搜索引擎技术分析与研究

Technology analyse and research about search engine

陈　丹　（浙江大学城市学院 杭州 310015）
郭伟青　（浙江工业大学之江学院学院 杭州 310024）

摘　要： 互联网技术的发展，使得互联网上的信息量急剧增加，越来越多的网络用户依靠搜索引擎技术，从浩瀚的信息海洋中获取信息。使得搜索引擎蕴涵着巨大的商业价值，甚至出现了“搜索力经济”的说法，目前搜索引擎技术研究在国内外发展得非常迅速。本文详细介绍分析了在这一领域多种新技术。

关键词： 互联网 搜索引擎 新技术 信息

1　搜索引擎的系统架构

搜索引擎的技术门槛包括网页数据的快速采集、海量数据的索引和存储、搜索结果的相关性排序、搜索效率的毫秒级要求、分布式处理和负载均衡、自然语言的理解技术等等，这些都是搜索引擎的门槛。对于一个复杂的系统来说，各方面的技术固然重要，但整个系统的架构设计也同样不可忽视，搜索引擎也不例外。

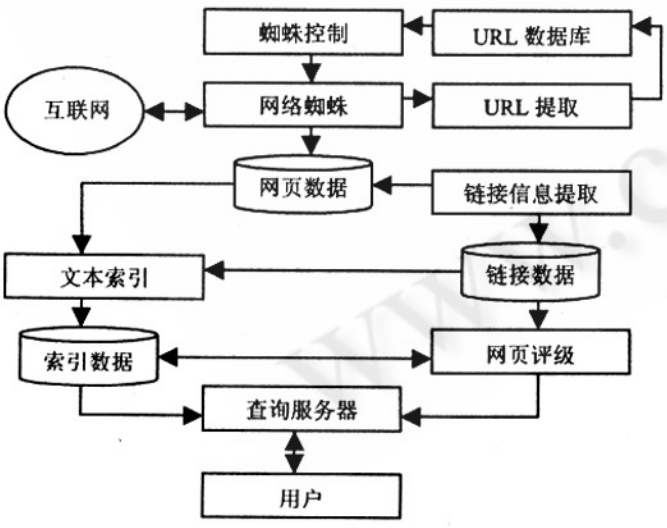

图 1　搜索引擎系统架构图

搜索引擎的实现原理，可以看作四步：从互联网上抓取网页→建立索引数据库→在索引数据库中搜索→对搜索结果进行处理和排序。图 1 是搜索引擎的系统架构图，搜索引擎的各部分都会相互交错相互依赖。

2　搜索器的技术分析

衡量一个搜索引擎很重要的一条标准，就是其搜索信息的“海量”，要保证不遗漏某些重要的结果，而且能找到最新的网页，这需要搜索引擎有一个强大的搜索器，一般称为“网络蜘蛛”，也叫“网页机器人”。

(1) 网络蜘蛛基本原理

网络蜘蛛是通过网页的链接地址来寻找网页，从网站某一个页面（通常是首页）开始，读取网页的内容，找到在网页中的其它链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去，直到把这个网站所有的网页都抓取完为止。如果把整个互联网当成一个网站，那么网络蜘蛛就可以用这个原理把互联网上所有的网页都抓取下来。在抓取网页的时候，网络蜘蛛一般有两种策略：广度优先和深度优先。

(2) 内容提取

对于网页内容的提取，一直是网络蜘蛛中重要的技术。整个系统一般采用插件的形式，通过一个插件管理服务程序，遇到不同格式的网页采用不同的插件处理。搜索引擎建立索引，处理的对象是文本文件。对于网络蜘蛛来说，抓取下来网页包括各种格

式，如 html、图片、Doc、Pdf、多媒体、动态网页及其它格式等。这些文件抓取下来后，需要把这些文件中的文本信息提取出来。

对于 Doc、Pdf 等文档，这种由专业厂商提供的软件生成的文档，厂商都会提供相应的文本提取接口。网络蜘蛛只需要调用这些插件的接口，就可以轻松的提取文档中的文本信息和文件其它相关的信息。HTML 等文档不一样，HTML 有一套自己的语法，通过不同的命令标识符来表示不同的字体、颜色、位置等版式。提取文本信息时需要把这些标识符都过滤掉。同时，对于 HTML 网页来说，会有许多广告链接以及公共的频道链接，也需要过滤掉。对于多媒体、图片等文件，一般是通过链接的锚文本（即，链接文本）和相关的文件注释来判断这些文件的内容。

Term1	Doc_i	位置1	位置2	Doc_j	位置1	-------	Doc_m	位置1	位置2
Term2	Doc_j	位置1	位置2	Doc_k	位置1	-------	Doc_n	位置1	位置2
Term3	Doc_j	位置1	位置2	Doc_k	位置1	-------	Doc_n	位置1	位置2

图 2　典型的倒排索引

动态网页一直是网络蜘蛛面临的难题。由于开发语言不断的增多，动态网页的类型也越来越多，如：asp、jsp、php 等。这些类型的网页对于网络蜘蛛来说，可能还稍微容易一些。网络蜘蛛比较难于处理的是一些脚本语言（如 VBScript 和 JavaScript）生成的网页，如果要完善地处理好这些网页，网络蜘蛛需要有自己的脚本解释程序。对于许多数据是放在数据库的网站，需要通过本网站的数据库搜索才能获得信息，这些给网络蜘蛛的抓取带来很大的困难。对于这类网站，需要网站设计者提供一定的方法协助网络蜘蛛收集数据。

（3）定期更新策略

由于网站的内容经常在变化，因此网络蜘蛛也需不断的更新其抓取网页的内容，这就需要网络蜘蛛按照一定的周期去扫描网站，查看哪些页面是需要更新的页面，哪些页面是新增页面，哪些页面是已经过期的死链接。

3　索引器的技术分析

索引器的功能是分析和理解搜索器所搜索的信息，从中抽取出具有代表意义的索引项，建立相应的索引库。

索引项有客观索引项和内容索引项两种：客观项与文档的语意内容无关，如作者名、URL、更新时间、编码、长度、链接流行度（Link Popularity）等等；内容索引项是用来反映文档内容的，如关键词及其权重、短语、单字等等。内容索引项可以分为单索引项和多索引项（或称短语索引项）两种。单索引项对于英文来讲是英语单词，比较容易提取，因为单词之间有天然的分隔符（空格）；对于中文等连续书写的语言，必须进行词语的切分（分词）。

索引库的结构一般采用倒排索引（Inverted Indexing），即由索引项（Term）定位相应的文档，如图 2 所示。相关信息一般包括‘索引项”、“索引项所在文件位置信息”以及“索引项权重’，例如，索引项“Term1”的位置信息为“文档 Doci 中第位置1”。建立索引库的准则是：易于更新、检索速度快。

索引器可以使用集中式索引算法或分布式索引算法。当数据量很大时，必须实现即时索引（Instant Indexing）或称增量式索引（incremental indexing），否则不能够跟上信息量急剧增加的速度。索引算法对索引器的性能（如大规模峰值查询时的响应速度）有很大的影响。一个搜索引擎的有效性在很大程度上取决于索引的质量。

4　检索器的技术分析

检索器的功能是针对用户的查询请求在索引库中快速检出文档，采用一定的信息检索模型进行文档与查询的相关度评价，对将要输出的结果进行排序、聚类等操作，并实现某种用户相关性反馈机制。

信息检索模型有以下几种:布尔逻辑模型、模糊逻辑模型、向量空间模型以及概率模型、混合模型等。

布尔逻辑模型,是最简单的信息检索模型,用户可以根据检索项在文档中的布尔逻辑关系提交查询,搜索引擎根据事先建立的倒排文件结构,确定查询结果。标准布尔逻辑模型为二元逻辑,所搜索的文档要么与查询相关,要么与查询无关。查询结果一般不进行相关性排序。如查询"计算机",只要文档中出现关键词"计算机",则全部包含在查询结果中。为了克服布尔型信息检索模型查询结果的无序性,在查询结果处理中引进了模糊逻辑运算,将所检索的数据库文档信息与用户的查询要求进行模糊逻辑比较,按照相关的优先次序排列查询结果。例如,查询"计算机",那么出现"计算机"较多的文档将排列在狡前的位置。

康奈尔大学的 Salton 等人提出的向量空间模型将查询和文本表示成标引项及其权重的向量。一个例子是:<信息,3,检索,5,模型 1>,然后通过向量之间的相似度比较来计算每个文本的相似程度。最典型的向量空间模型原型系统是康奈尔大学的 SMART,它提供源代码开放下载,目前已经被成千上万的研究者所用。

概率检索模型是通过概率的方法将查询和文本联系起来。最经典的概率检索模型是英国伦敦城市大学的 Robertson 和剑桥大学的 Sparck Jones 提出的二元独立概率模型(Binary Independence Retrieval, BIR)。它主要通过计算查询词中每个标引项和文本的相关概率来计算整个查询和文本的概率。

混合模型又称为扩展的布尔模型(Extended Boolean Model)。该模型与向量空间模型一样,将文档表示为 n 维空间中的向量,不同的是它用两个向量之间一般化的标量乘积衡量文档和查询的相似度。随着一般化参数的变化,该模型可以变为向量空间模型、模糊集合模型和布尔模型。

虽然不同检索模型使用的方法不同,但所要达到的目标是相同的,即按照用户要求,提供用户所需的信息。实际上,大多数检索系统往往将上述各种模型混合在一起,以达到最佳的检索效果。

5 词法分析技术

词法分析是对自然语言的形态进行分析,判定词的结构、类别和性质的过程。对于以英文为代表的形态丰富的语言来说,英文的词法分析的一个重要过程是形态分析,即将英文词还原成词干。而汉语形态变化很少,其主要的问题在于书写时词与词之间没有空格。所以通常中文词法分析的关键是分词,分词往往是后续进一步处理的基础。

(1) 英文形态分析

英语的词常常由前缀、词根、后缀等部分组成。具体到句子中,词还有性、数、格以及时态引起的词形变化。英文的形态分析的主要目标是将句子中的词从词形还原到词甚至词根。英文的形态分析常常也称为 stemming,分析器称为 stemmer。形态分析常常采用基于自动机的规则方法,即将词形变化的规律总结成规则,然后通过自动机的方法对词形进行转换。转换的过程当中可使用或者不使用词典。目前使用最广泛的 Stemmer 是 Martin Porter 提出的 Porter Stemmer。

(2) 中文分词技术

中文分词方法可以总结为两大类:基于机械匹配和基于概率统计的分词方法。前者通过对已有词典的机械匹配来得到分词结果。后者不需要任何词典就可以得到分词结果,或者对粗切分结果进行基于概率统计的后处理来得到最终的分词结果。

中文分词技术面临的两个最大问题是切分歧义和未定义词问题。前者要解决在上下文环境下不同切分结果的选择;后者要解决词典中未收录词(如人名、地名、机构名等)的识别。可以在机械匹配的基础上通过规则的方法来求解上述两个问题。然而规则方法很难穷尽真实文本的各种现象。目前比较主流的方法是通过对真实文本的概率统计来求解切分歧义和未定义词问题。一些研究(如微软研究院)将中文分词的一部分歧义问题延到后续句法分析阶段,利用更加丰富的信息加以解决并进行反馈,实现了基于这一新思路的分词系统。

6 链接分析技术

WEB 搜索中各页面之间的链接关系是一项可以利用的重要信息。基于这种信息的技术被称为链接分析技术。绝大部分链接分析算法都有共同的出发点:更多地被其他页面链接的页面是质量更好的页面,并且从更重要的页面出发的链接有更大的权重。这个循环定义可以通过迭代算法巧妙打破。最著名的链接分析算法是 Stanford 大学提出并应用到 Google 搜索引擎中的 PageRank 算法以及 IBM 用于 CLEVER 搜索引擎的 HITS 算法。

(1) PageRank 算法

PageRank,有效地利用了 Web 所拥有的庞大链接构造的特性。从网页 A 指向网页 B 的链接被看作是页面 A 对页面 B 的支持投票,Google 根据这个投票数来判断页面的重要性。可是 Google 不单单只看投票数(即链接数),对投票的页面也进行分析。"重要性"高的页面所投的票的评价会更高。根据这样的分析,得到了高评价的重要页面会被给予较高的 Page Rank(网页等级),在检索结果内的名次也会提高。

该分值的计算过程是一个迭代过程,最终网页将依照所得的分数进行排序并将检索结果送交用户,这个量化了的分数就是 PageRank 值,其计算公式如下:

其中 PR(A)是网页的页面级别,d 为界于(0 ,1)区间的衰减系数,一般取 0 .85 左右,T1、T2. . . Tn 为指向网页 A 的其它网页,C(Tn)是网页 Tn 中向外指出的链接数目。

(2) HITS

HITS 是 IBM Almaden 研究中心开发的另一种链接分析算法。它认为每个 WEB 页面都有被指向、作为权威(Authority)和指向其他页面作为资源中心(Hub)的两方面属性,其取值分别用 A(p)和 H(p)表示。A(p)值为所有指向 p 的页面 q 的中心权重 H(q)之和,同样,页面 p 的中心权重 H(p)值是所有 p 所指向的页面 q 的权威权重 A(q)之和,如下式:

(其中 qi 是所有连接到 p 的页面)

(其中 qi 是所有页面 p 所链接到的页面)

7 小结

搜索引擎已成为一个新的研究、开发领域。因为它要用到信息检索、人工智能、计算机网络、分布式处理、数据库、数据挖掘、数字图书馆、自然语言处理等多领域的理论和技术,所以具有综合性和挑战性。同时搜索引擎有大量的用户,有很好的经济价值,引起了越来越多研究单位、厂商的关注,目前的研究、开发十分活跃,新技术也不断出现,特别是分布式体系结构将会是今后的一个很好的发展方向,我们也会在这方面加以分析研究。

参考文献

1 李孝明、曹万华,文本信息检索的精确匹配模型,计算机科学,2004 ,31(9):100 - 102.

2 张彦、邵志清,具有概念联想功能的特定领域分词词典的自动构建,计算机工程,2004,30(20):148 - 150.

3 Inmon W H. Building the Data Warehouse Second Edition,机械工业出版社,2000.

4 孙晋文、肖建国,自动文本分类中的智能处理技术,计算机科学,2003,30(8):18 - 20.

5 网页搜索引擎,[2007 - 07 - 7]. http://wiki. huihoo. com/index. php? title = Search_Engine_Technology/