信息检索技术

基于WWW的全文检索系统设计与实现

李广建 黄永文

（中国科学院文献情报中心 北京 100080）（北京师范大学信息技术与管理学系 北京 100875）

【摘要】 在阐述全文检索系统的涵义、特点及发展现状的基础上，介绍了“基于WWW的教育信息全文检索系统”的概况，并探析了实现该全文检索系统的关键技术。

【关键词】 WWW 教育信息 全文检索

Design and Implement of Full—Text Retrieval System Based on WWW

Li Guangjian

(The Documentation and Information Center of CAS, Beijing)

Huang Yongwen

(Information Technology and Management Department of Beijing Normal University, Beijing)

【Abstract】 After discussing the definition, characteristics, development and current condition of the full—text retrieval system, the article introduces the general situation of “Educational Information Full—Text Retrieval System Based on WWW”, and then analyzes the key components with which the system is implemented.

【Keywords】 WWW Educational information Full text retrieval

1 概 述

全文检索系统是指用户可以使用自然语言对全文进行检索并能获取原文的系统。它主要包括三层涵义：①允许用户使用自然语言进行检索；②允许用户从全文的任意篇、章、节、句、词组、词、字查找；③能直接提供全文。

与书目检索系统、事实或数值数据检索系统等相比，全文检索系统具有以下特点。

①用户检索负担轻

全文检索系统主要采用原文献中的文字即自然语言作为检索的依据，用户无需学习选词规则，没有要掌握人工语言的额外负担。

②检索彻底、详尽，查全率、查准率高

全文检索系统允许对文献中的任何章节、段落、句子、词或字进行检索，提供的标引得深度达到了顶点，从而使得检索极为彻底，不受标引的限制。

③检索结果具有直接性、可靠性和原始性

全文检索系统提供的是原始文献本身，它们未经过任何加工，具有显著的直接性、可靠性、原始性和客观性。

收稿日期：1999—11—10

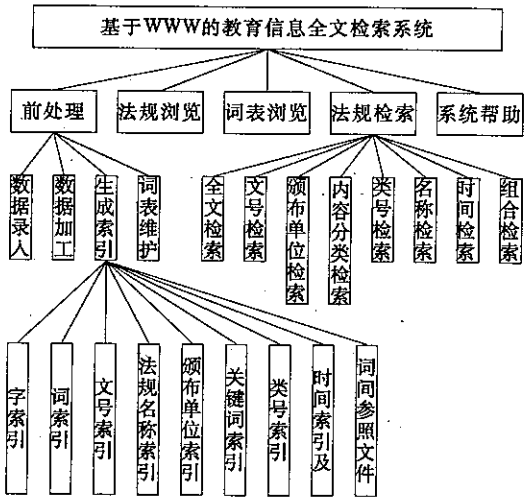
全文检索是本世纪50年代末出现的一种新型信息检索技术。1959年，美国匹兹堡大学卫生法律中心研制出世界第一个全文检索系统——法律情报检索系统，开创了全文检索的先河。此后30多年，尤其是80年代以来，全文检索得到了迅猛发展，已成为国外文字型信息检索的主流，主要表现在：全文数据库数量迅速增长，类型和范围不断扩大；全文检索技术日益完善，出现了集合、位逻辑等算法；出现了图像、文字、声音一体化的趋势。

与国外相比，我国全文检索系统的研究起步较晚，始于70年代末80年代初。1979年，由武汉大学计算机系和中文系联合研制的《骆驼祥子》全文检索系统是我国最早的这类系统。近年来，经过国内众多学者、专家的共同努力，全文检索系统的研究卓有成效，已建成一定数量和规模的全文检索系统，如清华信息系统公司的CAJR、北京南辰电脑公司的FTR7.0、北京海文电子信息系统公司的QUICK等。同时，随着网络的发展，我国也出现了网络版全文检索系统，如中国科技信息研究所的万方数据系统、中国期刊网专题全文数据库全文检索系统、北京大学开发的中国法律检索系统、上海交通大学研制的网上音乐数据库全文检索试验系统等。

2 基于WWW的教育信息全文检索系统的实现

“基于WWW的教育信息全文检索系统”是“九五”国家

重点科技攻关项目子专题"高等教育教学管理信息系统"的一个组成部分，目前已实现了对教育法规的全文检索。系统模块如下：



前处理模块为后台作业，主要完成数据录入、数据加工、索引生成和词表维护等功能。

法规浏览模块的主要功能是分库浏览系统中收入的所有法规文献，如果用户不了解系统的检索途径或感到无从下手，该浏览检索功能具有很好的指导作用。法规检索模块提供全文、法规名称、颁布单位、内容分类等7个检索入口，并支持7种途径的组合检索，如"法规名称＋颁布单位＋全文"等；

词表浏览模块按音序或字数的升、降序显示现有的词表，可指定显示词汇的字数，用户也可由此入手进行全文检索。这一功能的加入可为用户构造检索词提供帮助；

系统帮助模块详细介绍系统中每一个命令的使用，供用户参考。

目前本系统收入教育类法律法规信息，内容涉及教育改革、各级教育、教育仪器设备诸多方面。系统根据法规文献的颁布和实施日期，将所收入的法规分入四个库——"教育法规文献库(1945－1985)"、"教育法规文献库(1986－1990)"、"教育法规文献库(1991－1995)"以及"教育法规文献库(1996－至今)"，实行分库检索。系统采用浏览器/服务器结构，服务器平台为Windows NT＋IIS4.0，客户端平台为Windows 9X系列或NT Workstation、NT Server(便于与服务器NT Server协调工作)，浏览器为IE或Netscape。

本检索系统已于1999年10月1日开始在"中国高等教育教学信息网"(http://www.crct.edu.cn)上试运行，所有信息均可免费得到全文。

3 基于WWW的教育信息全文检索系统"的开发技术

3.1 采用字、词结合的系统模式扩展系统的检索结果，提高系统的查全率

汉字全文检索系统现有两种模式：单字无标引系统和全文后控系统。单字无标引系统避开了词的切分问题，把研究重点放在单字一级，实现起来比较简单，但系统查全率不高。例如，在查找有关计算机方面的文献时，输入检索词"计算机"，系统则只能检索出含这三个字的文献，而那些出现"微机"、"大型机"、"小型机"等词汇的文献就会漏检，单字无标引系统的这种缺点主要是由于缺乏控制而引起的。为了弥补这些缺点，改善检索性能，出现了全文后控系统。全文后控系统的基本思想是采用后控词表对系统加以控制，在后控词表中反映出各词的同义词、相关词等，从而提高查全率，但由于后控系统过分依赖后控词表，在一定程度上削弱了全文检索的优势。

本系统把上述两种模式有机地结合起来，吸取二者的优点，在保留单字无标引模式的基础上，借鉴了后控词表的思想，抽取《教育主题词表》中涉及高等教育的片段，生成了一个功能与后控词表类似的词库，在词库中反映出词汇间的代、用、属、分、参关系，并根据字索引和词库生成词索引文件。

这样，系统在检索时就有两条检索路线：如果用户输入的是字，则直接查找字索引文件，若找到，表明检索成功，反之，检索失败；如果用户输入的是词，则先查找词索引文件，若找到，表明用户使用规范词进行检索，此时判断该词是否具有参照词汇，若存在，则查找词间参照关系文件，取出其参照文献号和直接命中文献一并输出；反之，则只输出直接命中文献；如果在词索引文件中没有找到该词，表明用户输入的是自由词，此时通过查找字索引进行字组配，组配成功，则输出命中结果，反之检索失败。

这样，就使得系统具有自动扩检和缩检功能，系统可在用检索词查找文献的同时，根据该词的用、代、属、分、参等关系，进行扩检和缩检。例如，在本系统用"高等教育"一词进行全文检索的结果是：直接命中302篇文献，下位词"本科、学历教育、专科"命中81篇文献和相关词"高等学校"命中148篇文献。

3.2 字组配技术

所谓"字组配"是指通过组成词的每个字的位置信息，组配出词的位置信息的操作。

现有的字组配算法基本上可概括为两类：一类是单字匹配算法，另一类是集合算法。

单字匹配算法只查找首字的地址集合，并不查找检索词(字)中的每个单字的地址信息，然后根据首字地址信息依次判断首字出现位置的相临处是否存在整个检索词。当检索词较长时，这种算法会显示出一定的优势，但当首字地址信息较长时，此算法则会耗费大量的时间，同时由于此算法每次都要深入文献内部比较，故需将文献调入内存中，在原文献库较大的情况下，不断定位文献并将其读入内存会占用许多时间，从而影响运算速度。

集合算法是在用单汉字检索后对其结果集合进行交运算，这种算法的关键在于如何进行交运算以提高速度。本系统采用了集合运算算法，但利用了字索引文件的有序性来对

算法进行优化。

设字A在文献中出现了n次，位置分别为$a_1a_2...a_n$，字B在文献中出现了m次，位置分别为$b_1b_2..b_m$。用集合A、B分别存储上述位置信息，且集合A和B内的元素均已排序，用集合AB来存储组配结果，A_i、B_i、AB_i分别表示集合A、B、AB中的第i个元素，组配算法如下：

```
Procedure merge(A,B)
begin
    i:=1;
    k:=1;
    j:=1;
    l:=max(m,n)
    while (i<=l or j<=l)
    begin
    switch Aj
        case<Bi:
            j:=j+1;
        case=Bi:
            begin
              ABk:=Aj;
              i:=i+1;
              j:=j+1;
              k:=k+1;
            end
        case>Bi:
              i:=i+1;
    end
end
```

上述算法充分利用了系统前处理过程中地址信息自然形成的有序化排列，通过两组指针的同步移动，提高了运算的速度。实践证明，这种匹配算法的响应速度是很快的。

3.3 词加权技术

本系统运用词加权技术来提高查准率。词加权就是根据词在文献中的重要程度来判断词与文献的相关度，并据此决定文献的输出位置，使系统提供的检索结果本身就暗含了一种顺序，从而提高了文献内容与检索需求的相关度，更好地满足用户需求。

在全文数据中，词可能出现在标题中、摘要、章节标题及正文各段落中，如果知道词在全文中的位置，又能根据全文各部分表述文献内容的相对重要性为其设置合适的权值，就可以在全文检索中计算文献与词的相关度。为实现这一点，本系统主要作了以下几个方面的工作：

①识别文献的语句结构

目前已有若干国际标准对全文文本结构进行规范化描述，规定了专门的标识符号来标记文献章节段落和标题等，有的更为详细地规定了文献逻辑结构和物理表达方式，包括章节段落、标题、附录、图表等。本系统结合这些标准根据所收录文献的实际情况设计了一套标记格式，分别对法规名称、颁布单位、内容分类、文号、正文等进行标记。

②标记词的位置

由于本系统中词的位置信息是通过“字组配”算法得来的，所以仅对字的位置进行标记。本系统按下列规则自动标记字的位置：如果字出现于篇名中，则“文中位置”项置为“T(Title)”；如果字出现于内容分类中，则“文中位置”项置为“S(Subject)”；如果字出现于正文中，则“文中位置”项置为“C(Content)”；如果字出现于其它位置，则“文中位置”项置为“N”。

③设置权值

本系统的权值设置原则为：凡文中位置为“T”或“e”的词汇，其权值设为1；文中位置为“C”或“N”的词汇，则根据词汇的数量与文献内容的关系规律将其在文献中出现的次数(即词频)与文献长度的比值作为权值。

通过以上步骤完成词的加权后，在检索时即可按照词权值的高低顺序输出命中文献，从而使最相关的文献排在最前面，以提高检索结果与用户需求的符合程度。

除上述几个方面以外，本系统还具有非文本信息的处理能力、网络通讯能力，这均属于已知技术，本文不再赘述。

4 结束语

“基于WWW的教育信息全文检索系统”是我们在这个领域的初步尝试，目前，该系统正在进行改进，对现有功能进行完善并增加一些新的功能，例如，拟增加分页处理以提高网络传输速度、改进词表质量，增加逻辑检索等等，使之更趋于完善。

参考文献

1 邓玲怡. 基于Internet的教育信息全文检索系统. 北京师范大学硕士学位论文，1999.5

2 濮德敏. 国内中文搜索引擎的检索技巧. 中国信息导报，1999，(7)

3 周全明. 全文检索系统后控关键词采集政策研究. 情报理论与实践，1996，(4)

4 顾耀芳. 综述全文检索系统. 现代图书情报技术，1992，(1)

5 http://www.crct.edu.cn

6 http://compass.net.edu.cn:8080

7 http://pccms.pku.edu.cn:8000/gbindex.htm

8 http://www.tonghua.com.cn/

9 http://search.gznet.edu.cn/search/index.html

10 http://www.gbchinese.com.cn

11 http://www.sohoo.com.cn

12 http://infonavi.cei.gov.cn