

搜索引擎百度与 Google 的比较分析

张 岚

(沈阳工程学院图书馆,辽宁沈阳,110136)

摘 要:介绍了搜索引擎百度与 Google 的现状,分析了搜索引擎的工作原理及其核心技术——超链分析技术,并比较了百度与 Google 搜索功能的异同点。

关键词:搜索引擎;百度;Google;超链分析

中图分类号:TP393.02 **文献标识码:**A

1 百度与 Google 简介

百度 1999 年底成立于美国硅谷,创始人是北京大学的两位毕业生李彦宏和徐勇,“百度”一词来源于辛弃疾脍炙人口的词句“众里寻他千百度”,象征着百度对中文信息检索技术执著的追求,百度现已成为全球最优秀的中文信息检索与传递技术供应商。

百度搜索引擎是目前世界上规模最大的中文搜索引擎,可供搜索的中文网页已达 8 亿个之多,每天处理来自 100 多个国家的超过 1 亿人次的搜索请求,其流量居全球中文网站的首位,在全球排名居第 5 位。在中国所有提供搜索引擎的门户网站中,超过 80% 以上都由百度提供搜索引擎技术支持,现有客户包括新浪、263、腾讯、21cn、上海热线、新华网等。

从信息量上看,8 亿网页的抓取量大约相当于整个因特网上中文网页总数的 1/3。如果网民 1 min 可以看完一个网页的话,那么 8 亿个网页需要不吃不喝不睡阅读 1 500 年,如果用普通的 A4 纸打印的话,厚度将超过 40 km。

Google 成立于 1998 年,创始人是美国斯坦福大学的两位博士研究生 Larry Page 和 Sergey Brin。Google 一词由英文单词“googol”变化而来,表示 1 后边带有 100 个零的数字,使用这个词显示了 Google 欲整合网上海量信息的远大目标。Google 被公认为全球最大的搜索引擎,目前 Google 可搜索的网页高达 80 亿个之多,其中中文网页约 5 亿多个,图片 10

一样的。用户首先可以下载或者购买数字签名软件,然后安装在个人电脑上。在产生密钥对后,软件自动向外界传送公开密钥。由于公共密钥的存储需要,所以需要建立一个鉴定中心(CA)完成个人信息及其密钥的确定工作。鉴定中心是一个政府参与管理的第三方成员,以便保证信息的安全和集中管理。用户在获取公开密钥时,首先向鉴定中心请求数字确认,鉴定中心确认用户身份后,发出数字确认,同时鉴定中心向数据库发送确认信息。然后用户使用私有密钥对所传信息签名,保证信息的完整性、真实性,也使发送方无法否认信息的发送,之后发向接收方;接收方接收到信息后,使用公开密钥确认数字签名,进入数据库检查用户确认信息的状况和可信度;最后数据库向接收方返回用户确认状态信息。不过,在使用这种技术时,签名者必须注意保护好私有密钥,因为它是公开密钥体系安全的重要基础。如果密钥丢失,应该立即报告鉴定中心取

亿个,Google 界面可用的语言有 100 多种,Google 搜索结果所采用的语言也有 35 种。

2 搜索引擎的工作原理

搜索引擎是一个提供“信息检索”服务的网站,它使用某些程序把因特网上的所有信息归类以帮助人们在茫茫网海中搜寻到所需要的信息,一般由以下 3 个部分构成:

(1)在因特网上抓取网页。专门用于检索信息的机器人程序如蜘蛛程序在网络间爬来爬去,自动收集、访问网页,并沿着网页中的所有 URL 爬到其他网页,蜘蛛程序不断重复这一过程,把到过的所有网页收集回来,随着因特网呈几何级数的迅速发展,使得检索所有新出现的网页变得越来越困难,传统的蜘蛛程序工作原理发生了一些改变:既然所有网页都可能存在连向其他网站的链接,那么从一个网站开始,跟踪所有网页上的所有链接,就有可能检索整个因特网。

(2)建立索引数据库。由分析索引系统程序对收集回来的网页进行分析,提取相关网页信息,包括网页所在 URL、编码类型、页面内容包含的所有关键词、关键词位置、生成时间、大小、与其他网页的链接关系等,根据一定的相关度算法进行大量复杂计算,得到每一个网页针对页面文字中及超链中每一个关键词的相关度或重要性,然后用这些相关信息建立网页索引数据库。

消认证,将其列入确认取消列表之中。其次,鉴定中心必须能够迅速确认用户的身份及其密钥的关系。一旦接收到用户请求,鉴定中心要立即认证信息的安全性并返回信息。

Internet 的迅猛发展使电子商务成为商务活动的新模式,电子商务正在被越来越多的人所认识和接受,数字签名在电子商务活动中扮演着重要的角色,随着电子商务的蓬勃发展,数字签名技术也将不断成熟,为商务活动和人们的生活提供可靠、便利的服务。

(责任编辑:薛培荣)

第一作者简介:鱼双健,男,1978 年 8 月生,2003 年毕业于兰州交通大学,助理工程师,中铁建设集团有限公司,北京市丰台区吴家村路 1 号,100040。

Detailed Description of the Digital Signature

YU Shuang-jian

ABSTRACT:The digital signature, which is a new electronic signature technology, plays an important role in the commercial activities. This paper introduces the concept, principle, algorithm, function and application of the digital signature.

KEY WORDS:digital signature; encryption technique; EDI

(3)在索引数据库中搜索排序。当用户输入搜索关键词后,由搜索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。因为所有相关网页针对该关键词的相关度早已算好,所以只需按照现成的相关度数值排序,相关度越高,排名越靠前。最后,由页面生成系统将搜索结果链接地址和页面内容、摘要等信息组织起来返回给用户。

3 百度与 Google 的核心技术

百度与 Google 的核心技术是超链分析技术,超链分析是新一代搜索引擎的关键技术,已为世界各大搜索引擎普遍采用,据称百度总裁李彦宏是超链分析专利的唯一持有人。

超链分析技术的基本原理充分考虑了网页与网页之间的链接,提出了同引与同被引的概念。假设有 3 篇论文,论文 A、论文 B 和论文 C,如果论文 B 和论文 C 同时引用了论文 A,直观上看,论文 A 的权威性要比论文 B 和论文 C 高,这叫同引;如果论文 A 同时引用了论文 B 和论文 C,那么我们可以说论文 B 和论文 C 所探讨的主题内容有相关关系,这叫同被引。将同引与同被引的概念引入网页分析,推导计算出中心页和权威页,可以提高检索结果的准确度。

百度和 Google 这类基于超链分析的搜索引擎的基本思想即是:在某次搜索的所有结果中,被其他网页用超链指向得越多的网页,其价值就越高,就越应该在结果排序中排到前面。超链分析的结果可以反映网页的重要程度,利用某网页被其他网页的引用情况推导出该网页的权威性。这类类似于文献计量学中的引文分析,从而给用户提供更重要、更有价值的搜索结果,保证了用户在使用搜索引擎搜索时,越符合用户检索要求的内容排名越靠前。

超链分析是一种引用投票机制,对于静态网页或者网站主页,它具有一定的合理性,因为这样的网页容易根据其在因特网上受到的评价产生不同的超链指向量,超链分析的结果可以反映网页的重要程度,从而给用户提供更重要、更有价值的搜索结果。

但是,超链分析本质上是针对一种公开的、通行的价值评估体系的。当用户搜索的目的是寻找关于某些关键字的站点资源或网站入口时,它是有效的;但当用户搜索的目的是寻找关于某些内容的有效信息本身时,超链分析的结果不仅没有参考价值,而且会破坏用户搜索结果的精确度。

用户搜索关于某些内容的有效信息时,最大的特点是各异性,就是说,没有绝对意义上的“好”网页或“坏”网页,只有“有用的”网页和“无用的”网页。有用的网页是包含了满足用户的搜索目的,能够提供给用户足够信息的网页,而无用的网页是与用户搜索目的不相关或不能够直接提供用户所需信息的网页。从普遍意义、通用意义上的价值来评估的搜索结果,对用户查找具体的资料和信息没有什么用处。某个被普遍引用的网页从绝对意义上来说,可能是更有价值的,但对用户来说却是无用的(例如各种门户和入口网页);而某个很少被引用的关于某个具体问题的文章的网页,对于某个用户的搜索目的来说,可能是最佳的结果。

超链分析技术的应用导致了用户搜索到的不是更符合自己需要的网页,而是找到那些最热门的网页。用户通过搜索寻找自己想要信息的主动过程,变成了接受一种根据某种标准排名次的网页的被动过程。在超链分析的影响下,搜索引擎的发展从追求对用户寻找到最有用信息的技术研究,演变成了各个网站想尽办法追求网页排名的商业活动。从这个意义上来说,超链分析也许从商业上来说是有价值的,但从搜索引

擎的基本用途来看,已经有些走入歧途。

4 百度与 Google 功能比较

登录百度(<http://www.baidu.com>)与 Google(<http://www.google.com>)的首页,两者的界面风格、内容非常接近。百度可进行新闻、网页、贴吧、MP3、图片、网站及更多搜索;Google 可进网页、图片、新闻、论坛、网页目录和更多搜索,Google 有 3 个选择:搜索所有网页、搜索所有中文网页、搜索简体中文网页。

百度与 Google 运用的都是关键字搜索,一次输入若干个关键词,中间用空格隔开,系统会自动进行逻辑与运算。

发展到今天的搜索引擎已不仅仅是网页搜索工具,百度与 Google 都开发了许多新功能,如拼音—汉字转换、自动纠错、计算器、度量衡转换、股票查询、英汉互译、专业文件搜索。

此外,百度还可进行列车时刻表、飞机航班的查询,Google 具有汉字简繁自动转换以及天气、邮编、区号、手机号归属地查询等功能,还可以查询词或词组的定义,只需键入“define”或“定义”,接着键入一个空格,然后键入要查询的词,如输入 define HTML 可以查询 HTML 的定义。

百度与 Google 都具有网页快照的功能,解决了用户上网访问经常遇到的死链接的问题。搜索引擎已预先浏览各网站,拍下网页快照,这就贮存了大量应急网页,以备在用户找不到原来的网页时使用。

百度与 Google 都支持元词检索(meta words),即通过把元词放在关键词的前面,告诉搜索引擎你想要检索的内容具有哪些明确的特征。例如,在搜索引擎中输入“title:清华大学”,就可以查到网页标题中带有清华大学的网页,在键入的关键词后加上“domain:org”,就可以查到所有以 org 为后缀的网站,其他元词还包括:image:用于检索图片;link:用于检索链接到某个选定网站的页面:url:用于检索地址中带有某个关键词的网页。

还有一点值得注意,Google 搜索不区分英文字母大小写,所有的字母均当作小写处理。例如:搜索“google”“GOOGLE”或“GoOgLe”,得到的结果都一样,而百度则区分英文大小写字母。百度的精确匹配检索可以处理双引号和书名号,Google 只能处理双引号。

同样,访问百度与 Google 的高级检索界面,两者也是几近雷同。

据“中国互联网络信息中心”(CNNIC)的统计,搜索引擎已成为与电子邮件并重的最重要的网络应用,“给我一个关键词,我可以查询整个世界”这句话一点也不过分,运用搜索引擎,探讨搜索技巧,信息检索会变得易如反掌。

参考文献

- [1] 向桂林.复合型 Web 信息检索系统[J].情报学报,2003(5):545-549.
- [2] 吴江.WWW 超链分析技术及其应用 [J]. 中国信息导报,2004(3):58-60.

(责任编辑:薛培荣)

Comparative Analysis on the Search Engines of Baidu and Google

ZHANG Lan

ABSTRACT: This paper introduces the present situation of the search engines of Baidu and Google, analyzes on the working principle of the search engines and their core technique—the hyperlink analysis, and compares the differences and similarities between the two search engines of Baidu and Google.

KEY WORDS: search engines; Baidu; Google; hyperlink analysis

第一作者简介:张 岚,女,1973 年 12 月生,1996 年毕业于北京师范大学信息与管理学系,馆员,沈阳工程学院图书馆,辽宁省沈阳市道义经济开发区正义三路十八号,110136.