

RMS系统架构与情报检索系统的功能需求研究^{*}

□ 吴广印 / 中国科学技术信息研究所 北京 100038

摘要: 文章对RMS的发展历史、系统架构和设计理念进行了详细介绍,同时对情报检索系统的功能需求进行研究和分析,最后对情报检索系统的进一步发展和应用进行了展望。

关键词: RMS, 情报检索, 系统架构, 索引技术, 变长存储管理

DOI: 10.3772/j.issn.1673—2286.2013.06.006

1 前言

RMS非结构化数据库管理系统是上世纪90年代初期针对图书、情报的信息服务而开发的专业情报检索软件。20多年来,随着计算机处理技术和网络通讯技术的不断发展,该系统经历了单机、网络、HTTP Web、Web Service、RMS Cloud等版本发展,跨越了本地服务、在线服务、互联网服务、云服务等历史阶段;从一个单一应用系统发展到一个可以提供系统管理、COM组件服务及各种方式API的非结构化数据库管理支撑平台;从支持单机情报检索应用到分布式企业级情报检索应用。系统应用涉及图书、情报、政府、企业、科研管理部门,从公安系统的综合查询、安全的外部基础信息管理到万方数据的资源服务系统,都在使用RMS作为底层数据库管理和搜索引擎底层支持。作为RMS系统的总体设计师和系统开发者,作者利用本文首次介绍RMS的系统架构及主要功能设计,目的是让读者了解情报检索的特殊需求及解决方法。

2 RMS的系统的主要功能及架构分析

为了更好地向读者介绍RMS系统,本文首先简要介绍一下RMS系统的发展及应用历史。

2.1 RMS的发展历史及功能回顾

RMS系统于1990年正式立项开发,1991年投入市场,经历20多年的发展历史,大致可以分为如下几个阶段。

● **DOS单机阶段。**1990年,中国科学技术情报研究所为了满足对科技文献二次文献的计算管理与检索服务,提出开发一个自主知识产权的情报检索软件,要求这一软件能够满足科技文献的著录、存储管理、检索、输出等功能,这项工作由我本人承担。当时几乎没有任何可供参考的产品和资料,PC机应用也是刚起步,DOS英文操作系统仅支持简单的汉字输入/输出,开发人员更是寥寥无几。仅有的开发基础是本人刚刚完成“七五”公关项目中的“大型情报检索软件CDS/ISIS汉化”课题^[1],CDS/ISIS情报检索软件是联合国教科文组织的CDS部门开发的一个运行在大型计算机IBM43系列机上的一个情报检索软件,由System 370汇编语言开发而成。通过对该系统的汉化,我掌握了CDS/ISI的系统架构及其实现方法,这一系统由当时国际上的知名情报检索专家Giampaolo Del Bigio先生设计开发,Giampaolo Del Bigio先生是我在UNESCO工作期间的直接领导者,也是我情报检索生涯的启蒙恩师。CDS/ISIS代表英文Computerised Documentation Service / Integrated Set of Information Systems,表示是由UNESCO CDS部门开发的集成信息服务系统。鉴于上述原因,我们把这一产品命名为ISTIC/ISIS,ISTIC是中国科学技术情报研究所的英文缩写,开发语言选用Pascal语言,开发工具Turbo Pascal 7.0。经过近一年的设计、开发、测试等工作,ISTIC/ISIS投入应用并进行市场推广,在图情机构、档案部门得到很快推广。中国企业产品数据库就是使用当时的ISTIC/ISIS作为数据库支撑工具。这时候的ISTIC/ISIS是单机DOS版,在汉化的DOS操作系统下支持汉字处理。ISTIC/ISIS主要支持如下功能:

^{*} 本文系国家高科技发展计划(863计划)“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题(编号:2011AA01A206)成果之一。

(1) 支持数据库结构的定义, 数据库字段可变长、重复并包含多值字段;

(2) 根据数据结构可定义数据库的录入工作单, 工作单内可支持缺省值定义;

(3) 数据库库存储支持变长字段存储, 包括字段数可变;

(4) 支持用户对可检索字段及检索方式的定义, 中文支持单汉字索引;

(5) 支持记录内全文检索、精确检索、字段限定检索、前方一致检索、基于布尔表达式的逻辑“与”、“或”、“非”的组配检索等;

(6) 通过格式化语言来定制用户的输出格式, 支持多字段的多级排序输出, 无需编程可实现任意格式的编目格式输出;

(7) 支持ISO-2709、CCF、MARC等图书、情报标准数据格式的交换。

注: 在汉化DOS操作系统下实现中文的检索和格式化排版输出需要开发单独的中文字符集处理模块, 尤其是中英文混合字符集的处理, 难度更大。

● Windows版本。1990年Windows 3.0版正式发布 (Windows的前两个版本只是MS-DOS的简单扩展), 随着中文版Windows操作系统发布, ISTIC/ISIS移植到Windows系统成为必然。1992年使用Borland公司的Delphi 1.0版, 完成了ISTIC/ISIS的Windows版本的工作, 同时产品更名为Quick IMS (Information Management System)。该系统面世后很快得到普及推广, 成为国内图书、情报、档案等部门的首选数据库管理软件。Quick IMS和以前的版本相比, 不再仅是一个情报检索管理软件, 已经成为一个包含数据库定义、数据库维护与管理、数据库检索等完善数据库管理功能的数据库管理系统 (支持局域网内的并发管理), 和DBase、Foxpro等关系数据库系统相比, 除了具备完善的关系数据库管理功能外, 还具备了数据库记录变长管理、字段多值、内容索引等内容管理系统所需要的功能, 因此作者提出Quick IMS是一个面向内容管理的非结构数据库管理系统。Quick IMS在中文信息管理方面的技术和市场优势得到中国软件行业协会的认可, 连续三年被评为“中国优秀软件产品”。同期Quick IMS的相关技术获得国家科技情报优秀成果“一等奖”, 及国家级重大科技成果认定。

● Web版。1995年中国公用计算机因特网开通, 在北京和上海建立了国际节点, 完成了与Internet的互联

和与国内公用数据网 (ChinaDDN) 的互联, 标志中国的互联网应用正式走向社会。大部分信息服务系统面临从本地服务走向Internet服务的模式转变, 软件通信技术也面临从TCP/IP到HTTP的技术转变。基于浏览器Web Server (BS结构) 的技术架构的信息服务系统成为发展的必然趋势。1997年, 经过近2年的研发, 我们推出了基于BS架构的通用信息服务系统IMS Web版本。这一版本的技术上主要包括如下特点:

(1) 将HTTP服务器直接与数据库服务集成, 无需应用服务器支持, 更不需要第三方其他工具支持, 比如IIS。

(2) 开发IMS Web数据库管理工具, 支持数据库定义创建、索引方式及索引技术定义, 支持10种索引方式; 支持数据库在线维护, 可定义数据库录入格式、支持关系数据库数据导入和挂接; 扩展支持XML数据交换服务等; 数据库字段扩展支持外部文件管理及全文索引。

(3) 开发IMS Web系统配置管理工具, 支持对数据库访问端口、数据库路径、访问认证、数据库挂接、数据库使用显示格式等数据库服务配置。

(4) 用户通过标准浏览器可直接访问数据库服务器提供的各类服务, 支持Open URL以及跨库检索。

(5) 扩展格式化语言, 支持HTML嵌入式格式化语言, 定义数据库显示网页、关联网页等, 无需编程; 扩展对外部文件内容的格式化处理和。

(6) IMS Web是一个基于HTTP协议的安全数据库服务系统, 可在裸操作系统下提供数据访问服务。

IMS Web的推出大大提高了系统的功能、效率和应用范围, 该系统很快扩张到政府及安全部门, 是国内第一个推出的内容管理软件, 万方数据资源服务系统就是使用该产品开发完成的。该系统获国家科委科技进步“二等奖”。

● Web Service版。Web Service是一项新技术^[2], 能使得运行在不同机器上的不同应用无须借助附加的、专门的第三方软件或硬件, 就可相互交换数据或集成。依据Web Service规范实施的应用之间, 无论它们所使用的语言、平台或内部协议是什么, 都可以相互交换数据。Web Service是自描述、自包含的可用网络模块, 可以执行具体的业务功能。Web Service也很容易部署, 因为它们基于一些常规的产业标准以及已有的一些技术, 诸如XML和HTTP。Web Service减少了应用接口的花费。Web Service为整个企业甚至多个组

织之间的业务流程的集成提供了一个通用机制。Web Service由IBM和微软公司在2002年正式提出,很快成为一种标准,并作为SaaS模式首选方案。2003年万方数据公司在IMS Web基础上推出Web Service版本和基于Windows平台的COM数据库服务组件,新的系统除了支持非结构化数据库管理外,可直接管理关系数据库和各类文档,因此把所有被管理的数据称之为资源,新的数据库管理系统命名为资源管理系统,简称RMS (Resource Management System), RMS的Web Service版被命名为R Service。R Service的主要技术特点如下:

- (1) 符合Web Service技术架构,支持SOAP、WSDL和UDDI;
- (2) 资源服务区和SOAP服务器直接集成,提高了系统效率 and 安全性;
- (3) 将资源管理中的数据库定义、数据维护、数据索引、数据检索、数据输出和交换定义为标准的Web Service服务接口。接口及接口描述可跨平台远程获取和调用;可在任意应用服务中调用R Service提供的任何接口服务(通过认证信息);
- (4) RMS系统内增加了使用中文分词技术的全文索引和检索功能,同时提供了基于中文分词的全文模糊检索功能;
- (5) 增加了基于全文内容的自动标引和分类服务功能,用于科技文献的辅助主题标引和学科分类(中图分类法)。

R Service自推向市场后在国内政府、图书、情报、科研部门得到广泛应用,公安部综合查询系统、各省市科技文献共享服务系统,中石油、民航总局等一些大型内部科技文献支撑服务系统都采用的是R Service系统。关于R Service详细技术架构以及RMS COM系统的一些特点本文不作介绍。R Service(万方数据资源服务于管理系统)2007年获北京市政府科技进步三等奖。

● RMScloud。以上介绍的RMS的系列版本都是集中式情报检索系统,传统的集中式的情报检索软件已经无法满足飞速发展的信息爆炸和普及化的海量用户需求,能够提供“云服务”的分布式情报检索系统已经成为必然。2011年底国家“863”计划重大专项“云计算一期”启动,并下达“以科技文献服务为主的搜索引擎研制”课题任务,北京万方数据股份有限公司为课题依托单位,北京万方软件有限公司负责技术开发工作。到

目前为止,基于“云计算架构的分布式搜索引擎RMS Cloud”已经研发完毕,正在通过“中国学术搜索网”进行全面测试。关于RMScloud参见本期《分布式学术搜索引擎研制及其大数据应用》一文。

2.2 RMS数据库系统架构

以上对RMS系统的发展历史及其各版本的功能及技术特点作了介绍,从中可以看出RMS为满足技术发展和用户需求的技术体系变更。尽管从技术体系上有了很大的变化,但RMS作为一个专业情报检索系统,其核心功能及其系统架构并没有发生本质性变化。下面介绍RMS系统的数据库管理架构,使读者了解情报检索系统的工作原理,见图1。

RMS的每一个数据库包括两部分,其一是变长存储管理的数据文件(简称主数据文件),其二是为数据的快速检索而建立的索引文件系统(在RMS系统中称之为倒排文件)。RMS数据库记录是变长存储的,这里的变长是指记录中每一字段的长度是变长的,记录中的字段数也是可变的;字段的变长存储是指按每一字段的实际长度存储数据,对于文献数据库而言可大大节省存储空间(关系数据库的字段长度是定长的,字段数也算固定的,数据按表格方式存储)。在RMS系统中每一数据库的记录格式通过数据库的字段定义表(FDT)来描述。FDT表格式如下:

- (1) 数据库记录中的每一个字段用一个FDT行来描述;
- (2) 每一个FDT行中包括字段名称、字段号、字段长度、字段类型、是否可重复、子字段、字段缺省值等说明项。其中字段名称的最大长度为30个字节,字段号是给出的数据字段的标识,字段标识是规范化的,可参照ISO-2709、MARC等数据交换标准,RMS系统内部以字段号为标识(RMS系统为专业图书、情报服务系统的特征);字段长度是一个象征,因为RMS的字段存储是变长的;字段类型除常规数据类型外,还支持重复字段(或者是多值字段,比如作者、成果完成人等),这是一个图书、情报系统所必需的功能;子字段是图书、情报系统著录必需的数据类型;缺省值用于数据著录,提高著录效率。

RMS系统通过数据库定义的FDT表来创建数据库的主数据文件,主数据文件是由一系列变长记录数据所组成的。由于RMS数据记录是变长的,所以数据

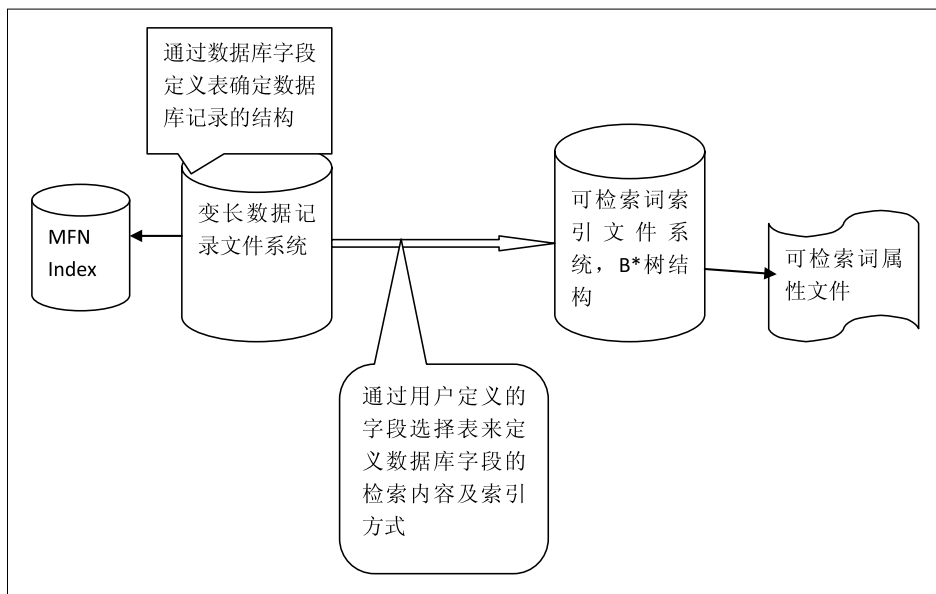


图1 RMS系统架构图

库记录的维护相对复杂，但记录库数据的变长存储对于迅速增长的文献数据库而言是很重要的，尽管目前存储设备的性价比得到大大提高。为了提高主数据文件记录的直接获取效率，系统为主数据文件建立了一个定长的记录号索引，每一个记录号索引中包括对应记录在数据主文件中的块号和起始位置，通过记录号和对应的索引信息可以快速找到对应的记录信息。RMS系统提供了数据库记录的增、删、改等基本功能。

RMS系统和其他数据库系统一样，最核心的部分是数据检索，直接对主数据文件顺序检索显然是不行的，需要预先为数据库的可检索项建立索引，通过索引文件可以快速找到可检索项并通过对检索项的属性列表处理，形成检索结果。RMS系统通过字段选择表（简称FST）来定义数据库的可检索字段、索引方式和索引内容。每一个FST行定义一种可检索方式，包括字段号、索引技术和索引内容三项参数，三项参数分别解释如下：

(1) 字段号：这里的字段号是字段限定检索时使用的标识号，可不同于主数据文件中记录字段的字段号；

(2) 被索引字段的内容，用RMS格式化语言来描述，被索引的内容可以是某一字段或多个字段，也可以是其中的一部分；

(3) 对被索引字段使用的索引技术，它决定了被检索内容的检索方式和系统整体检索功能。RMS索引技

术多达10种，几个主要索引技术阐述如下：

.索引技术0：把被索引的内容整体作为一个整体检索；

.索引技术1：把被索引的内容中每一个用子字段标识的内容作为整体检索；

.索引技术2, 3：分别把被索引内容中的用<...>和/.../标识的检索项作为一个抽取出来建立索引，这一技术允许标引人员直接在文摘或正文中进行标引；

.索引技术4：把被索引内容中的每一个英文单词或中文单字抽取出来作为检索词，然后可通过相邻检索实现全文检索。这一技术可提高全文检索的查全率，弥补汉字分词误差造成的影响，但同时降低了系统的检索效率。单汉字检索对于一些特殊字段是必须的。

.索引技术10：把被索引内容中的英文单词或中文内容通过分词技术形成的词作为可检索词（禁用词除外），然后通过位置检索实现全文检索。

RMS可对同一字段采用多种索引方式，比如作者字段（字段号为10），在FST中的定义如下：

10 0 v10; //对10字段建立整字段索引，检索字段号为10；

101 4 v10; //对字段10按单字索引技术索引，检索字段号为101；

102 4 rpin(v10); //首先将字段10的内容转换成拼音，然后按单字索引，索引字段号为102；

RMS系统通过FST和相关索引技术即可抽取数据库中的可检索词，系统为了实现不同的检索方式在抽取检索词的同时，为每一个检索词生成了一些属性信息，RMS系统建立索引时为每一检索词获取的属性包括所在记录的记录号、记录中的字段号和字段内的词序，关于这些属性信息的用途后面作详细分析。为了实现可检索词的快速查找，需要为数据库中的所有检索词建立一个索引文档，索引文档的组织有多种方式，大部分搜索引擎采用的是倒排文件组织（Inverted File）。倒排文件的组织通常以B+树结构为主^[3]，但RMS系统采用的

是B*（多叉平衡树）树结构存储管理可检索词。RMS系统倒排文件结构如图2所示。

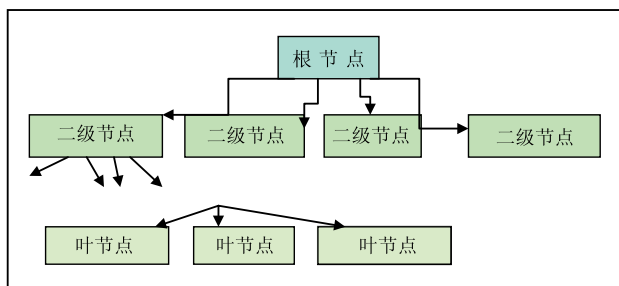


图2 RMS系统B*树结构图

图2给出的是RMS系统采用的B*树结构的倒排文件示意图，B*树是一种子节点数不固定的平衡树结构，在RMS系统中B*中一个关键词的比对最多需要7次。倒排文件中每一个节点中除代表一个关键词外还有一个属性列表指针，指向每一关键词的属性列表。

图3是倒排文件中关键词属性列表的示意图，每一

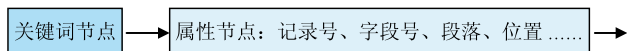


图3 节点属性列表示意图

个属性列表中的属性内容确定了检索系统的检索功能。简单解释如下：

(1) 如果属性值只包括关键词的来源记录号，检索系统中只支持基于记录级的布尔检索；

(2) 如果属性之中包括字段属性，检索系统可支持限定字段的布尔检索；

(3) 如果属性之中包含段落信息，检索系统可支持限定段落内的布尔检索；

(4) 如果属性之中包含位置信息，检索系统可支持位置检索（相邻检索），全文检索实际上是位置检索的一种特殊形式。

通过上面的解释也可以回答一个检索系统评价中倒排文档和数据文档比例的问题；同时倒排文档的组织 and 物理结构直接影响到检索系统速度，倒排文档中关键词的属性决定了检索系统的检索功能。非情报检索系统把倒排文件称为索引文件（Index file），是因为它们为每一个字段建立了一个倒排文件，甚至没有字段区分。而本文所提及的倒排文件是一个多字段的复合索引文件，尤其在多字段组配检索方面具有较高的检索效率。

2.3 RMS系统的检索流程

RMS系统的核心功能是情报检索，情报检索系统的检索功能实现流程如下：

图3是RMS系统的检索流程图，现对流程图中的各功能点解释说明如下：

(1) 用户提交检索表达式给系统（一般通过应用系统获取用户检索需求并传递给RMS系统，比如R Service），检索表达式可以是自然语言或标准的检索语言，比如CQL，系统中的检索表达式首先对检索表达式进行预处理，包括检索词分词（和索引建立时的分词技术和方法要一致，这一点很重要）、词表扩展、跨语言转换、表达式规范处理等；

(2) 检索处理引擎对规范后的检索表达式进行检索处理（这是搜索引擎的核心技术）。处理过程中首先将检

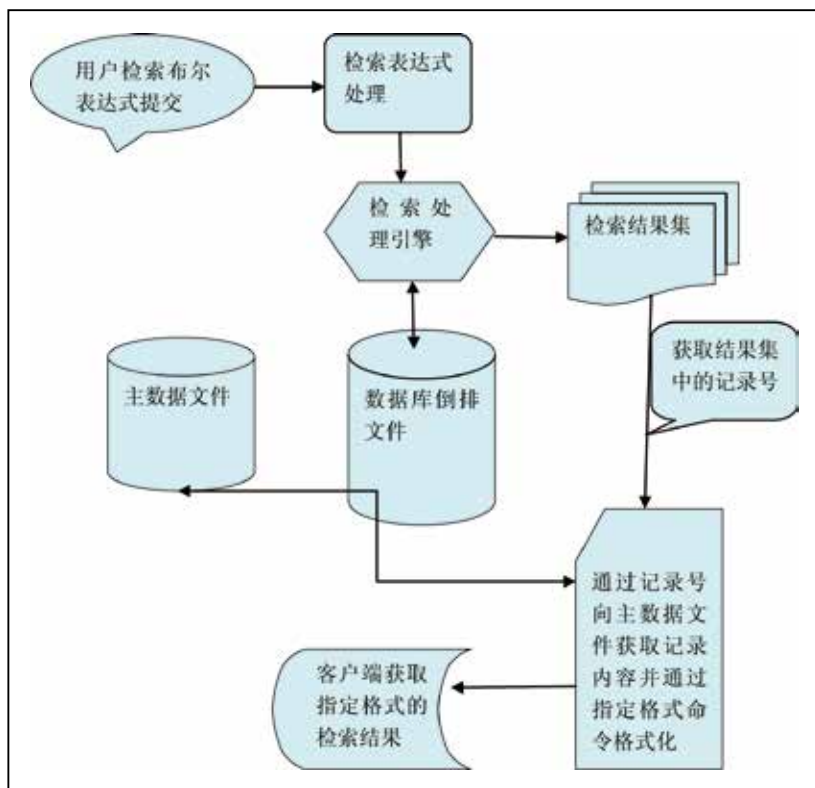


图4 RMS系统检索流程图

索表达式中的每一个检索词进行处理,也就是通过对倒排文档中检索词的快速定位,然后获取检索词的属性列表,然后根据检索词间的逻辑关系及运算优先级进行逻辑运算,最后形成检索结果,即命中记录号的集合。系统对检索表达式的逻辑关系支持,取决于倒排文件检索词的属性内容;

(3) 可通过检索命中记录的集合获取每一命中记录的记录号,通过记录号可从主数据文件中获取记录的全部内容,然后通过系统内部的格式化处理形成用户需要的输出格式。格式化处理是RMS系统的独特功能,用户可以通过嵌入式格式化语言定义多个显示格式,由系统通过格式化处理,获取自己所希望的格式。

3 情报检索系统的核心功能研究与实现

计算机情报检索系统经过近40多年的发展历程,历经多次重大技术变革,但情报检索系统核心功能并没有发生本质变化,本人从事情报管理与检索技术的研究20多年,是RMS系统的总设计师和核心技术开发人员,多有心得和感慨;尤其是近10年全文检索系统和搜索引擎系统的迅速发展,甚至一些情报检索的专业人员,似乎忘记了我们情报检索的专业需求,普通搜索引擎大有取代专业情报检索系统之势。在这种环境下借助本文,作者对情报检索系统的核心功能进行研究总结,同时介绍RMS系统的实现方法,以唤起大家对专业情报检索的重视。

3.1 情报检索系统的核心功能

情报检索系统除了具备数据检索功能外,还要包括一些数据维护与管理的功能,确切讲它首先应该是一个数据库管理系统,其主要功能包括数据库定义、数据维护与管理、索引文档维护、检索、输出及数据交换等功能。各项功能说明如下:

(1) 数据库定义:用户可根据需要定义数据库记录的结构、可检索项及检索方法、输出格式及字段间的关联属性。数据库定义是情报检索系统的基础,也是顶层设计。数据库记录结构的设计原则上要符合专业标准,比如Marc、CCF、XML_DC等。可检索项及检索方法确定了数据库索引文件的结构及内容,检索只能检索到索引文件中存在的内容,因此这一定义及实现决定检

索系统的核心指标“查全/查准确率”。输出及交换格式在情报检索系统中较为复杂,比如卡片格式、详细格式、参考文献格式等。字段间的关联属性决定了数据库内、外部逻辑关系,更体现了整个服务的数据关联性。

(2) 数据维护:数据维护的基本功能包括数据记录的建立、修改、删除等功能。但是,情报检索系统处理数据复杂,从编目数据、二次文献数据甚至到全文数据,字段多、类型多、长度变化更大,如果采用结构化表格数据库去处理,浪费空间巨大。因此情报检索数据库要求能够管理非结构化的数据和变长存储。

(3) 数据检索:数据检索是情报检索系统的核心,其特点是数据量大、检索功能复杂(很多功能无法用SQL来描述)，“查全/准确率要求高”。为了提高系统的检索效率,需要对可检索项进行预处理并建立倒排文件,其最核心的特点是要求对数据按内容进行索引(关系数据库是对字段的属性值进行索引)。检索功能包括检索词精确检索、模糊检索、前方一致/后方一致检索等检索词检索、全文检索,同时还要有对检索词的字段限定及逻辑组配检索(逻辑关系包括“与”、“或”、“非”等)。检索表达式用布尔表达式描述,一般情况下采用CQL为标准。对最终用户而言,其输入的检索词可以是自然语言,也可能是专业CQL,用户检索需求的处理通常是应用系统层面的事情。

(4) 数据输出:情报检索系统最大的特点是输出格式多样化,不同类型的数据输出格式差异较大,因此要求系统能够灵活定义数据输出的格式。除了输出格式有特殊要求外,系统还要求有灵活的排序方式,比如命中相关性、日期、文献影响力、文献类型等。

(5) 数据交换:数据交换是图书、情报文献系统的特殊需求,而且有国际标准格式,比如ISO-2709、Marc、CCF等。图书馆联合编目是需要数据交换的典型应用。

3.2 RMS系统的情报检索与管理功能

从前面的RMS发展历史回顾中可以看出RMS系统是一款专为情报检索与管理设计的一款非结构化数据库管理系统。下面以RMS系统的Web Service版^[4]R Service为例,介绍其技术实现方法。

(1) R Service系统中包含一个控制台软件,被命名为RMS MIC(我的信息中心),它既是一个单机版数据库管理软件,同时也可在服务器端安装用来定义及

管理服务器端的数据库。

(2) R Service的数据库管理的核心是变长管理的数据主文件和通过字段选择表配置管理生成的B*倒排索引文件。数据库定义支持使用MARC、ISO-2709、CCF、DC等标准,通过字段选择表可定义字段任意内容的索引技术和方式,对于每一个可检索项包含记录号、字段号、和字段内序号信息,因此可支持字段限定和字段内位置检索(包括相邻、相邻模糊匹配等功能)和精确的全文检索。

(3) R Service的数据检索支持布尔表达式,通过B*树结构的倒排文件可快速实现检索项定位,同时对检索项的属性列表采用多项优化技术,支持千万级数据库检索的秒级响应。

(4) R Service支持嵌入式格式化语言,无需编程可实现索引内容的控制、记录格式输出(包括任意格式的XML-DC输出)等功能。

(5) R Service系统首先集成了我们自主研发的SOAP服务器,支持所有相关的Web Service标准,同时提供了实现上述所有情报检索与管理功能的Web Service接口,用户可通过任意平台调用相关接口,开发自己的应用服务系统。

关于R Service的具体实现技术本文不作介绍,本文的目的是通过对RMS系统的技术架构和功能分析,使大家充分了解一个通用情报检索系统应该具备的功能和相关技术。

4 情报检索与服务系统展望

计算机情报检索系统经过40多年的发展历史,经历多次技术变革,但从原来的技术体系来看基本成熟,尤其是近2年分布式“云计算”架构下分布式搜索引擎的出现,使大家更清晰地看到了未来情报检索系统的发展方向。受国家“863”计划课题“以科技文献服务为主的搜索引擎研制”的资助,基于“云计算”架构的分布式系统RMSCloud云搜索引擎已经研发完成,并建立示范应用“中国学术搜索网”。随着“大数据”时代的到来,情报检索系统发挥的作用越来越大,主要体现在如下几个方面:

(1) 专业情报检索系统和普通搜索引擎系统会相互影响,互相进步,最终实现整合,统称搜索引擎。从Google学术搜索系统可以印证这一点,该系统引入了大量情报检索服务功能。

(2) 在大型信息服务系统中传统的“集中式”情报检索系统将被“分布式”搜索引擎所取代,但对于一些中小型信息服务系统,仍然需要原来的集中式系统;分布式搜索系统需要大量高端服务设备,而且需要更复杂的技术支持。

(3) 情报检索系统特有的内容管理(索引)功能,能够较好地用于非结构化数据的数据挖掘与分析功能;“大数据”应用的核心是挖掘与分析,挖掘与分析除了技术上支撑外,更重要的数据规范,从前面介绍大家可以看到,情报检索系统提供了更有针对性的支持。

另外,目前的情报检索系统的核心技术是对可检索内容的预处理形成“倒排文件系统”,然后通过检索词的快速匹配形成检索结果,显然这种架构的检索具有很大的局限性。多年来许多专家学者试图改变这种现状,提出了“向量检索”、“模糊匹配检索”、“基于人工智能的自然语言检索”等多种检索方法,但受“倒排文件系统”技术的影响,一直无法获得突破。众所周知,现代“分布式”搜索引擎的指导思想是在传统“集中式”搜索引擎的基础上通过Map/Reduce实现的。作者认为“Map/Reduce”的指导思想^[5],可以打破现有的“倒排文件”技术瓶颈,实现更为“智能化”的检索处理,我们拭目以待。

结语

本文借此专刊介绍了RMS系统的发展历史及其为实现专业情报检索服务系统所采用的技术架构,目的只有一个,就是希望国内青年专业人员更好地理解情报检索的专业需求,不要盲目相信“搜索引擎”所带来的种种影响。国内现有的文献服务系统大都看似提供了专业化的信息服务功能,但其结果大都经不起推敲,究其原因还是开发者不太了解情报检索系统的需求所致。本文不当之处,敬请批评指正。

参考文献

- [1] 吴广印.微机通用信息管理系统Micro CDS/ISIS的开发[J].情报学报,1990(1):25-31.
- [2] 百度百科.Web Service[OL]. [2013-04-20]. <http://baike.baidu.com/view/1086510.htm>.
- [3] 程序人生.倒排索引-搜索引擎的基石[OL]. [2013-04-20]. <http://blog.csdn.net/hguisu/article/details/7969757>.
- [4] 吴广印.基于Web Service构架的资源共享技术研究与实现[J].情报学报,2007(6):851-857.
- [5] MapReduce: Simplified Data Processing on Large Clusters [OL]. [2013-04-20]. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/zh-CN/archive/mapreduce-osdi04.pdf.

作者简介

吴广印 (1965-), 男, 中国科学技术信息研究所研究员, 北京万方软件有限公司董事长。RMS系统的总体设计师和主要开发人员, “863”专项课题“以科技文献为主的搜索引擎研制”的技术负责人。研究方向: 非结构数据库管理系统和中文信息检索。E-mail: gywu@wanfangdata.com.cn

The RMS System Structure and Function Demand Research of Information Retrieval System

Wu Guangyin / Institute of Science and Technology Information Research of China, Beijing, 100038

Abstract: In this paper, the development history of RMS and the system architecture and design idea are introduced in detail, and the functional requirements of information retrieval system of research and analysis, finally, the further development and application of information retrieval system are discussed.

Keywords: RMS, Information retrieval, System architecture, Indexing technology, Variable-length storage management

(收稿日期: 2013-05-02)