

# Web 搜索引擎技术研究

申 健, 柴艳娜

(长安大学 教育技术与网络中心, 陕西 西安 710064)

**摘 要:**科技的进步导致了互联网中的信息以指数级速度增长。如何有效地管理和组织信息,帮助用户在海量的信息里获取有用的信息,并快速定位和索引,既是搜索引擎的目标,也是搜索引擎能够成为网络用户不可或缺的基础工具的原因。对搜索引擎技术进行了研究,讨论其内在原理和运行机制,分析其技术架构和信息抓取方法,并从工作原理上对其采用的算法和策略进行了分析。同时,对实际中 Google 搜索引擎所采用的核心技术和算法进行研究并与传统技术进行了对比,分析其所具备的先进性。另外,对搜索引擎工作流程涉及到的索引问题、SEO 等都分别进行了探讨。指出信息检索工具对于海量信息数据处理的重要性,以及在信息检索方面搜索引擎体现的优越性,它的不断发展必将带动信息科学的进步。

**关键词:**搜索引擎;蜘蛛;检索排序;SEO

**中图分类号:**TP31

**文献标识码:**A

**文章编号:**1673-629X(2016)12-0030-05

**doi:**10.3969/j.issn.1673-629X.2016.12.007

## Research on Web Search Engine Technology

SHEN Jian, CHAI Yan-na

(Education Technology and Network Center, Chang'an University, Xi'an 710064, China)

**Abstract:**Information in Internet is exponential growth with the development of science and technology. There should be a tool to help users to manage the big data effectively and get the useful information what they want, and locate and index information quickly and properly, which is the target of search engine, and why search engine has been an essential tool in daily life. The search engine technologies are researched and their internal principle and mechanism are discussed, and their technical architecture and the information retrieval are analyzed. In the working principle, the relative algorithm and strategy is studied. At the same time, the core technology and algorithm adopted by Google's search engine are studied and compared with the traditional technology, analyzing their superiority. In addition, the indexes and SEO the search engine working process involves are discussed respectively. It is pointed out that the information retrieval tools are important for huge amounts of information processing and advanced in information retrieval, the development of which will drive the progress of information science.

**Key words:**search engine; spider; index sorting; SEO

## 0 引言

全球互连网革命所引发的信息浪潮已经使互联网成为海量信息的重要来源地。搜索引擎作为互联网用户必不可少的信息获取工具,其主要作用是运用专门的策略和程序从网络上寻找、收集、提取、汇总、排序和处理信息,向用户提供数据信息检索服务和导航服务,将最终内容显示给用户的系统。经过调查,网络信息搜索在互联网服务中已经成为继 E-mail 后的第二大应用<sup>[1]</sup>。

目前,常用的搜索引擎有全文索引、目录索引、元

搜索引擎等,其中 Google、Bing、Yahoo 和 Baidu 等则是搜索引擎的代表。

## 1 搜索引擎架构

搜索引擎(Search Engines)是指在互联网环境中能够响应用户提交的搜索请求,通过已经制定好的策略和程序从互联网上搜集信息,对信息进行处理和归纳,并将检索相关的结果展示给用户的提供检索服务的系统。这类系统一般由搜集、整理和查询三个模块组成<sup>[2]</sup>。

收稿日期:2016-01-06

修回日期:2016-05-11

网络出版时间:2016-11-22

基金项目:陕西省信息化重点建设项目(2171-20120042)

作者简介:申 健(1980-),男,硕士,工程师,研究方向为计算机网络技术。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20161122.1233.050.html>

在搜索引擎的结构和执行模式的设计中,将信息检索系统内许多有价值的经验吸收进来,并且通过两种系统使用用户的不同,针对他们的特点进行了许多修改。搜索引擎系统的内容处理功能和查询功能同一般信息检索系统类似,在对繁杂数据对象的处理方面搜索引擎对系统结构进行了针对性的调整,以适应处理数据和用户查询的需要<sup>[3]</sup>。图1为搜索引擎系统架构。

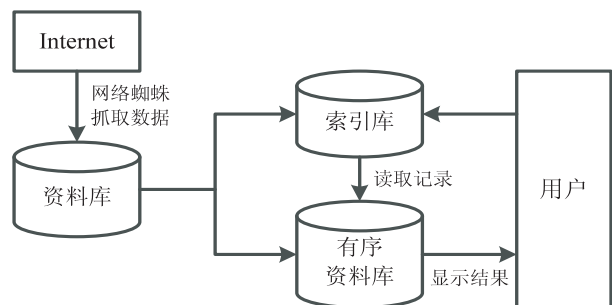


图 1 搜索引擎架构

### 1.1 搜索引擎的工作原理

搜索引擎的工作原理如图2所示。

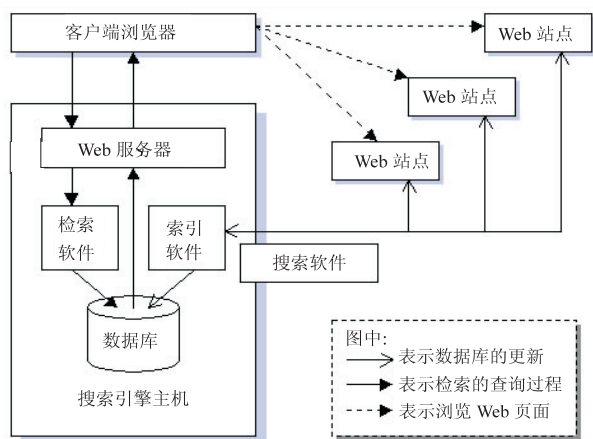


图2 搜索引擎工作原理示意图

首先,利用网络蜘蛛进行全网搜索,自动抓取网页;其次,对获取到的网页信息进行索引,同时记录与检索有关的信息(如果是中文搜索引擎就需要中文分词);最后,接收用户查询请求,按照设定好的参数对索引文件进行计算,并将结果向用户显示。简单概括为:抓取网页→建立索引数据库→在数据库中排序→结果反馈。搜索引擎抓取数据与分析过程如图3所示。

## 1.2 网络蜘蛛

它是按照一定的规则,自动抓取万维网信息的程序或者脚本的半自动化资源获取方式<sup>[4]</sup>,因为尚未对获取的数据进行处理,所以只能称作是一种半自动化的资源而不是信息。半自动化是指需要人工指定起始网络资源(Uniform Resource Locator, URL)进行搜索,并按照URL指向获取网络资源,然后分析、获

取与该资源有关的所有其他资源。例如 Google, 它利用蜘蛛程序获取资源, 先由一个管理程序进行任务分配并处理结果, 然后由多个分布式的蜘蛛程序接受任务, 最后将获取的资源作为结果返回, 再重新获得任务。

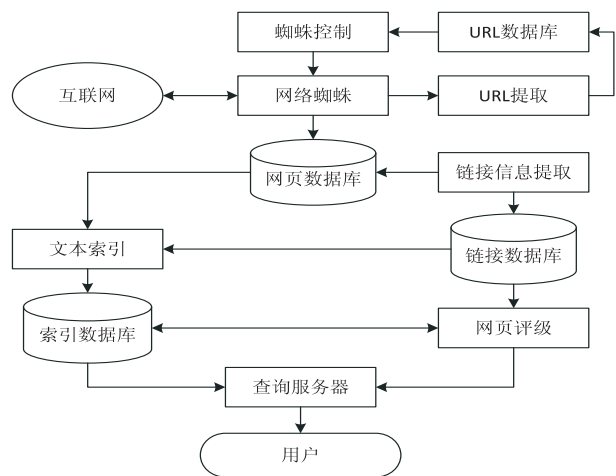


图3 搜索引擎的数据与分析过程

搜索引擎的蜘蛛抓取网页有一定的规律,主要有以下几种策略:

(1)深度优先搜索策略。网络蜘蛛通过页面发现的一个链接,顺着链接的页面又发现一个链接,并且将发现的页面全部抓取。

(2) 宽度优先搜索策略。先搜索完一个 Web 页面中所有的超级链接,然后再继续搜索下一层,直到底层为止并进行抓取。

(3) 权重优先策略。即深度优先+宽度优先。参照链接的权重进行网络抓取,对权重高的链接采用深度优先策略,而对权重低的链接则采用宽度优先策略。也就是综合层次的多与少以及这个链接的外链多少与质量等因素获取链接的权重。

(4)重访抓取策略。包括全部重访与单个重访。

### 1.3 建立索引

建立索引数据库的过程是利用索引器从搜索器搜索到的资源中抽取信息,建立检索所需的索引表<sup>[5]</sup>。

通常情况下,网络蜘蛛捕获的资源需要去掉控制代码和其他不相关信息,提取有用信息并通过模型将信息表示出来,这样能够使查询结果更为准确。就像网页上的信息是以 Web 形式进行表现,在查询结果的页面中网页要生成摘要,摘要会向用户显示网页的大概内容,并将模型化的信息存放在临时数据库中。网页上的数据量非常巨大,为提高检索效率,搜索引擎会按照设定好的规则对资源建立索引。不同的搜索引擎会分别按照全文索引、无用词汇过滤,或者根据 meta 信息建立索引。在该过程中,需要进行的资源分析处理可概括为以下几个方面:

- (1)网页结构化。即将 html 代码全部删掉,提取出内容。
- (2)消噪。留下网页的主题内容,删掉没用的内容。
- (3)查重。由搜索引擎查找重复的网页与内容,如果找到重复的页面与内容,即删除。
- (4)分词。提取出正文的内容,将其分成  $N$  个词语,并排列出来,存入索引库,同时计算该词在页面出现的频率。
- (5)链接分析。分析页面的反向链接数、导出链接数以及内链数,然后链接加上权重等。
- (6)用户查询(Query)解析。最大可能地分析出用户想要表达的查询目的,然后将用户的需求转化成信息模型供数据库检索使用;根据用户的需求模型,在索引数据库找出结果;对结果进行排序。由于 Web 数据的内容量大、结果模糊性高,检索结果通常很多,如何将用户感兴趣的结果排在前面去设计结果集的排序算法十分重要。

2 搜索索引

搜索索引的核心结构是倒排索引,如图 4 所示。倒排索引实际应用中需要根据非主属性(也叫副键)值来查找记录,其特殊性在于不是由记录来确定属性值,而是由属性值来确定记录的位置,带有倒排索引的文件称作倒排文件,即次索引。倒排索引是以

文档的关键词作为索引(就像普通书籍中索引是关键词,页面符号是目标),文档就是索引目标的一种结构。

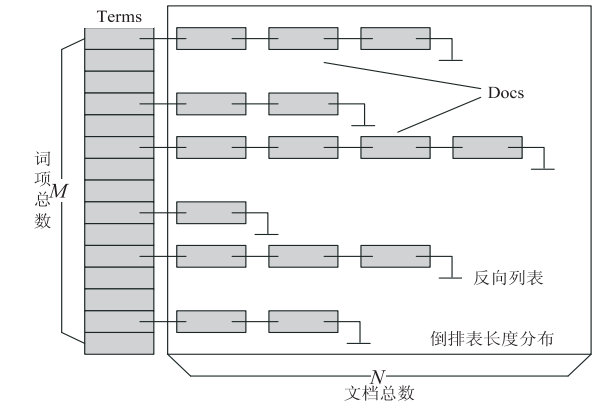


图 4 倒排索引

倒排索引包括所有副键值,并列出相关的记录主键值,它主要应用于复杂查询。与通常的结构化查询语言(SQL)的差别在于,搜索引擎收集完数据后在预处理的步骤,通常利用高效的数据结构来提供检索服务,而现阶段“倒排索引”就是效率最高的数据结构。

2.1 构建索引

1)简单法。

构建索引就是从正排表到倒排表的建立过程。首先对网页进行分析,建立以网页为主码的索引表<sup>[6]</sup>;其次在索引建立完成后得到倒排表。构建倒排索引的具体流程如图 5 所示。

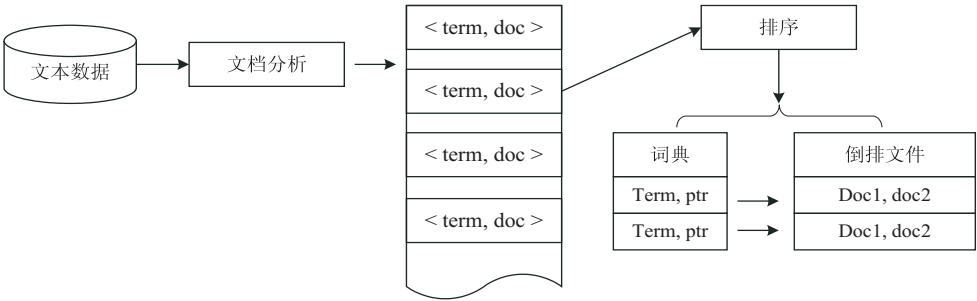


图 5 倒排索引构建示意图

流程描述如下:  
(1)将文档分析用 term 标记;  
(2)利用 hash 去重单词 term;  
(3)生成倒排列表。  
倒排列表就是文档编号 DocID,不包含其他信息(词语的频率、位置等),这就是简单索引。简单索引的功能可以用在数据量小的内容,例如对几千个文档进行索引。不过它有两点限制:

- (1)需要足够大的内存空间存储倒排表。对于搜索引擎来说,都是以 G 为单位的数据量,在其规模不断扩大的同时不能确保内存的空间能够得到相应的增长。 万方数据

(2)算法是按照一定顺序来执行,对于并行处理造成不便。

2)合并法。

即归并法。每一次内存数据在写入磁盘的时候,包括词典在内的所有中间结果信息都被写入磁盘,这样内存中的所有内容都被清空,之后建立的索引可以使用全部的内存空间。合并流程如下:

- (1)页面分析。首先生成临时倒排数据索引 A 和 B,一旦索引 A 和 B 占满内存空间后,将索引 A 和 B 写入临时文件来生成临时倒排文件。
- (2)多路归并。对已经生成的临时文件来执行多路归并,得到最终的倒排文件(inverted file)。

在创建索引的过程中,页面分析特别是中文分词是消耗时间的主要步骤,而第二步就快得多了,对创建算法进行优化重点在于提高中文分词的效率。

2.2 更新策略

包括四个方面:完全重建策略、再合并策略、原地更新策略以及混合策略。

(1)完全重建策略。一旦新增文档满足一定数量标准,就对新增文档和原文档进行整合,再对生成的文档创建静态索引,保留新建索引并删除原索引。此法代价高,但是主流商业搜索引擎一般采用此方式来维护索引的更新<sup>[7]</sup>。

(2)再合并策略。对进入系统的新文档进行解析,更新内存中保留的临时索引,在文档中每个单词的倒排列表末尾追加倒排列表项;当临时索引占消耗完指定内存后,进行索引合并,这里需要倒排文件里的倒排列表存放顺序是按照索引单词字典顺序由低到高排序,这样按顺序可以直接扫描合并。其缺点是:在生成新的倒排索引文件时,会将老索引倒排列表中很多未发生变化的单词也取出并写入新索引中,这样增加了对磁盘的消耗。

(3)原地更新策略。基本出发点,可以认为是试图改进再合并策略的缺点,在原地合并倒排表,这需要预留空间给未来插入,如果预留的空间不够就要进行迁移。迁移的过程中会破坏老索引中某些单词的连续性,不能顺序进行读取,并且需要足够大的磁盘连续存储。实际操作中表明,其原地更新的效率比再合并策略要低。

(4)混合策略:其目地是将不同策略的优势结合到一起,混合其他索引更新策略,形成一种更加高效的方法。

3 Google 搜索引擎

3.1 Google 技术

“完美的搜索引擎”是 Google 坚持的开发目标。正如公司创始人之一 Larry Page 所定义的那样,可以“确解用户之意,切返用户之需”。为了能够达到这个目标,Google 坚持“不受现有模型限制,不断追求创新”,通过开发具有自身特色和突破性的服务基础结构和 Page Rank 技术,从而根本性地改变基于互联网的信息搜索方式。

为此,Google 开发人员采用了一种全新的服务器设置,利用相互链接的 PC 来快速查找每个搜索答案,以最快的速度为用户提供最精确的搜索结果的设计理念,从而避免了因使用少量大型服务器导致搜索引擎在访问高峰期相应速度会减慢的缺陷。应用这种技术能够降低成本、缩短响应时间、提高可扩展性。与此同时,

Google 对其内部技术的持续改进使得该技术的效率得到不断提升。

Google 搜索技术的特点是利用的软件能够在同一时间进行一系列运算,且都能在很短的时间内完成;Page Rank 技术通过对整个网络链接进行检查,依据每个网页的重要性进行排序;进行超文本匹配分析,判断出预指定搜索有关联的网页;综合考虑特定查询与整体重要性的相关性,将关系最密切并且可靠性最强的结果放在首位。与此不同的是,普通搜索引擎一般都是以网页上文字的出现频率高低作为排序的重要依据。

3.2 Google 搜索关键技术

1)Page Rank 技术。

Page Rank(网页排名)是根据网页之间相互的超链接计算的技术,让链接来“投票”。其特点是:

(1)不计算直接链接的数量,而是将从网页 A 指向网页 B 的链接解释为由网页 A 对网页 B 所投的一票<sup>[8]</sup>,页面的超链接就表示对该页面投一票,页面的重要性由它的“得票数”来决定;

(2)通过对投票价值的评估,拥有较高投票价值的网页可以获得较高的评价;

(3)重要网页的网页排名高,显示在搜索结果的较高处;

(4)利用反馈的综合信息确定单个网页的重要性;

(5)没有人因素干扰到搜索结果。

Google 能够成为一个公正的、得到用户信任的、不受付费排名影响的客观信息来源,这个技术起到了重要的推动作用。

Google Page Rank 技术的 PR 值算法如式(1)所示<sup>[9]</sup>:

$$PR(A) = \left( \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots + \frac{PR(N)}{L(N)} \right) q + 1 - q \tag{1}$$

其中,PR(A)指网页 A 的佩奇等级(PR 值);PR(B),PR(C),...,PR(N)表示链接网页 A 的网页 N 的佩奇等级(PR);N 是链接总数,这个链接可以是来自任何网站的导入链接(反向链接);L(N)是网页 N 往其他网站链接的数量(网页 N 的导出链接数量);q 是阻尼系数,介于 0~1 之间,Google 设为 0.85<sup>[10]</sup>。

2)超文本匹配分析。

Google 的搜索引擎也对网页文本内容进行分析。它并不仅仅只局限于网页文本的扫描方式,还对包括本网页和相邻网页的字体、分区和文字精确位置等等内容进行分析,以确保向用户反馈查询最匹配的



结果<sup>[11]</sup>。

对于通过便携式终端访问网络的用户,Google 推出了行业内第一款无线搜索技术,将 HTML 即时转换为针对 WAP、I-mode、J-SKY 和 EZ Web 优化的格式<sup>[12]</sup>,保障用户能够快速获得精确的搜索结果,这是一项并不限于台式机的创新。

### 3) 查询的全过程。

Google 查询过程需要在短时间内(一般不超过 0.5 s)完成多个步骤,而后将搜索结果向用户显示。

(1) 服务器将查询内容发送给索引服务器。索引服务器所包含的内容与索引目录相似,即显示与查询内容匹配的都有哪些网页。

(2) 查询内容传输到文档服务器,后者检索存储的文档,然后生成描述结果的摘录。

(3) 返回用户需要的搜索结果。

## 4 SEO 优化

SEO(Search Engine Optimization, 搜索引擎优化)是指在了解搜索引擎自然排名机制的基础上,利用搜索引擎的搜索规则来提高目前网站在有关搜索引擎内的自然排名,以获得更多流量,实现网络营销及品牌建设的目标<sup>[13]</sup>。它能够使网站更适合搜索引擎的索引原则,这样不仅在用户面前提高了搜索引擎的效果,还会使显示的网站相关信息对用户来说更具有吸引力。

搜索引擎 SEO 的搜索方法如图 6 所示。

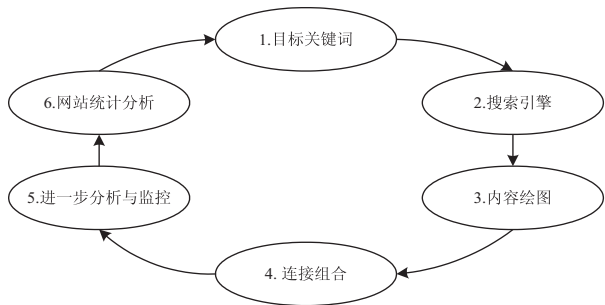


图 6 SEO 搜索方法

## 5 结束语

Internet 提供了多种不同的检索工具,它们有各自的语言、数据库、检索功能和显示方式,对于用户来说

了解这些工具的性能,掌握检索技巧,提高检索命中率是最重要的。掌握了方法与技巧并且经常进行实践操作,就能够方便快捷地利用搜索引擎获取更多符合需求的有价值的信息。

目前,搜索引擎在扩大覆盖范围的同时,正在趋向个性化、智能化、专业化、多媒体、多语言搜索和实用性的模糊检索方面发展,并已取得了较大技术进步。随着需求的提高和互联网技术的发展,不断应用新的技术和策略,搜索将会向着更加方便、快速、准确的目标前进,这已成为搜索引擎的发展方向<sup>[14]</sup>。

### 参考文献:

- [1] 梁 斌. 走进搜索引擎[M]. 北京:电子工业出版社,2007.
- [2] 吴泽欣. 搜索引擎优化入门与进阶[M]. 北京:人民邮电出版社,2008.
- [3] 卢 亮. 搜索引擎原理、实践与应用[M]. 北京:电子工业出版社,2007.
- [4] Lawrence S, Giles C L. Accessibility of information on the web[J]. Nature, 1999, 400(6740): 107-109.
- [5] Lawrence S, Giles C L. Searching the World Wide Web[J]. Journal of the American Society for Information Science & Technology, 1998, 280(1): 8-14.
- [6] 张园园. 基于用户兴趣的个性化搜索引擎的分析与研究[D]. 秦皇岛:燕山大学,2006.
- [7] 王 涛. 基于行业的个性化搜索引擎的应用[D]. 北京:北方工业大学,2008.
- [8] Vazirgiannis M, Drosos D, Vlachou A, et al. Web page rank prediction with Markov models[C]//WWW 2008. Beijing, China: ACM, 2008.
- [9] Wills R S. Google's page rank: the math behind the search engine[J]. The Mathematical Intelligencer, 2006, 28(4): 6-11.
- [10] Lo S. 全球最强搜索引擎谷歌 Google[M]. 上海:上海财经大学出版社,2007.
- [11] 林 中. Google 搜索引擎的关键词检索[J]. 中国信息导报, 2003(3): 60-61.
- [12] 陈 钢. 搜索引擎优化宝典[M]. 北京:清华大学出版社, 2009.
- [13] 周元兴. Google 入门与实例教程[M]. 北京:电子工业出版社, 2007.
- [14] 万胜林, 王祖荣. 搜索引擎的类型及其功能分析[J]. 中国信息导报, 2003(5): 52-54.