中文全文检索技术研究*

刘开瑛
山西 太原 030006 山西大学

文摘：本文根据中文全文检索系统技术上的特点，以自然语言处理理论和技术应该是情报检索自动化的基础的观点出发，主要讨论了中文全文检索系统的评估标准，文本切词标引原则和概念语义词典等研制技术。

计算机情报检索，实际上包含有情报的检索和存贮两大部分。从检索手段来看，从传统的书名、作者检索和分类检索起步，发展到当今实用的主题词检索和主题词组、自由词检索，下一步将向自然语言检索发展。从存贮内容来说，早期以文献目录库为主。主要存贮包括标题、作者、出版单位、主题词、索引号等著录事项。由于文摘也是文献的一种标识引导，具有简法表达原文内容的能力，可以帮助读者快速掌握世界文献发展的动态，或初步了解文献的内容，所以，现行检索系统，大都基于书目数据库且以文摘为主。这类检索系统所获得的信息有极大的局限性，当用户要想获得命中记录的全面的、详细的情报信息的话，往往还须按检索到的文献索引号再到书库中去翻阅、摘引大量的文本。读者迫切希望建立全文检索系统改变这种只有文摘表达状况。随着计算机软硬件技术的进步，给建立全文数据库提供了良好的环境。全文检索系统存贮内容上既包括了文献的全文，又包括了著录事项和文摘等内容，在检索功能上读者可直接从联机终端上快速、可靠的获得文献全部信息，极大地方便了用户，受到广大用户的欢迎。当今国外全文数据库的数目不断增加，例如：美国 Dialog 情报检索系统的 300 多个数据库中，1990 年就有 100 多个库提供全文检索服务。

随着我国汉语语言信息处理技术的长足进步，特别是当前电子出版物的普及，给中文全文数据库的建造技术及其应用技术研究有了良好的支撑环境，现在武汉、北京、太原等地相继推出全文检索系统，并正在向更深层进展。早在 1986 年武汉大学开始接受国家教委文科博士点科研项目《湖北省地方全文检索系统》，此后建立了“湖北省地方志大事记”和“中国人民银放军大事记”两个全文数据库。尔后全国已推出多种方案，主要有北京文献服务处(BDS) 已着手研制的“基于自然语言处理的中文情报检索和处理系统(Chinese Iuformation Retrieval and Pro—cessing System—based on NLP，简记为 CIRPON)，该系统主要承担 BDS 的文献标题和文摘处理，经过两次大规模选库(第一次文献总数 6.5 万篇，第二次 15 万篇)实践，经标引行家评估查全率和查准率大体相当于手工标引质量。1990 年初，北京信息工程学院与人民日报社合作开发的 Biti FTRS (Full Text Retrieval System)，已实现商品化，并在人民日报社等单位使用。我们山西大学在多年从事语言处理技术基础上采用自动分词、自动分类、自动标注词性等技术，并在 1991 年推出“中文全文检索软件系统”，现已被南京金陵石化总公司精细石化文献检索系统和山西省政府办公厅和太原市政府办公厅的政务信息处理系统中采用。并在 1992 年参展了“深圳 92 计算机高技术产品展示会”。

学习国外先进技术，1987 年中国科学技术情报研究所由瑞典 PARALOG 引进了全文数据库检索管理系统(TRIP)，它是专门处理西文全文文献的。此后与对方联合，对 TRIP 进行汉

* 国家自然科学基金资助项目

化，1990 年初实现了中西文兼容的全文检索系统，达到了实用化程度，已在经济报社的 VAX 机上运行使用。

下面我们对中文全文检索系统的评估标准、文本标引和概念语义词典等研制技术提出一些看法。

一、全文检索系统的评估标准

全文数据库是将取自报刊杂志、来往公文、政府法规、资料书籍等文献的全部内容输入计算机，使之成为计算机可阅读和处理的文本。全文检索系统是在全文数据库的条件下，文献的文本内容是一种可以从各种角度进行检索、选取、组合、排序的，即在一定层次上进行知识加工的“活”体，从而各文本中的各种大小知识元——关键词，人名、地名、事件以至文本中的一个词语、一个字以至一句话得以激活。一般说来出现在全文数据库中的相应字段中（例如标题、作者、文本内容等）每一个单词，都可以作为检索的入口点。显然全文数据库检索技术同文献目录库及其受控检索有很大的差异。对全文检索系统进行正确评价，是提高其技术水平的达到实用的目标的一个重要方面。这也是当前大力开展全文数据组建中一个急待解决的问题。通常，检索系统的评价是多方面的，包括信息的价值和覆盖面；系统响应时间；输出的结果形式；操作的方便性；查全率和查准率等。其中最主要的可归结为从检索效用（Effectiveness）和检索效率（Efficiency）来考察，前者通常用查全率和查准率来衡量，后者通常用检索时间和费用来衡量。在检索效用上，全文数据库首先是对原材料的覆盖面和其信息价值的评价，这是一项基础工程。否则，查全率和查准率的讨论将无意义。全文检索的优点是可直接提供原文献，并有不存在词汇滞后问题，不损失专指性，易于实现自动化，能有效地克服组配等优点。但也存在一些缺陷，如采用自由词字面组配、误组和不正确的词间关系比以概念组配为主的受控检索多，容易降低查准率。而且由于词汇不受控制，查全率较低，这是评判一个全文检索系统的检索效用要密切关注的问题。由机型和软件配置不同，引起的检索时间和费用差异很大。但我们认为，若不考虑支撑环境，仅就检索效率来说，检索时间是其主要关键，我们应从库容量在一定规格之上，例如，库容量在 100MB 以上时响应速度应在秒级内比较合理。有时某个系统虽然从功能上说具有全文检索的内容，但由于检索时间太长也不能确认其为全文检索系统。据报导，在 VAX 机上建立的关系全文数据库 Rdb，本身具有某个字段的字符串匹配（Matching）检索功能，设数据库的数据量为 100MB 量时，经核算其检索响应时间大于 10 分种，这显然不能认为建造的全文库具有全文检索功能。国内现有几个系统在响应时间上是比较理想的。又如北京信息工程学院中文检索专家系统（ExpCIR）对容量为 100MB 数据库，平均响应速度为 10 秒左右。汉化 TRIP 软件，一般估计库容量为 50490 条记录（每条记录平均 500 个汉字）估计容量在 25MB 以上 8 个用户同时响应时间最长为 3 秒。据经济日报社统计年入库量占 60 兆字节时，检索响应时间在 20 秒以内。

二、中文全文检索系统应采用切词标引

汉语中以什么为最小检索单元，是字还是词（包括主题词）是中文全文检索中一个主要的讨论点。虽然目前国内的系统按字、按词以至混合使用的都有产品，而且都是可行的。但是我们认为从中文全文数据库检索的性能综合评价来说，应该采用切词标引也就是自动分词按词标引的技术方案才是正确的。

1、从理论上讲，情报检索是以概念为基本单元的，词是概念的基本组成部分。同西文一样，中文词也应该是基本的最小检索单位。在情报检索中，概念的提取有必要采用知识库和推理机制技术。一个名词性概念有代用、相关、属分关系，（这些关系同主题词表的三种关系可以定义

一致)；动词性概念有方式、工具、程度、时间和原因等谓词框架。通过知识库和推理理制来确认和提取概念并确认其代用、相关、属分关系或谓词框的有无。所以全文检索中以词为基本单位，进行切分标引，在处理同义词、成语、缩略语和同形多义现象，以至推出蕴含概念有其明显的理论基础。北京语言学院的 Exp CIP 系统，采用切词标引技术，在建造领域知识库和策略规则库中充分显示其具有坚实的理论基础。

2、从应用角度来看，切问标引是自动分词技术的应用。自动分词是汉语语言信息处理技术中一个关键技术，研制能适合各自专业领域特点的一个切分正确率高和切分速度快的自动分词软件是当前机器翻译、人机接口和自然语言理解以及情报检索等重要的课题领域中十分迫切需要的。所以国家六五、七五科技攻关项目均列为重点，进行研究。历经十年，成绩卓著，已提出多种实用的方案。当今情报检索技术的发展采用基于自然语言处理技术已成为必然趋势，从情报检索自动化正向自然语言检索和知识提取的方向发展的时候，用汉语语言处理的自动分词、自动标注词性、自动分类等智能技术，逐步完善切词标引技术，才能保证同其他更高级检索技术的接轨。

3、采用按单汉字建立位置索引，具有实现方法简单、查全率高以及在处理新词(如人名、地名、机关名、专业名)和自然汉字串具有与其他方法不可比拟的优点。但是作为此采用按单汉字杯引其缺点是十分明显的，首先是随着数据库容量的增加，标引量急骤上升，耗费时空太大。其次按字全文检索虽然实现方法简单，查全率高，但速度慢，检索效率低，只有通过对检索词语的后控处理才能达到提高检索效率的目的。需要指出，自动分词系统研究中对新词处理技术已作为重点，有的已经提出解决方案，如 EXP CIP 系统已提出了一种基于整篇文献词串频度统计与构词基本规律相结合的新词识别算法。当前国内正在研究的分词新方法中如基于知识的分词方法、基于人工神经网络的分词原理等都在新词切分和歧义切分处理上提出了一些解决方案。毫无疑问，随着智能技术的应用，自动分词系统最终可以圆满解决这类问题的。

三、概念语义词典是提高情报检索效用的极其重要的手段。

情报检索是以概念为基本单元的系统，词(或词组)是概念的基本组成部分，所以词应该是基本的最小检索单位。我们现行的自动分词系统，基本上都是按照《词典匹配＋岐义处理＋补充新词》来处理的，其核心是以词典匹配技术为主体，这样对汉语文本的切词标引，就以很难准确地提取文献的隐含概念。因为，同一个词可能有不同的语义和表达法，相同的事情即相同的概念可用不同的词来表达。例如：同一概念有许多不同叫法(如鸦片、烟土、洋烟、烟毒、毒品等)；实名和指称、指代有等价关系(他、该年、该地、同年、次年、该厂等)；实名与缩写词(人大、政协、日伪军等)；简称与全称(晋/山西、京津/北京、天津等)；地名今昔不同(北京/北平、沈阳/奉天等)。对这些情况共性指称指代外，必须根据系统处理的领域不同，认真搞清楚同义词、类义词、反义词、关联词的关系，实际就是通过概念及其语义关系的集合组成概念语义词典。这方面的工作我国几个实用系统，已取得成效。北京文献服务处的 CIRPON 系统中，为了构造同义词典和相关词典问题，设计了相关标引模块. 根据已经由标引专家标引了主题词，著录了范畴号的国防科技文献语料为基础统计出每个主题词与每个自由词同时出现在同一篇文献的频率值，认为具有高频度值的词相对集中，其自由词与该指定的主题词将形成同义、近义和相关关系。他们引入相似优先比的概念并通过运算，成功地构造了同义词典和相关词典。同义词典建立之后，使检索人员不考虑所有表达同一概念的词，系统可以根据检索入口词自动扩展。借助相关词典，达到编检、扩检等功能，有助于提取隐含概念。这种以文献语料为和统计模型相结合的语言分析方法，就是七十年代兴起的语料库语言学，它摆脱了某些受限的子语言中处理语言

问题的困难，在具有丰富的、真实的基础语言资料的计算机资源上，采用统计模型的方法，来分析和处理语言现象。这是一项有光明前景的技术路线。北京语言工程学院开发的 BitiFTRS 系统中，根据现行的主题词表的基础上成功地重拳设计和构造了适合新闻语料库的概念语义词典。它和一般词典的区别在于概念本身并没有定义，而是概念之间的四种语义关系来明确其含义。这四种语义关系为①同义关系（也称用代关系）②属分关系（即上下位关系）③相关关系（也称参关系）④词组/成分关系。如“情报检索”由子概念“情报”和“检索”组成，它们之间为词组/成分关系。这种技术实际是就是中文全文检索中专家系统技术的应用。这也是一个很有乐观前景的方法。

在受控检索中，根据这类关系整理和编写的词典叫主题词表（THE－ASURUS），编辑主题词表是一项相当艰巨的工作。我国现有的《汉字主题词表》共 30 卷，10 分册，108568 个主题词，从 1975 年 7 月开始，组织了 500 多个单位 1370 个人，耗时四年编辑出版。主题词表对整个情报检索系统有极其重要的作用。主题词表的编制及确定关键词采用的聚类（CLUSTER-ING）主法，对全文检索中概念语义词典研制中也是一个有参考价值的、有效的方法，今后应着重研究。

情报检索自动化进程中，以自然语言处理技术为基础是必由之路，汉语自然语言处理不编句子分析方法有多少主要仍处于词法，语法理论和应用研究阶段，语义知识刚刚起步，语用知识尚未涉及，所以对全文检索技术尚无坚实的基础，许多问题尚待继续努力研究。

参考文献

1、长尾真（日本），《语言识别》，中国计算机报，1991 年 8 月 6 日技术专题版

2、陈光祚，关于全文数据库的结构、检索功能及其文本标引的探讨，知识工程，1991 年 2 期

3、刘开瑛、郭炳炎，《自然语言处理》，科学出版社 1991 年出版

4、曾民族、陈豫，基于自然语言处理的中文情报检索和处理系统，计算机世界，1993 年 6 月 2 日专题综述

5、施水才、苏东庄，中文全文检索专家系统，计算机世界，1993 年 6 月 2 日，专题综述

6、刘启业，自动分词、自动分类和全文检索技术，计算机世界，1993 年 6 月 2 日，专题综述

7、施水才、苏东庄，ExpCIR —— 中文全文检索专家系统，Pro－ceeding 1992 International Conference On Chinese Information Processing（2） P. 163

Techniques Research For Chinese Full Text Retrieval System

Liu Kaiying
Shanxi University
Zip Code 030006 Taiyuan Shanxi

ABSTRACT:

In accordance with the technical features of chinese fulltext retrieval system, and from the point of view thatnatural language processing theory and technique should bethe foundat—ion of information retrieval automation , this article mainly discusses the evaluation standards of Chinese full text retri—eval system, the principles of automatic segmentation of freewords, conceptual semantic dictionary and other development te—chniques.