

分布式学术搜索引擎研制及其大数据应用^{*}

□ 吴广印 / 中国科学技术信息研究所 北京 100038

摘要: 受“搜索引擎”流行的影响, 目前大家已经习惯把图书情报领域使用的“情报检索系统”称之为“学术搜索引擎”。无论从技术层面上还是应用层面上, 尽管二者有很大的共同点, 但也有很大差异。传统的集中式的搜索引擎已经无法满足飞速发展的信息爆炸和普及化的海量需求用户, 能够提供“云服务”的分布式搜索引擎已经成为必然。文章主要内容包括学术搜索引擎涉及的关键技术、分布式搜索引擎的架构, 以及分布式搜索引擎在大数据领域的主要应用价值三个方面, 最后给出了分布式搜索引擎RMSCloud的典型应用介绍。

关键词: 学术搜索引擎, 分布式检索, 大数据应用, 云服务, RMSCloud

DOI: 10.3772/j.issn.1673—2286.2013.06.003

1 前言

情报检索系统 (information retrieval systems) 是对情报资料进行收集、编辑、管理和检索的系统。现代情报检索系统是由电子计算机、通信网络和终端设备等组成的自动化系统, 可进行情报资料的收集、标引、分析、组织、存储、检索和传播等工作。计算机情报检索可分为数据检索、文献检索、图谱检索、事件检索等类型。传统计算机情报检索的服务方式又可分为三类: ①定题情报服务。它是针对相对固定的用户提出的要求, 定期对新到文档进行检索, 及时向用户提供所需信息。②回溯情报检索。它是根据用户的要求, 对过去某段时间内积累收藏的全部文献, 进行主题检索, 一般采用脱机批处理方式。③联机情报检索。它采用人机对话的方式, 用户在计算机终端上经过通信线路直接与计算机对话, 能在短时间内获得检索结果。而现代情报检索服务系统, 借助先进的互联网

技术能够提供全方位一体化的在线服务功能^[1]。情报检索系统的核心是“检索” (retrieval), 它不同于目前搜索引擎中的“搜索” (search)。目前通用的搜索引擎, 以Google、百度为代表, 实际上仅提供了以文本全文检索 (字符串匹配) 为主的Search功能和一些简单的智能扩展服务。情报检索系统的主要考核指标是“查全/查准率”, 应该能够在浩瀚的文献海洋中, 快速准确地“捞针”。现在, 几乎所有的互联网用户每天都在使用搜索引擎去上网搜索网上的信息, 搜索的结果大家只能再次在近似海量数据里去做人工选择。另外网上搜索引擎的搜索主要对象是网页, 很难评价其真实性, 但突出的优点是信息获取及时。情报检索系统的检索对象是文献, 基本上都是正规的出版物信息, 是经过专家或权威机构评审过的内容, 可直接参考引用。近几年Google和百度也相继推出了“Google Scholar”、“百度文库”等类似的文献搜索平台, 但技术上仍然是“搜索”。近几年“情报检

索系统”一词几乎被人们淡忘了, 被“学术搜索引擎”所替代。所以本文也以“学术搜索引擎”为题, 去探讨我们业内“情报检索系统”的相关核心技术和应用。

2 学术搜索引擎的关键技术研究

从存储在不同类型的数据库中去发现满足自己所需数据称之为“数据搜索”, 数据可以存储在各类数据库中, 其中包括关系数据库、非结构化数据库以及近几年流行起来的NoSQL数据库 (Not only SQL)^[2]。科技文献类数据库由于其自身特点, 比如变长、多值、字段数量变化等特点, 通常使用非结构化数据库来管理。因此, 搜索引擎不同于专业数据库管理系统, 它只是构架在数据库管理系统之上的搜索功能模块 (搜索引擎通常也提供简单的数据管理功能), 因此本文讨论的是学术搜索引擎 (以下简称搜索引擎) 相关的关键检索技术。

^{*} 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。

2.1 搜索引擎的技术架构

搜索引擎的主要功能是提高常规数据库管理系统的检索效率,主要包括速度、效率、检索“查全/查准率”等。图1为搜索引擎的技术架构图,说明如下:

a.通常搜索引擎通过数据库管理系统提供的接口获取数据库的数据,比如关系数据库提供的ODBC、BDE等数据驱动方式,RMS系统通过Web Service方式对外提供数据访问接口。

b.搜索引擎将从数据库获取数据建立索引数据库以满足未来数据的快速检索。搜索引擎建立索引数据库时涉及索引策略、数据分词、存储结构等关键技术。索引数据库是搜索引擎的核心,直接影响搜索引擎的检索速度、查全/查准率等关键功能指标。

c.用户提交检索需求,检索需求可以是自然语言或者是符合CQL标准的查询语言,搜索引擎收到检索请求后,按照相同的分词规则对检索请求进行分词和扩展处理,扩展处理一般包括词表扩展、跨语言扩展等。词表扩展对于提高系统的查全/查准率尤为重要,词表包括主题词表、机构词表、人物词表、分类词表等^[3]。

d.搜索引擎根据检索需求,对索引数据库进行检索,形成命中结果,命中结果的数据是命中记录号集合,命中记录号是常规数据库中的数据记录唯一标识符,通过命中记录号可从数据库中获取数据。检索时包含多项检索技术,比如单检索词结果的优化提取、检索结果的逻辑组配策略等。

e.检索结果输出。检索结果输出一般分为两类,聚类输出只输出命中结果数。

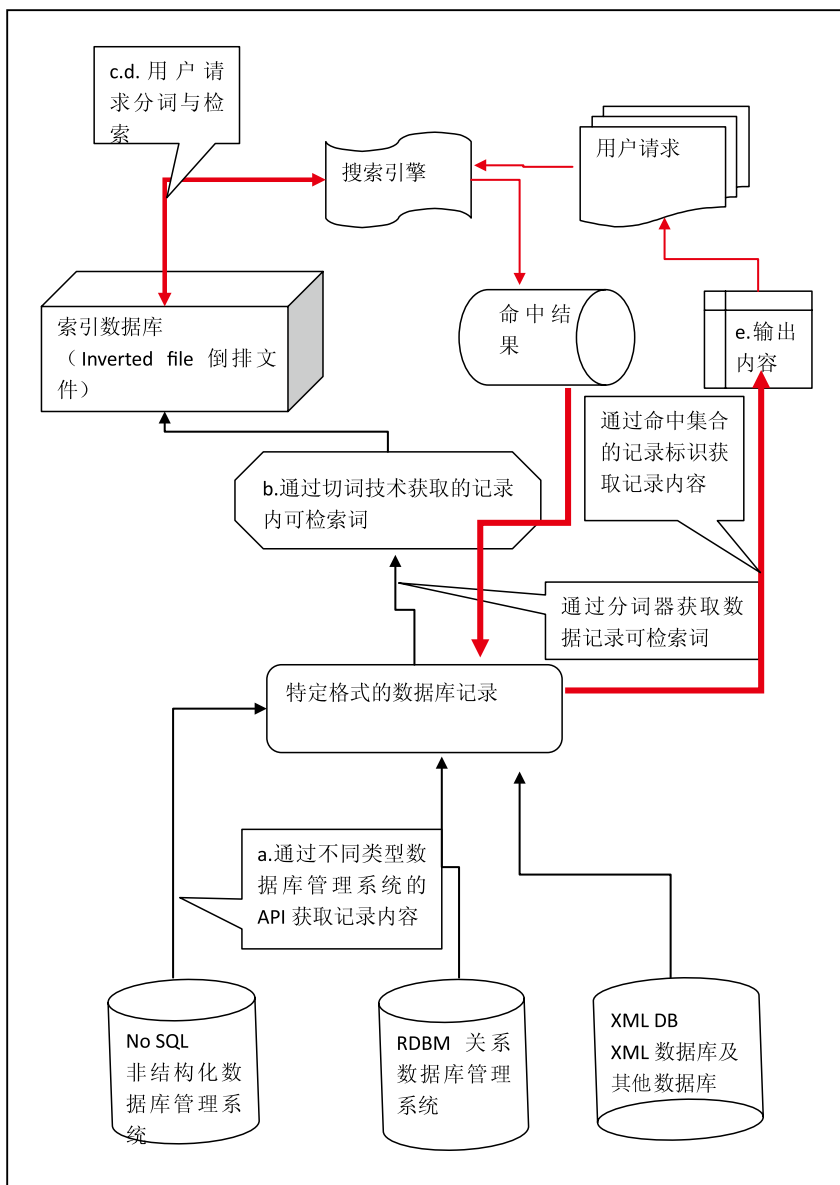


图1 搜索引擎技术架构图

另一类是要求输出命中结果中记录的内容按页输出。输出结果的展现要有排序,通常包括时间、相似度、论文被引次数、地区、国别、语种、类型等指标排序输出。排序输出涉及多项核心技术,下面章节介绍。

2.2 搜索引擎研制中涉及的关键技术

本节对上一章节中提到的关键

技术进一步说明,同时对我们如何实现这些技术作简单介绍,更加详细的说明在其他专门的文章里进行论述。

2.2.1 数据获取技术

搜索引擎是构架在传统数据库管理系统之上的专门负责数据检索的搜索系统,其目的是为了提高传统数据库管理系统的查询效率和功能。因此搜索引擎

必须有从传统数据库中获取数据去处理。传统数据库大致可以分为三类数据库：关系数据库（Oracle、DB2、MS SQL、MySQL等）、NoSQL数据库（Cassandra、MongoDB、CouchDB、Redis、Riak、Membase、Neo4j和HBase等）和专业非结构化数据库管理系统^[4]，比如我们研发的RMS非结构化数据库管理系统。关系数据库由于起步较早，提供了工业化数据存储接口标准，通用标准为ODBC。NoSQL数据库是近几年才发展起来的非结构化数据库系统，典型代表是大型分布式开源数据库Hbase（Hadoop database）^[5]，基于Google的BigTable开源项目，目前国际上大部分数据库是构架在该数据库基础上的。NoSQL数据库大都支持HTTP访问协议，很容易从数据库系统中获取数据。其他类型的一些专业情报检索系统也都提供了标准数据访问接口，比如万方软件推出的非结构化资源服务系统R Service除了提供标准本地COM+服务外，还提供了跨平台的Web Service服务，通过任何开发平台都可以方便地获取格式化数据。

2.2.2 数据抽取技术

搜索引擎的主要目的是为了提提高传统数据库管理系统的检索效率，提高数据库检索效率的唯一手段就是为数据库中定义的可检索项建立索引文件（情报检索系统中一般称之为“倒排文件”）。为可检索词建立索引文件的基础是根据定义从数据库中抽取出检索词，抽取检索词是由系统需要提供的检索功能和方式所决定的。通常的情报检索系统需要支持如下几种检索方式：

- 整个字段检索：这是所有数据库系统的检索功能所必须支持的，把整个字段作为一个检索词处理，通常该字段的长度不超过30个字节，所涉及的数据抽取技术非常简单。

- 主题词检索：一般指的是主题词字段的检索，主题词字段是对本文记录的主题描述，是文献检索的核心功能，一般来讲主题词是由文献创建者自己给出的；一个主题词字段中一般包括3-5个主题词。提高一个文献检索服务系统的“查全/查准率”主要靠的是主题词检索，关于主题词检索在检索技术中介绍。由于主题的给出传统上是很专业的，要求作者了解本领域的主题词表（主题词必须是领域词表中的主题词），但大部分作者并不按标注标识，给出的“主题词”大都不符合要求；大量的文献需要图书情报专业人员进行二次标注，专业上称为“标引”，单人工标引费时、费力，投入较大，因此产生了自动标引这一研究领域。自上世纪70年代，自动标引一直是情报服务自动化系统的热门研究领域，全文检索技术或者说搜索引擎的出现，降低了人们对专业情报检索的要求（通用搜索引擎的普及影响造成，并非专业需求降低），主题词自动标引的研究热度有所降低，取而代之的是宽泛的“自动标引”。无论是手工标注的主题词还是机器自动标引的主题词，抽取时需要将每一主题词独立抽出处理，抽取技术相对简单。一般在专业情报检索系统中把这类字段称之为“重复字段”或“多值字段”。在数据库中作者、单位、参考文献等字段也有类似要求^[6]。

- 全文检索：全文检索是上世纪90年代初期问世的一种检索系

统，其特点是数据库的任何字段内容都可以进行检索，看似通过字符串的全文比对获取的检索结果，所以被大家统称为全文检索。实际上全文检索系统首先对可检索字段的内容进行切词，然后为切词后的每一个词（禁用词Stop Word除外）建立索引文件以实现快速检索。全文检索的核心技术是对内容的切词处理，通常对于西文数据，采用识别每一个单词为检索词，类似the、of、at等虚词为禁用词。中文内容的切词技术比较复杂，通常是基于一个基础词库和正向、逆向扫描分析法来进行切词；还有一种单汉字切词方式，即把每一个中文字符当作一个检索词，检索时通过相邻检索来实现检索词的检索。全文检索的优点是节省人工标引的工作量，缺点是无法实现主题检索，查准率降低。后来人们采用后控词表检索技术来弥补^[4]。

- 数值检索：搜索引擎所处理的数据绝大部分都是文本数据，但随着搜索引擎在数据挖掘、分析中的应用，数字字段或文本中的数字也需要检索，尤其是数值区间的检索，因此从数据中抽取数字也是搜索引擎应有的功能。数值检索一直是搜索引擎甚至常规数据库管理系统的短板，有待进一步研究提高。

2.2.3 索引数据库技术

通常搜索引擎先从数据库记录中抽取检索词，每个检索词包含一些特定的检索词属性（属性包括记录号、字段、位置信息等），把数据库中抽取的所有检索词和属性一起组织成一个数据文件；当用户检索时，可快速确定检索词的位置及

其属性内容,然后根据属性内容形成检索结果,通常我们把这类文件称之为“索引数据库”或“索引文件”。索引数据库在搜索引擎中的功能与作用如图2所示。

图2所示搜索引擎索引数据库建立流程是通用的,不同的搜索引擎只是在技术实现上有所差异,下面将介绍RMS系统技术实现及其特点。

- RMS是北京万方软件有限公司开发的一个非结构化数据库管理系统^[7],它可以通过格式化语言输出任意格式的文本数据被搜索引擎处理,也可以建立自己的搜索引擎。

- RMS数据库支持整字段、多值字段甚至子字段的精确检索、单汉字全文检索、基于字典的分词检索;同时支持对字段内容的主关键词标引和分类(中图分类)等10种检索词索引方式。

- RMS可对同一字段实现多种索引方式以满足不同的检索需求,根据不同的索引方式抽取检索词和检索词属性。RMS系统抽取检索词时生成的属性包括记录号、字段号和记录内的位置信息,检索词属性直接决定搜索引擎的功能。没有位置信息无法实现精确的全文检索,没有字段信息无法实现字段限定检索,只有记录号就只能实现记录内的检索词逻辑检索。

- RMS对生成的检索集合进行排序处理,然后建立一个基于检索词的B*树(平衡树)结构的检索词索引文件,每一个检索词指向一个不定长检索词属性列表,属性列表涉及变长数据块维护、存取、修改等多项技术。检索词的索引文件和属性文件构成了索引数据库,RMS的数据库文件和索引库文件

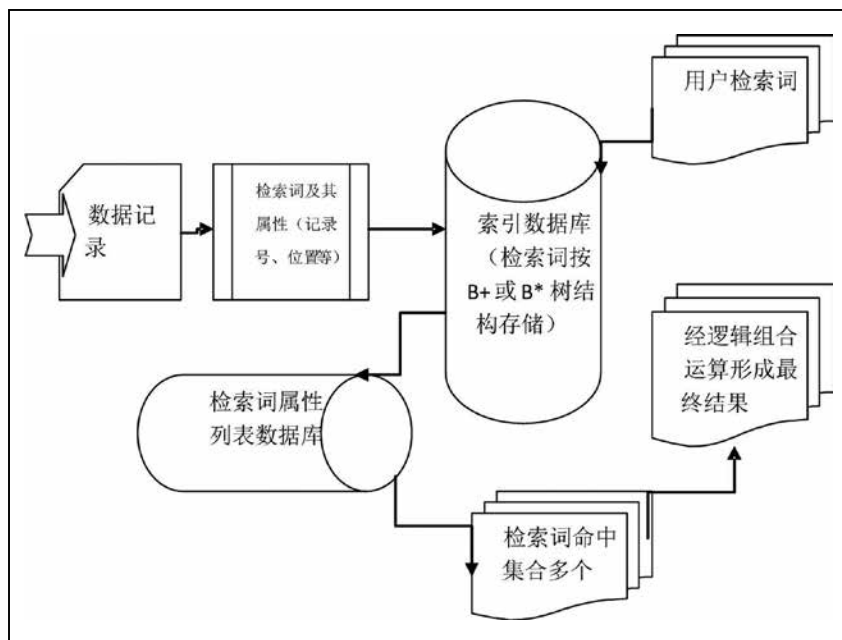


图2 倒排数据库建立及功能流程图

物理上独立。为了节省索引文件的存储空间,RMS建立了两个索引文件,一个用于短词(小于等于10个字节),另外用于长词(大于10个字节,小于等于30字节),RMS系统每一独立检索词的长度不得大于30个字节。

- RMS处理用户检索需求时分三种情况处理检索词:精确检索词,系统不做扩展处理。全文检索词:系统对用户提交的检索词进行系统切分,然后对切分后的检索词按位置相邻进行扩展。值得强调的是这里的分词技术一定要和数据库索引建立切词时使用同一分词技术和方法。模糊检索词:系统对用户提交的检索词进行系统切分,然后对切分后的检索词按位逻辑“与”进行扩展,同时模糊检索也支持前方一致扩展。

- RMS系统通过索引文件可以快速找到检索词的位置并获取检索词属性列表,RMS的B*树结构确定一个检索词的位置最多只需要

7次I/O,也就是说RMS的检索词的查找不受数据量多少的影响。然后系统对检索词的属性列表进行处理,包括去重、组合运算等,属性列表的大小和组合运算的算法直接影响系统的效率和功能。RMS系统支持字段限定、位置检索、模糊等多种检索方式和逻辑“与”、“或”、“非”及其复杂布尔表达式(CQL)的组配计算。

2.2.4 中文自动分词技术

正如2.2.2节所述,搜索引擎的主要功能是全文检索,全文检索的核心是分词,英文分词技术比较简单,中文分词较为复杂,是中文信息检索的核心,不同的搜索引擎在分词技术上有所不同,分词技术的效果直接影响搜索引擎的“查全/查准率”。本节主要介绍RMS系统所采用的分词技术及其实现。

中文分词技术大概分为三种:基于词典匹配的分词方法、基于统

计的分词方法、基于人工智能的分词方法。目前能够达到使用水平的只有基于字典匹配的分词方法,另外两种方式还处于研究实验阶段,没有看到实用产品问世报道。词典虽然存在效率不高、歧义处理困难以及无法囊括所有词等等不足(对于不在词典中的词RMS使用单汉字分词方法),但它实现简单,分词效率很高,所以大多数的系统以该方法为主来实现。一个完善的词典匹配法包括两部分:一是词典,二是相应的匹配算法,二者相互依存。目前常用的匹配算法有“长词优先法”和“短词优先法”,例如对于“有机化学合成法”,“长词优先法”得到的结果是“有机化学”、“合成法”两个词,而用“短词优先法”得到的结果是“有机”、“化学”、“合成”、“法”四个词。显然,“长词优先法”的优势在于专指程度高,可以消除大多数的歧义词,更适合于文献标引;它的劣势在于匹配效率低,而且程序的实现比较复杂,短词分词效率高,更适合全文检索,查准率较差^[8]。如果在字典里面有“计算”,但没有“计算机”这个词,全文检索中检索“计算”一词就会检索出含有“计算机”的文献,“计算”不同于“计算机”,系统查准率出现问题,这类错误在通用搜索引擎里被认为是允许的,但作为一个情报检索系统是绝对不允许的;如果字典里包括“计算机”,就不会出现上述扩检错误。显然搜索引擎的“查全/查准率”取决于系统的分词字典,尤其是中文搜索引擎。RMS系统的分词字典采用通用基础词、作者三字主题词构成(我们把近20年的5000多万期刊文献作者自标主题词进行了提取)、作者姓名、地名等去重后的

复合词典,分词匹配以逆向扫描为主,姓名、地名类采用正向扫描为主,同时采用了自主研发的快速匹配算法,大大提高了RMS系统的分词速度和切分准确率。详细的算法本文不作介绍。

2.2.5 检索表达式处理

检索表达式处理是搜索引擎面向用户的核心功能,如何正确理解用户需求并快速获取正确检索结果是搜索引擎的本质。用户向搜索引擎提交检索需求一般有两种方式:自然语言和查询语言,查询语言是标准化的CQL(Common Query Language)公共查询语言^[9]。对于用户提交的自然语言系统会将其切分成检索词,然后执行相邻或逻辑“与”运算。分词有两种,一种按全文检索分词方式进行分词,然后执行相邻运算,即全文检索;另一种是从用户提交的自然语言里提取主题词,然后按逻辑“与”进行组合检索;一般专业情报检索同时使用两种方式进行检索,然后根据用户需要选择检索结果,单独或混合使用。CQL检索表达式的使用者一般是专业情报检索人员,比如查新专家,使用的检索词是主题词,对检索结果要求较高。无论何种用户需求,系统解析完后还需要进行检索词的词表扩展。比如用户提交的检索词是“计算机”,系统应按词表中“计算机”词条的相关词、用代词、英文词等信息进行扩展,对于人名、机构名都有这类需求。词表扩展也称为“词表后控检索”,是专业文献检索所必须的,现有的文献服务系统都不支持这一功能,主要是搜索引擎的检索性能瓶颈所引起的,Google类搜索引擎更不具备这

一功能。RMSCloud是基于云服务架构的RMS系统的升级版本,支持这一功能(后面章节介绍)。

2.2.6 检索结果处理优化

搜索引擎一般通过索引数据库可以快速定位一个检索词并获取检索词属性列表,检索结果的获取速度主要取决于对于属性列表的处理;在一个海量数据库中,检索词的属性列表是非常庞大的,属性列表的处理涉及数据缓存技术、属性去重技术(一个记录中可能包含多个相同的检索词,但在检索结果中只能出现一个记录)等。为了提高系统的处理效率,大部分搜索引擎将检索的属性列表一次性读入内存,然后在内存进行处理,此举主要是避免多次I/O读取造成的效率瓶颈;但检索词的属性列表小者几十MB多者上百GB,仅靠扩展内存是不可行的,必须使用优化I/O读写算法才能保持系统的可扩展性(包括文件缓存和多路组配优化技术)。对于海量数据库的检索服务,检索结果的优化处理是关键。RMS系统虽然在检索词定位时速度不受数据量大小的影响,但检索词命中结果的优化处理仍然是关键,RMS对包括逻辑“与”、“或”、“非”、字段限定、相邻等逻辑运算都进行了优化处理。

2.2.7 检索结果记录输出

一般搜索引擎除了完成检索并形成命中结果外,还需要提供检索结果的输出功能;检索结果的输出包括检索结果导出、列表显示,特别为了提高输出结果的可读性,需要对结果进行排序,并格式输出每

一个记录的相关内容。检索结果的排序通常情况下有如下几种方式:

- 按时间排序: 为了让用户获取最新的文献数据, 通常倒序并分页显示, 比如期刊文章按出版日期排序; 这一功能看似简单, 但目前Google、百度等搜索引擎都没有提供这功能, 原因之一是这些搜索引擎的处理对象是网页, 没有明确的发布日期, 更重要的是搜索引擎的命中文献数量有时很大, 甚至百万级以上, 无法实现动态排序。

- 按命中匹配度排序(相似度): 目前搜索引擎的检索处理应该说只是简单的检索词匹配处理, 很难精准地获取到用户所需文献。比如: 用户请求检索式为“计算机”, 通常系统把包含“计算机”一次的文献作为命中结果; 显然“计算机”一词出现在题目字段、主题词字段、摘要字段其匹配程度是不一样的。为了让用户首先看到匹配度最高的文献, 系统需要提供按匹配度由高向低的排序功能。匹配度的计算有多种方式, 最主要的是字段加权法, 检索词出现在不同的字段其匹配度权重不一样, 词频也是一个影响因素。还有一些搜索引擎试图使用人工智能技术计算匹配度, 但结果都不理想。目前一些科技文献搜索引擎, 包括“同方知网”、“万方数据知识服务平台”都提供了相似度排序输出功能, 但显示结果都经不起仔细推敲。

- 引文量排序: 该项功能目前几乎仅用于期刊文献的检索排序输出。引文量反映了该文献被其他学者参考引用的情况, 直接反映了该文献的学术影响度。引文量排序可让用户首先看到学术上最好的文章, 是学术文献搜索服务中不可缺少的核心功能。该项功能实现的核

心不是技术, 而是要求每一篇文献记录都包括完整的参考文献信息, 文献覆盖量要尽量多, 而文献引文量是一个客观指标, 它直接影响到文献作者的学术影响力评价。

- 用户喜好排序: 该项功能旨在向用户通过搜索推荐同类用户最为喜爱的文献, 在学术搜索中也很重要; 需要系统能够完整地记录每一个用户的浏览、下载对应文献的次数, 目前大部分搜索引擎不提供此项功能。

综上所述几种排序输出算法在大部分学术搜索引擎里都能看到, 我个人认为最好的输出算法是几个排序指标的多级排序, 直接显示给用户, 无需用户选择排序算法。

3 分布式搜索引擎研制中的关键技术

上一节所介绍的搜索引擎所涉及的关键技术是针对集中式搜索引擎的, 所谓集中式搜索引擎是指搜索引擎为某一数据库系统建立的搜索引擎, 是在一台计算机上完成的、多个集中式搜索引擎可以构成分布式集群服务系统。随着搜索引擎所处理数据库数据量的快速增长, 以中国学术搜索网数据为例, 2010年包含的科技文献数据量不到1亿条记录数据, 使用的搜索引擎为RMS系统, 单机检索速度为秒级响应, RMS系统是业内公认的国产优秀情报检索系统; 到2012年年初学术搜索网的数据量已经超过2亿条记录, RMS存在的检索瓶颈暴露, 加之用户量的剧增, 全国有近20个省市科技信息研究所的文献共享保障平台直接通过万方提供的“云服务”平台, 访问RMS管理的元数据中心, 中国学术搜索网的检

索速度变得越来越慢, 不能较好地满足用户快速访问的需求。中国学术搜索网(<http://www.sciinfo.cn>)不仅仅是一个传统概念上的搜索网, 它内置了大量数据挖掘、分析、发现和数据关联的功能, 其计算量是爆炸式的。为了解决中国学术搜索网面临的种种技术瓶颈, 同时满足全国科技信息服务机构的同等需求、实现全国范围内的软件、数据资源、硬件资源及网络资源的共享服务要求、借助国家“863”重大专项“云计算及其关键技术(一期)”的课题“以科技文献为主的搜索引擎研制”课题的国家支持, 2012年初我们正式启动“RMSCloud”项目的研发工作, 同时提出“基于云服务架构的国家科技文献共享服务”总体思路, 到2012年年底“RMSCloud”搜索引擎研制基本取得成功, 并成功用于“中国学术搜索网”。下面将对“RMSCloud”研制过程中解决的关键技术给以简单介绍(RMSCloud技术的详细介绍参见本人在《情报学报》报发表的《分布式检索系统架构及核心技术研究》一文)。

3.1 Map/Reduce指导思想

Map/Reduce是Google提出的一个软件架构, 用于大规模数据集的并行运算。概念“Map(映射)”和“Reduce(化简)”以及它们的主要思想, 是目前被广泛采用的分布式云计算处理方法^[10]。Map是将一个大任务拆分为很多个小任务, 每个小任务可以交由一个虚拟机来执行完成。Reduce则将每个虚拟机的计算结果进行收集和汇总, 提炼出大任务的最终结果。就是说通过Map/Reduce方法把原来集中式系

统的一个完整数据处理流程分割到多个计算机系统中,然后将不同系统的结果进行汇总(和集中式系统处理结果一样)返回客户端。对于分布式搜索引擎而言就是将集中式搜索引擎的每一个处理步骤进行分割(Map),这一过程包括分割数据获取,并建立相应的分段索引数据库,检索时把用户检索请求Map到由系统管理的多个分布式的索引数据库进行检索处理,由于每个索引数据库都是一个较小的库,可以快速完成,理论上如果每一个Map处理是均衡的,整个系统的检索效率几乎和每一个Map的效率一致,然后就是对每一个Map结果的汇总处理。因此Google成功地利用普通PC机大规模群集,解决了海量数据分布式和并行处理问题,并在降低运营成本的同时,提高了整个系统的效率。

3.2 RMSCloud Map/Reduce服务

RMSCloud首先是把RMS构建成一个集群服务。以中国学术搜索网的期刊数据库为例,目前中国学术搜索网的期刊数据包含中文期刊5600万,英文期刊9800万。首先将该数据库分割成1000万条记录1个数据库,每个数据库分布到一台服务器上。1000万只是一个策略配置,Map/Reduce服务在建立索引数据库时会自动完成任务分发,告知各独立搜索引擎的数据获取对象及其策略,各搜索引擎将按照指令执行自己的索引数据库;当用户提交一个检索请求时,Map/Reduce服务器将把检索请求分配给对应的搜索引擎,各搜索引擎分别进行检索处理,最后将检索结果返回Map/

Reduce服务器,该服务器负责对各子检索结果的合并。Map/Reduce服务器和各服务器之间的通讯协议及其带宽对系统的Reduce效率影响较大,目前集群内服务连接一般采用TCP/IP协议,广域网协议采用Web Service(HTTP协议),以满足“云计算”架构下的分布式搜索服务需求。对于上一节中涉及的用户检索需求扩展处理、排序输出等技术大部分在Map/Reduce服务器中完成。

3.3 分布式搜索引擎的技术架构

将集中式搜索引擎构建搜索引擎集群在Map/Reduce服务器的统一管理下构成分布式搜索引擎,

是目前企业级搜索引擎通用技术架构,如图3所示。

RMSCloud完全符合上述基于“云服务”的分布式搜索引擎技术架构,RMS搜索引擎作为基础搜索引擎,Map/Reduce服务器中还包括了软硬件资源调度服务,这也是云计算的关键技术之一,由于篇幅有限,本文不作详述。

4 分布式搜索引擎的大数据应用

搜索是当今时代所有应用的基础,当下最热门的“大数据”中的各类应用的基础也是搜索。本人认为“云计算”或“云服务”解决的是技术架构问题,主要目的是充分利用各类软硬件资源,以实现真正

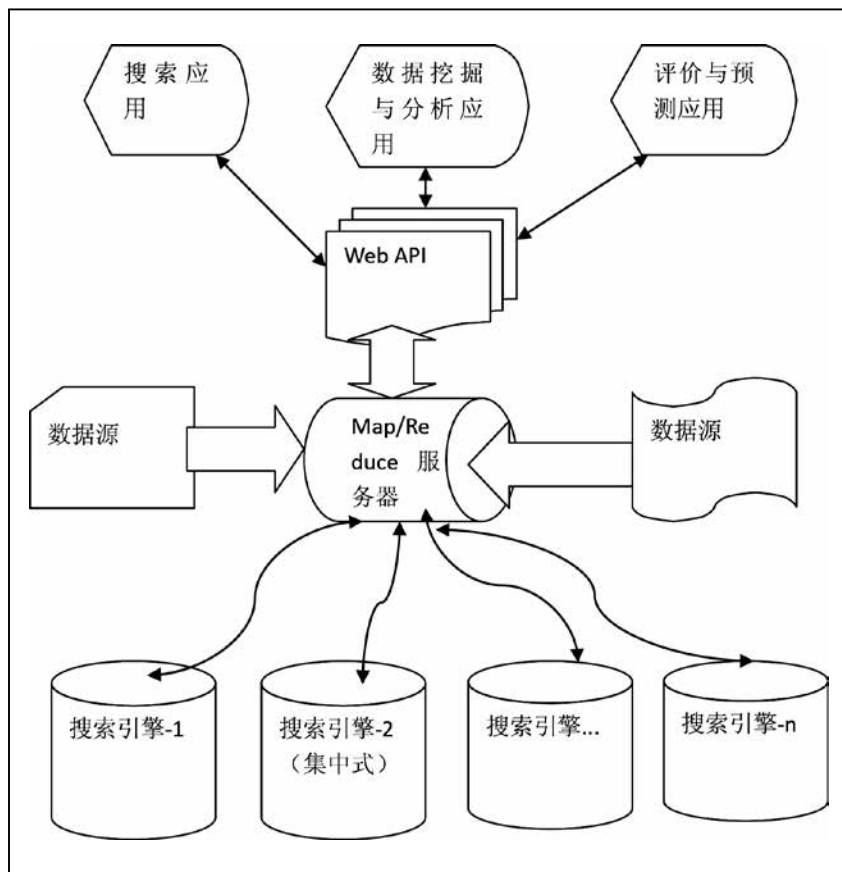


图3 通用分布式搜索引擎技术架构

意义上的资源共享;而“大数据”是以“云计算”为基础实现“大数据”各类应用,充分利用大数据,让数据说话,让决策有理有据。没有“云计算”的分布式计算,“大数据”的很多应用无法想象,一个最简单的例子就是“基于词表的扩展检索”,这里没有太多的技术难点,主要靠的是搜索引擎的处理能力。我们通常讲的“数据挖掘”、“聚类分析”、“评价、预测”的基础是获取检索命中数,然后再对数字进行处理,后者计算量较小,容易实现。RMSCloud系统研发完成后,按任务书要求建立相关的示范应用系统,目前已经投入应用的示范系统如下:

- 学术搜索,利用万方软件的中外文科技文献元数据仓储(共计5亿多条记录数据,每周更新数据量为10万级,支持公共云、交叉云和私有云混合使用)和RMSCloud系统提供的统一搜索和聚类接口,搭建完成了中国学术搜索网,2013年3月24日正式发布。系统提供了统一检索(包括导航、精确、模糊检索、词表自动扩展检索、跨语言检索等)、原文定位、基于“五要素的聚类分析”、主题发展趋势分析、机构知识仓储、个人知识仓储与创新能力分析等和数据挖掘相关的功能。另外20多家省市科技信息服务机构直接访问RMSCloud及数据仓储,构建地方科技文献共享服务

平台,地方信息机构无需购买安装大量的硬件和数据资源,是典型的“云服务”应用。系统详细功能访问<http://www.sciinfo.cn>。

- 科技创新辅助决策支持系统,简称“创新助手”。创新助手的服务模式为SaaS,用户只需下载客户端软件,通过注册支付服务费用即可。客户端软件和万方软件的仓储中心相连,如果软件的功能发生变化,系统将通知用户更新软件,系统数据资源自动更新,用户不必关心。该系统提供了“主题分析”、“人物分析”、“机构分析”、“学科分析”、“基金分析”和“科研查新”等数据挖掘和分析功能,为科研管理部门科研立项、专家评审、学科和基金科研产出分析提供了数据支撑,同时也为科研人员的开题、立项、查新等提供决策依据。使用该系统可以即时产生所需的相关分析报告,没有一个分布式搜索引擎作支撑几乎是不可能实现的。系统详细功能介绍及演示访问<http://stads.infosoftware.com/index.htm>。

- 中国机构创新能力透视,简称EIDeep。它是一款以RMSCloud+数据仓储为基础的中国机构创新能力透视系统,尤其以高校、科研机构、科研企业等事实性数据库支撑,通过强有力的机构规范名称词表为基础,利用100多个指标体系来分析机构的科研产出并进行量化评价,可以实现行业内、区域

内企业间的创新能力比对分析,包括专利、成果、标准、期刊论文等文献指标,该系统是典型的大数据应用。详细功能可访问<http://www.eideep.com>。

以上三个应用只是典型的科技文献服务的大数据应用,没用分布式搜索引擎作系统支撑是不可能的。RMSCloud这一“云计算搜索引擎”,可以在很广泛的领域内支持各类“大数据”应用,比如学科学术网络分析、产业链技术背景分析、专利关联分析、虚拟社会网络等。

结语

本文通过对搜索引擎的基础架构、关键技术全面阐述介绍集中式搜索引擎的工作原理及解决办法。集中式搜索引擎从整体架构上已经不能满足“云计算”的商业模式和“大数据”应用的发展需求,分布式搜索引擎应运而生。科技文献的搜索引擎不同于一般意义上的搜索引擎,有自身的功能需求和技术特点,在国家“863”课题“以科技文献服务为主的搜索引擎研制”支持下,我们完成了基于“云服务”的分布式搜索引擎RMSCloud的研制,可为未来科技文献系统的“大数据”应用提供自主知识产权技术支撑和保障。

参考文献

- [1] 丁蔚,倪波,等.情报检索的发展——情报学世纪回眸之一[J].情报科学,2001(1):81-86.
[2] 李冯筱,罗高松.NoSQL理论体系及应用[J].电信科学,2012(12):23-30.
[3] 曾建勋,常春.网络环境下新型《汉语主题词表》的功能定位与发展[J].情报学报,2010(6):973-977.
[4] 吴广印,胡亚莉.基于Internet的汉语后控全文检索系统的研究与开发[J].图书情报工作,2002(3):91-95.
[5] LIU Z Y, OUYANG C P, LI Y, et al. CloudMat: A Framework for Materials Web Resources Integration in the Resource Description Framework-Based Cloud [J]. Advanced Science Letters, 2012(7): 111-115.
[6] 吴广印.浅谈“查全/查准率”[J].数字图书馆论坛,2010(7):30-34.
[7] 吴广印.RMS程序员开发指南[M].北京:北京万方软件有限公司,2012.
[8] 奉国和,郑伟.国内中文自动分词技术研究综述[J].图书情报工作,2011(2):41-45.
[9] 张玲,石子夜,王胜海.WCQA:CQL的一个通用语法分析器[J].数字图书馆论坛,2006(7):19-22.
[10] MapReduce: Simplified Data Processing on Large Clusters [M/OL]. [2013-04-02]. 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

Distributed Academic Search Engine Development and Big Data Applications

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

Abstract: Influenced by the popularity of "search engine," we have been used to call "information retrieval system" in the field of book information "academic search engines." Both from the technical level and the application level, though both of them have a lot in common, they still have very big difference. Traditional centralized search engine has been unable to meet the rapid development of information explosion and the universalization of voracious demand of users, and it has become inevitable to provide "cloud services" of distributed search engines. This paper mainly includes the key technology involved in the academic search engine, the architecture of distributed search engines, as well as the main application value of distributed search engine in the field of big data. Finally it introduces the typical application of the distributed search engine RMS Cloud.

Keywords: Academic search engine, Distributed retrieval, Big data applications, Cloud services, RMS Cloud

(收稿日期: 2013-04-19)