先在每段中用各种手段先选出段主题词，然后以段的位置，长短及该词本身的词频等作为该段的加权代表去竞争作为全文主题关键词。

由于考虑到任何一个词在一个段中的频次不可能很高，因此，我们提出了"能反映主题的关键词一般应在某段中有较高的密度"的假说，这帮助了我们顺利地进行段主题词的提取工作。

4、选词适当聚类。

当有一组词形相近或词意相近时，可借助于信息量的分析或词典进行并合[5]。

<四>

继1990、1991年开发的中文文献主题自动标引系统之后，1992年我们又成功地开发了一个基本上基于上述认识和算法的中文文献主题的自动标引系统（CSAIS 2），该系统的总体结构大致如图1所示：

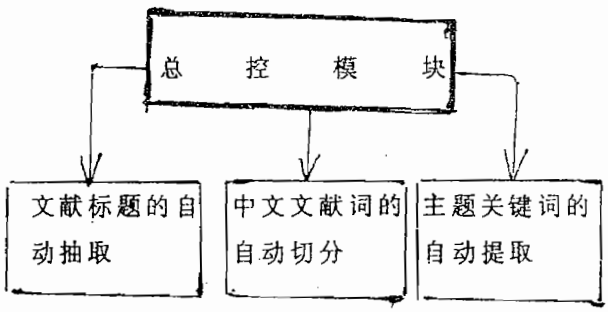

图1：中文文献主题的自动标引系统（CSAIS 2）总体结构

其中总控模块执行人机接口与各模块之间的协调、调度和切换等功能。

人机接口在设计时我们充份考虑了：由于电脑应用的日趋普及，电脑应用已从专业人员迅速发展到广大的非专业人员。因此，易学易用已成为我们的主要设计原则。

我们的屏幕采用的是目前流行的瓦式窗口显示方式。同一屏幕上，我们可同时提供下述六种窗口：

1、被标引的文献的选择

2、文献标题的自动抽取与显示

3、文献正文的逐段显示

4、自动抽出的文献10个主题关键词的显示

5、对抽出的文献关键词进行简单地编辑窗。在其中可以对已抽出的主题关键词进行增、删、改的编辑。当从已显示的10个主题关键词中删去一个时，隐蔽的可作为后补的主题关键词即可自动地补足之。

其中的"标题自动抽取"模块将自动地从文献中自动地识别并抽出大小标题。

其中的"中文文献词的自动切分模块"实际上是把我们早在1990年开发的中文自动切分系统当作一个模块来调用。

利用该系统抽出的词，实际上应称为一种混合了词、词组及少数字串的所谓"辞"，它更接近于所谓的概念词。

<五>

去年我们曾用CSAIS 2 对1986年的《情报学报》中的31篇科技论文进行了试标引。附录中提供了部分结果。

从这些结果人们不难发现：虽然该系统还有很多待改善之处，但总体上讲，效果是令人鼓舞的。

<六>

本文实际上报道了我们上海交通大学电脑应用技术研究所的很多人的集体劳动的成果。除作者们外，周国栋、莫燕、顾立帆等同志也参加了很多开发工作甚至作出了杰出的贡献。以周国栋为主力的1993年版本也已经基本完成，我们将力争在最短期间完成其商品化的任务，让整个社会可以分享我们的劳动成果。

参考文献

1、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，1992

2、储荷婷，索引工作的自动化：自动标引的主要方法，中国索引学会首届年会暨学术讨论会论文（编号37），1992，12(美国德瑞富尔大学)

3、曾　蕾，索引工作的自动化：计算机辅助标引及索引的编制，中国索引学会首届年会暨学术讨论会论文（编号36），1992，12（美国匹兹堡大学）

4、顾立帆、王永成，联想树分析法及其在无词库中文自动标引中的应用，情报学报，1992，2，No5

5、H.Borko, C.L.Bornier,(赖茂生、王知津译)，《文摘的概念与方法》，书目文献出版社，1991.6

Automatic Indexing on Subject of Chinese Text

Wang Yongcheng, Gu Xiaoming

Abstract

In this paper we describe the main idea, basic algorithm and basic structure of CSAIS 2--Automatic Indexing System of Subject of Chinese text--developed by us in 1992.

附录：

文献1标题：论我国数据库产业化的科技情报工作模式

作者：曾民族

抽出的主题关键词：数据库事业、科技情报工作、数据库化、数据库生产、微型数据库、数据库产业化、模式、数据库生产能力、数据库流通。

文献2标题：我国情报学的进展

作者：杨沛霆　王松益　赵宗仁

抽出的主题关键词：情报学、情报学者、情报学报、情报政策、科技情报学、理论、情报研究、情报学界、情报学史、基础情报学

（略）