汉语分词技术综述*

熊回香　夏立新
华中师范大学信息管理系　武汉　430079

〔摘要〕 首先介绍汉语自动分词技术及基于词索引的中文全文检索技术，接着分别从文献自动标引、文摘自动生成、文本自动分类、文本信息过滤、自然语言检索接口和智能检索等方面详细地阐述汉语自动分词技术在中文全文检索中的应用，并对目前汉语自动分词技术存在的局限性进行分析，提出发展思路，最后对汉语自动分词技术在中文全文检索中的应用前景进行预测。
〔关键词〕汉语自动分词　中文全文检索　文献自动标引　自然语言检索
〔分类号〕G354

The Review of Chinese Automatic Word Segmentation Technology

Xiong Huixiang　Xia Lixin
Department of Information Management, Central China Normal University ,Wuhan 430079

〔Abstract〕Fristly,this paper introduces Chinese automatic technology and Chinese full-text retrieval technology based on word index. Secondly, it detailedly expounds the applications of Chinese automatic word segmentation technology in Chinese full-text retrieval from several aspects, such as document auto-index , summarization auto-generation , text auto-classification,text information filtration, natural language retrieval interface and intelligent retrieval etc., and also analyzes the limitation of Chinese automatic word segmentation technology at present , comes up with developing method. Finally, it forecasts application prospect of Chinese automatic word segmentation technology in Chinese full-text retrieval.
〔Keywords〕Chinese automatic word segmentation　Chinese full-text retrieval　document auto-index　natural language retrieval

1　汉语自动分词技术

词是最小的能够独立活动的有意义的语言成分，自然语言的处理必须以词为单位，然而，汉语文本中词与词之间没有明确的分隔标记，而是连续的汉字串，因此理解和处理汉语的首要任务就是把连续的汉字串分割成词的序列，即自动分词。

近 20 年来，国内语言学界、人工智能领域和情报检索界的学者们对汉语自动分词这一研究领域给予了极大的关注，提出了许多解决汉语自动分词的方法，归纳起来主要有四种类型：基于词典的分词方法、基于统计的分词方法、基于理解的分词方法和基于人工智能的分词方法。这些分词方法各有其特点，分别代表着不同的发展方向。其中基于词典的分词方法由于其算法成熟，易于实现，是目前普遍使用的切分方法；基于统计的分词方法由于有良好的歧义切分能力和低频词识别能力，受到越来越多的研究人员的重视，发展较快，但实际使用中，单独使用的较少，一般都与基于词典匹配的分词方法结合使用；基于理解的分词方法是在分词的同时进行句法、语义的分析，利用句法信息和语义信息来处理歧义现象，因而具有良好的歧义切分能力，但因为要对语言自身信息进行更多的处理，因而加大了实现的难度[1]；基于人工智能的分词方法是目前理论上最为理想的分词方法，但是该类分词方法的研究还处于初级阶段，并且由于汉语自然语言复杂灵活，知识表示困难，所以对于这类分词技术还需要进行更深入和全面的研究。

2　基于词索引的中文全文检索

全文检索是一种面向全文和提供全文的检索技术，其核心是将文档中所有基本元素的出现信息记录到索引库中，在中文全文检索系统中，这些基本元素可以是单个汉字，也可以是词，因此存在两种基本的索引结构，即基于字的索引和基于词的索引。

基于词索引的中文全文检索系统首先必须进行汉语自动分词，其次是把文档中出现的所有有意义的词建立倒排索引，

* 本文系国家社会科学基金项目“基于中文 XML 文档的全文检索研究”（项目编号：04CTQ005）研究成果之一。
收稿时间：2007-10-08　修回日期：2007-10-24　本文起止页码：81-84　本文责任编辑：王传清

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

检索时将用户输入的检索要求按照一定的匹配机制与词索引库中的信息进行匹配，最后将检索结果返回给用户。

建立词索引库时，需要扫描整个文档，并利用自动分词技术对文档中的汉字串进行切分，对切分出来的每一个有效词，计算其在文档中出现的位置和频率，同时将该位置信息和频率的值以及所属文档号加入到词索引库中，建立基于词的倒排索引。

典型的基于词的倒排索引结构(见图1)包含两部分:①中文词组成向量（称之为词汇表），包含词的基本信息和词索引在索引文件中的偏移量 ；②对于词汇表中的每一个词，都有一个它出现过的文档列表，包含了出现文档编号和在此文档中该词的词频以及出现位置序列[2],也可以在词索引中记下段落号、句子号等。

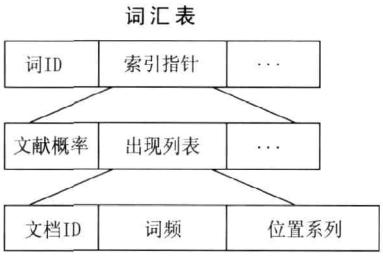

图1 词的倒排索引结构

词索引库建立之后，就可根据一定的检索模型来处理用户的检索请求，常用的信息模型有：布尔逻辑模型、向量空间模型以及概率模型等，其中布尔逻辑模型是目前中文全文检索系统采用最多的一种模型。

3 汉语自动分词技术在中文全文检索中的应用

3.1 文献自动标引

标引是对文献进行分析，提取关键信息，产生对文献的描述，它是全文检索实现的主要支持。目前，对网上日益丰富的信息资源进行人工标引变得越来越困难，因而利用计算机进行文献自动标引的需求也越来越迫切。要实现计算机自动标引，其重要的前提是汉语自动分词，只有正确地把具有检索意义的汉语词切分出来，才能提取足以描述文档内容的关键信息，并在此基础上进行文献的自动标引。

常见的文献自动标引的方式有：① 全文标引：将整篇文章中出现的所有具有检索意义的汉语词切分出来，统计词频并标注其位置信息，存入全文数据库；②主题词自动抽取：根据文献所论述和研究的具体对象和问题，赋予文献以恰当的主题词，其首要的工作是对需要处理的文献进行自动分词处理，去掉停用词，并计算词频和权值，然后进行排序，选出系统规定数量的词汇作为主题词[3]。

3.2 文摘自动生成

文摘自动生成是把文档内容从逻辑和语义上进行分析，缩写成有限的可读摘要，标志文章的主题内容，从而有助于用户快速评价检索结果的相关程度。文摘自动生成常用的是基于统计的方法，即首先对全文进行自动分词，然后计算文章中各个词出现的频率和权重，并按照某种准则确定出关键词，将关键词所在的语句抽取出来，再依据各种句子权重指标计算句子综合权重，选出一组最能代表文献主题内容的句子，并对句子进行排序作为文摘句，最后生成文摘[4]。

3.3 文本自动分类

文本自动分类的任务是基于内容将大量的用自然语言写成的文本按照一定的主题类别自动进行分类，它能将信息文档分类并自动将其归入事先给定的最接近的类中，从而使用户能够更加准确地查找所需的信息。

为完成分类任务，需要对文本进行必要的表示和预处理，在此基础上再运用分类算法对其进行分类。目前在信息处理方向上，文本的表示主要采用向量空间模型。向量空间模型的基本思想是以向量来表示文本：(W_1，W_2，…，W_n)，其中W_i为第i个特征项的权重，一般选择词作为特征项。因此，要将文本表示为向量空间中的一个向量，就首先要将文本分词以获取文本中所有的词，继而将文本用词频来表示[5]，形成表示文本的特征向量，用于以后的文本分类。

3.4 文本信息过滤

信息过滤是根据用户的信息需求，运用一定的标准和技术，从大量的动态信息流中将与用户无关的信息滤掉，把满足用户需求的信息提供给用户，从而提高用户获取信息的效率。信息过滤的首要工作是：对采集到的Web页面进行预处理，将HTML页面里的文本提取出来，然后使用中文分词技术将Web文本切分成单个的中文词语并进行词频统计，根据统计的结果从得到的中文词向量中提取出能够表达出该文本主题的特征向量[6]，这就是特征信息提取，它是信息过滤的基础工作。

3.5 自然语言检索接口

检索接口是连接用户和全文检索系统之间的桥梁，没有一个有效的用户接口，系统的功能就难以充分发挥。自然语言检索接口允许用户以自然语言的方式和机器交互，是一种人性化的智能接口，它的主要功能是分析用户用自然语句输入的查询请求，“理解”人们检索的真正意图。其工作原理是：首先对用户输入的查询语句进行分词，识别每个词的词性，提取关键词，然后从逻辑上进行语法语义分析，生成中间形式的表现形式，再经过翻译模块翻译成目标数据库查询语言表示的语句，最后对文本进行语义上的概念匹配。在此过程中，分词的准确性对查询效率的影响较大。

3.6 智能搜索

智能搜索是结合了人工智能技术的新一代搜索技术，它将信息检索从目前基于关键词层面提高到基于知识（或概念）层面，对知识有一定的理解与处理能力。它的主要任务是对

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

信息进行智能处理和智能理解用户的检索需求。而汉语自动分词技术正是使搜索具备“智力”的前提，它是自然语言理解、语法语义分析、概念匹配、机器翻译等技术的基础，将这些相关技术应用到全文检索系统中，可使检索系统更加深入细致地获取用户需求，从关键词的选择、检索范围的确定到检索结果的精确，系统都能帮助用户从知识的海洋中及时准确地获取所需信息。

4　汉语自动分词技术的局限及发展

经过十几年的研究，汉语自动分词技术取得了令人瞩目的成果，出现了一些实用的自动分词系统，如：北京航空航天大学的CDWS分词系统、清华大学的SEG分词系统和SEGTAG分词系统等，这些系统在分词的精确度（精度达到99%以上）和分词速度（速度达到千字/S）方面都具有相当的水平[7]，并在一些中文全文检索系统中得到了应用，如百度、北大天网、北京易用宝公司的TRS系统等都使用了汉语自动分词技术。但与此同时又应该看到目前汉语自动分词的技术还在一定的局限性，需要从以下几个方面加以进一步的研究。

4.1　分词算法

分词算法是汉语自动分词技术中的重点和难点，它是影响切分效率的关键因素，切分效率的衡量指标是分词速度和分词精度。现有的分词算法基本上都是基于规则和词典的分词方法，它们都必须在分词速度和精度之间做出选择。要提高速度，就要适当放弃精度的追求，缩减词典，减少匹配次数；而要提高切分精度，就得舍弃速度，无限扩充词典，匹配次数也会无限增加[8]。对此，目前还没有找到有效的破解方法。

分词的精度常常直接影响到对全文检索结果的相关度排序，分词的速度也会严重影响检索系统内容更新的速度，因此对于全文检索系统来说分词的精度和速度两者都需要达到很高的要求。传统的汉语自动分词要获得新的突破，只有在现有的切分算法的基础上，充分吸收自然语言处理、人工智能和专家系统的最新研究成果，着重从汉语句法和语义入手，并加强对汉字串统计性质的研究，将基于知识和推理的深层方法与基于统计等“浅层”方法结合起来，对汉语分词算法进行更加深入的研究，这是今后汉语自动分词努力的重要方向之一。

4.2　分词词典与分词规范

分词词典是汉语自动分词过程中的重要工具之一，目前，互联网上信息膨胀，各种概念说法繁多，如何使词典收录的词粒度适中，提高信息检索的查全率和查准率，是词典编制面临的一大挑战。另外，分词词典的组织方式、通用的核心词典和各个领域的专业词典的编制和更新也是未来需要进一步关注的问题。与此同时，词与词素、短语之间的概念模糊，给分词词典的规范化造成了困难。虽然目前已有《信息处理用现代汉语分词规范》指导分词，但该规范还不成熟，有很多地方有待商榷，需要改进[9]，需要计算机科学家和汉语言学家共同努力。

4.3　岐义消除

汉语词与词之间没有没有任何区分标志，加上汉语词理解的多义性、复杂性，因而岐义消除是自动分词过程中的一大难题，切分岐义的存在将严重影响着分词系统的切分精度，而目前的分词系统大多在消除岐义方面不理想，因而也就直接影响到中文检索的查准率和查全率。未来在岐义消除方面的研究除了完善分词词典以外，还需要深入细致地分析各种岐义产生的原因，针对不同类型的岐义提出不同的消岐方法；同时深入研究汉语的构词规则和词法规则，增强歧义判别的能力。

4.4　未登录词的识别

未登录词即是指未包含在分词词表中的词，包括各类专名(人名、地名、企业字号和商标号等)、某些术语、缩略语和新词等，由于专用术语繁多，新名词、新概念层出不穷，这些词一般很难全部收录到词典中，但这些词往往在一定时期内呈现较高的检索概率。因而未登录词识别也是中文信息处理中的一个难点，在大规模中文文本的自动分词中，未被识别的新词是造成分词错误的一个重要原因。

目前，未登录词辨识的研究基础还比较薄弱，同时拥有多种未登录词辨识能力的系统尚不多见，因此未登录词的综合识别问题还没有引起足够的重视，现行的识别方法主要是基于分解与动态规划策略的识别方法和基于语料学习的检测方法[10]，这些方法的识别能力还非常有限，未来的发展方向主要是探究新词自身的构成规律和特点，充分利用语料库等网上语言信息资源，提出更有效的识别新词的方法。

4.5　汉语语料库的建设和应用

汉语语料库对中文全文检索的辅助是必不可少的，目前，语料库对于信息检索的辅助作用还没有得到充分的发挥，未来对汉语料库的工作主要包括两方面：①充分利用现有的语料库资源，如国家语言文字工作委员会的“国家现代汉语语料库”，它是一个大型的国家级的、通用语料库。该语料库2005年通过鉴定，其中包有丰富的语料资源，这些语料信息使计算机能从中学到汉语的构成规律，也就增强了计算机自动识别的能力，这对汉语自动分词的切分精度有非常大的帮助。②进一步进行语料库的建设，尤其是大规模真实语料库的建设更为需要。

4.6　词索引数据库的结构

词索引数据库是全文检索系统实现的基础，由于全文检索系统通常处理的数据量很大，经过处理生成的索引数据也很大，这对系统的存储容量和检索速度都带来了极大的挑战，因此，未来还需要继续对词索引数据库记录内容的确定、数据库的逻辑结构和存储结构、数据库的压缩存储等方面进行

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

进一步的研究。

5 汉语自动分词技术在中文全文检索中的应用前景

汉语自动分词技术的每一次突破都会使中文全文检索的效率得到很大的提高，未来的中文全文检索技术必定是以提高其系统的查准率、查全率和查询速度为目标，因此，汉语自动分词技术在中文全文检索中的应用将会在以下方面得到进一步拓展。

5.1 文献信息的深度处理

信息搜索的真正对象是标引的结果，因而高性能的检索需要有效的索引支持。目前，中文信息处理的深度还不十分理想，随着汉语词的切分精度和自然语言处理水平的提高，未来的标引是按照一定的格式，建立词法、句法/语义层次的深度标引，与此同时，文摘自动生成和文本自动分类的准确性将会得到大幅度的提高，这些对中文全文检索的速度和效率都会产生很大的影响。

5.2 匹配机制的进一步优化

信息检索的目的是在信息收藏中查找包含用户所需的信息内容的文档，当前的全文检索系统采用自由词匹配，其优点是灵活，缺点是有大量的误检和漏检。未来的信息系统应当是概念匹配，即系统自动抽取能够描述文献内容的概念，用文中的关键词或与之相应的主题词加以标引；用户在系统的辅助下选用合适的词语表达自己的信息需求；在此基础上两者之间执行概念匹配[11]，匹配在语义上相同、相近、相包含的词语，使检索更逼近人的智能程度，以减少误检和漏检。

5.3 自然语言检索的智能化

自然语言应用于全文检索主要体现在两方面：一是用自然语言标引全文；二是向用户提供自然语言检索接口。目前，在自然语言标引方面多数限于词形或词汇层次，即使在词汇层次，也没有很好的解决由于词的同义、近义现象而需要扩展检索的问题，不能从语义上理解，因此漏检和误检问题非常严重；在自然语言检索接口方面，目前大多数中文全文检索系统在这方面的功能比较缺乏。要改变这一现状，一方面，必须将自然语言与受控语言进行有机的融合，融合的手段主要是通过有效的词汇控制技术，如停用词表、同义\近义词表、入口词表、后控词表等；另一方面，充分利用自然语言处理的最新研究成果，使自然语言检索具有更高的智能，其智能化主要表现在：①从内容上真正的理解文献所论述的主题；②使用适当的知识表示方法来充分体现各主题概念和标识之间的分、属、交叉等复杂关系；③能准确分析用户的自然语言提问，并通过人机交互推断出其真正需求[12]。

6 结束语

汉语分词是中文信息处理的基础，也是中文全文检索中的“瓶颈”问题，因而，中文全文检索系统的检索效率的提高，依赖于汉语自动分词技术的发展；依赖于对汉语的语词结构、句结构、语义等语言知识的深入系统的研究；依赖于对语言与思维的本质的揭示；同时，在很大程度上还寄希望于人工智能技术的突破。相信在不久的将来，随着相关领域知识的研究越来越成熟，未来的中文全文检索将最终达到真正的语义、语用、语境层次的智能信息检索，检索结果更加全面和准确。

参考文献：

[1] 熊回香.全文检索中的汉语自动分词及其岐义处理.中国图书馆学报,2005,31(5):54-57.
[2] 韩 升,刘广志.全文检索系统的数据预处理研究.计算机技术与发展,2006,16(3):208-210.
[3] 吴春玉.中文全文检索系统中实现主题词标引思路.情报杂志,2005,24(1):115-119.
[4] 董建设,任 丽,周燕玲.中文自动文摘在搜索引擎中的应用.情报科学,2006,24(2):267-269,309.
[5] 周 瑛,刘政怡.覆盖算法在文本分类中的应用.情报理论与实践,2006,29(1):115-117.
[6] 费洪晓,巩艳玲,黎 成.基于Agent的个性化信息过滤系统的设计与实现.计算机技术与发展,2006(12):1-3,6.
[7] 刘 迁,贾惠波.中文信息处理中自动分词技术的研究与展望.计算机工程与应用,2006(3):175-177,182.
[8] 文庭孝.汉语自动分词研究进展.图书与情报,2005(5):54-63.
[9] 孙 巍.一种面向中文信息检索的汉语自动分词方法.现代图书情报技术,2006(7):33-36.
[10] 周文帅,冯 速.汉语分词技术研究现状与应用展望.山西师范大学学报(自然科学版),2006,20(1):25-29.
[11] 吴慰慈.网络环境下信息存储与检索技术的发展.四川图书馆学报,2003(1):3-6.
[12] 张世红,胡佳佳,宋继华,等.网络环境下的自然语言检索.医学情报工作,2005(6):434-436.

〔作者简介〕 熊回香,女,1966年生,副教授,发表论文20余篇。
夏立新,男,1969年生,教授,发表论文数十篇。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net