

关于搜索引擎的研究综述

沈贺丹¹, 潘亚楠², 邵良杉¹

(1. 辽宁工程技术大学 系统工程研究所, 辽宁 阜新 123000;

2. 辽宁石油化工大学, 辽宁 抚顺 113001)

摘要:由于目前搜索服务被越来越多的用户所认识和青睐, 各样的搜索引擎也应运而生。文中阐述了搜索引擎的工作原理, 并对搜索引擎按照不同的依据对其进行分类。介绍并比较目前较为有名气同时其发展历史也推进了搜索引擎发展的几个搜索引擎, 最后提出目前搜索引擎所存在的问题。

关键词: Spider; 超链分析; 元搜索引擎

中图分类号: TP393.092

文献标识码: A

文章编号: 1005-3751(2006)04-0147-03

A Study for Search Engine

SHEN He-dan¹, PAN Ya-nan², SHAO Liang-shan¹

(1. System Engineering Research Institute, Liaoning Technical University, Fuxin 123000, China;

2. Liaoning University of Petroleum and Chemical Technology, Fushun 113001, China)

Abstract: Owing to search serve has been known and favoured by more and more Internet users, many kinds of search engines emerge as the times require. This paper sets forth the work principle of search engine, and sorts it on different basis. Then recommend some search engine companies which with great fame and have push the development history of search engine. In the end, bring forward some problems which exist in search engine at present.

Key words: Spider; hyperlink analysis; a meta search engine roundup

0 引言

目前,关于具体的搜索引擎的介绍及比较的综述比较多,但是笔者认为从理论层面对搜索引擎进行探讨更有意义,因为理论的意义在于它的前瞻性,特别是对搜索引擎这种实践性较强的学科,实时的理论关注会使搜索引擎的应用发展更理性、更科学。

1 搜索引擎的工作原理

搜索引擎并不真正搜索互联网,它搜索的实际上是预先整理好的网页索引数据库。搜索引擎也不能真正理解网页上的内容,它只能机械地匹配网页上的文字。

真正意义上的搜索引擎,通常指的是收集了互联网上几千万到几十亿个网页并对网页中的每一个文字(即关键词)进行索引,建立索引数据库的全文搜索引擎。当用户查找某个关键词的时候,所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。在经过复杂的算法进行排序后,这些结果将按照与搜索关键词的相关度高

低依次排列。

现在的搜索引擎已普遍使用超链接分析技术,除了分析索引网页本身的文字,还分析索引所有指向该网页的链接的 URL, AnchorText, 甚至链接周围的文字。所以,有时候,即使某个网页 A 中并没有某个词比如“考博”,如果有别的网页 B 用链接“考博”指向这个网页 A,那么用户搜索“考博”时也能找到网页 A。而且,如果有越多网页(C, D, E, F...)用名为“考博”的链接指向这个网页 A,或者给出这个链接的源网页(B, C, D, E, F...)越优秀,那么网页 A 在用户搜索“考博”时也会被认为更相关,排序也会越靠前。

搜索引擎的原理,可以看做三步: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排序。

(1) 从互联网上抓取网页:利用能够从互联网上自动收集网页的 Spider 系统程序,自动访问互联网,并沿着任何网页中的所有 URL 爬到其它网页,重复这过程,并把爬过的所有网页收集回来。

(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、与其它网页的链接关系等),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

收稿日期:2005-07-03

基金项目:辽宁省教育厅自然科学基金资助项目(20022154)

作者简介:沈贺丹(1980-),女,辽宁抚顺人,硕士研究生,研究方向为 Web 内容挖掘;邵良杉,教授,博士生导师,辽宁工程技术大学副校长,研究方向为计算机应用技术和科学管理。

后用这些相关信息建立网页索引数据库。

(3) 在索引数据库中搜索排序: 当用户输入关键词搜索后, 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的链接地址和页面内容摘要等内容组织起来返回给用户。

2 搜索引擎分类

搜索引擎按照不同的分类方式可分为以下几种类型:

(1) 按照检索方式分为独立型搜索引擎和元搜索引擎。独立型搜索引擎: 拥有自己的索引数据库, 检索在自身数据库进行, 并根据数据库的内容提供有关信息或连接站点; 元搜索引擎 (A Meta Search Engine Roundup): 用户只需提交一次搜索请求, 由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎, 并将各独立搜索引擎返回的所有查询结果, 集中起来处理后再返回给用户 (注: 元搜索引擎概念上好听, 但搜索效果始终不理想, 所以没有哪个元搜索引擎有过强势地位)。

(2) 根据搜索引擎的不同时期的研究重点和服务性能, 可以将搜索引擎分为三代。第一代搜索引擎出现于 1994 年。这类搜索引擎一般都索引少于 100 万个网页, 极少重新搜集网页并去刷新索引。而且其检索速度非常慢, 一般都要等待 10s 甚至更长的时间。在实现技术上也基本沿用较为成熟的 IR (Information Retrieval)、网络、数据库等技术, 相当于利用一些已有技术实现的一个 WWW 上的应用。大约在 1996 年出现的第二代搜索引擎系统大多采用分布式方案 (多个微型计算机协同工作) 来提高数据规模、响应速度和用户数量, 它们一般都保持一个大约 5000 万网页的索引数据库, 每天能够响应 1000 万次用户检索请求。自 1998 年到现在, 出现了一个搜索引擎空前繁荣的时期, 一般称这一时期的搜索引擎为第三代搜索引擎。第三代搜索引擎的发展有如下几个特点:

- * 索引数据库的规模继续增大, 一般的商业搜索引擎都保持在几千万甚至上亿个网页。

- * 除了一般意义上的搜索以外, 开始出现主题搜索和地域搜索。很多小型的垂直门户网站开始使用该技术。

- * 由于搜索返回数据量过大, 检索结果相关度评价成为研究的焦点。相关的研究又可以分为两类: 一类是对超文本链的分析, 在这方面 Stanford 大学的 Google 系统和 IBM 的 Clever 系统作出了很大的贡献; 另一类是用户信息的反馈, DirectHit 系统采用的就是这种方法。

- * 开始使用自动分类技术。NorthernLight 和 Inktomi 的 DirectoryEngine 都在一定程度上使用了该技术。这一阶段的发展为搜索引擎拓展了生存空间, 同时提高了搜索的质量和效率, 为以后的发展奠定了坚实的基础。

(3) 按照索引方式的不同可以分为目录系统和搜索引

擎系统。第一类是目录系统, 它通过有专业知识的网页编辑人员对网上的网页进行精选, 建立一个索引目录, 来为用户提供服务。这类系统的优点是提供的网页准确率高, 但覆盖的范围小, 其典型代表是 Yahoo。第二类是搜索引擎系统, 它通过程序自动地从网上搜集和分析网页, 建立索引, 为用户服务, 其典型代表是 AltaVista。这类系统的优点是涵盖的网页数量巨大, 但搜索的准确率相对较低。

3 著名的搜索引擎

搜索引擎自 1993 年出现发展至今, 已取得了长足的进步, 信息检索工具搜索引擎也是层出不穷, 以下是与搜索引擎发展历史息息相关的几个搜索引擎:

(1) Fast (All the web) 公司创立于 1997 年, 是挪威科技大学 (NTNU) 学术研究的副产品。1999 年 5 月, 它发布了自己的搜索引擎 All The Web。Fast 创立的目标是做世界上最大和最快的搜索引擎, 几年来庶几近之。Fast (All the web) 的网页搜索可利用 ODP 自动分类, 支持 Flash 和 Pdf 搜索, 支持多语言搜索, 还提供新闻搜索、图像搜索、视频、MP3 和 FTP 搜索, 拥有极其强大的高级搜索功能。

(2) Teoma 起源于 1998 年 Rutgers 大学的一个项目。Apostolos Gerasoulis 教授带领华裔 Tao Yang 教授等人创立 Teoma 于新泽西 Piscataway, 2001 年春初次登场, 2001 年 9 月被提问式搜索引擎 Ask Jeeves 收购, 2002 年 4 月再次发布。Teoma 的数据库目前仍偏小, 但有两个出彩的功能: 支持类似自动分类的 Refine; 同时提供专业链接目录的 Resources。

(3) Wisenut 由韩裔 Yeogirl Yun 创立。2001 年春季发布 Beta 版, 2001 年 9 月 5 日发布正式版, 2002 年 4 月被分类目录提供商 looksmart 收购。Wisenut 也有两个出彩的功能: 包含类似自动分类和相关检索词的 WiseGuide; 预览搜索结果的 Sneak-a-Peek。

(4) Gigablast 由前 Infoseek 工程师 Matt Wells 创立, 2002 年 3 月展示 pre-beta 版, 2002 年 7 月 21 日发布 Beta 版。Gigablast 的数据库目前仍偏小, 但也提供网页快照, 一个特色功能是即时索引网页, 你的网页刚提交它就能搜索 (注: 这个 spammers 的肉包子功能暂已关闭)。

(5) Openfind 创立于 1998 年 1 月, 其技术源自台湾中正大学吴升教授所领导的 GAIS 实验室。Openfind 起先只做中文搜索引擎, 曾经是最好的中文搜索引擎, 鼎盛时期同时为三大著名门户新浪、奇摩、雅虎提供中文搜索引擎, 但 2000 年后市场逐渐被 Baidu 和 Google 瓜分。2002 年 6 月, Openfind 重新发布基于 GAIS30 Project 的 Openfind 搜索引擎 Beta 版, 推出多元排序 (PolyRankTM), 宣布累计抓取网页 35 亿, 开始进入英文搜索领域, 此后技术升级明显加快。

(6) 北大天网 是国家“九五”重点科技攻关项目“中文

编码和分布式中英文信息发现”的研究成果,由北大计算机系网络与分布式系统研究室开发,于1997年10月29日正式在CERNET上提供服务。2000年初成立天网搜索引擎新课题组,由国家973重点基础研究发展规划项目基金资助开发,收录网页约6000万,利用教育网优势,有强大的FTP搜索功能。

(7)2000年1月,两位北大校友、超链分析专利发明人、前Infoseek资深工程师李彦宏与好友徐勇(加州伯克利分校博士)在北京中关村创立了百度(Baidu)公司。2001年8月发布Baidu.com搜索引擎Beta版(此前Baidu只为其它门户网站搜狐新浪Tom等提供搜索引擎),2001年10月22日正式发布Baidu搜索引擎,专注于中文搜索。Baidu搜索引擎的其它特色包括:网页快照、网页预览/预览全部网页、相关搜索词、错别字纠正提示、新闻搜索、Flash搜索、信息快递搜索。2002年3月闪电计划(Blitzen Project)开始后,技术升级明显加快。

(8)Northernlight公司于1995年9月成立于马萨诸塞州剑桥,1997年8月Northernlight搜索引擎正式现身。它曾是拥有最大数据库的搜索引擎之一,它没有Stop Words,它有出色的Current News,7100多出版物组成的Special Collection、良好的高级搜索语法,第一个支持对搜索结果进行简单的自动分类(注:2002年1月16日,Northernlight公共搜索引擎关闭,随后被divine收购,但在Nlresearch,选中“World Wide Web only”,仍可使用Northernlight搜索引擎)。

(9)1995年9月26日,加州伯克利分校CS助教Eric Brewer、博士生Paul Gauthier创立了Inktomi(UC Berkeley Announces Inktomi),1996年5月20日,Inktomi公司成立,强大的HotBot出现在世人面前。声称每天能抓取索引1千万页以上,所以有远超过其它搜索引擎的新内容。HotBot也大量运用cookie储存用户的个人搜索喜好设置(注:Hotbot曾是随后几年最受欢迎的搜索引擎之一,后被Lycos收购)。

(10)Google在1998年10月之前,只是Stanford大学的一个小项目BackRub。1995年博士生Larry Page开始学习搜索引擎设计,于1997年9月15日注册了google.com的域名,1997年底,在Sergey Brin和Scott Hassan,Alan Sterenberg的共同参与下,BackRub开始提供Demo。1999年2月,Google完成了从Alpha版到Beta版的蜕变。Google公司则把1998年9月27日认作自己的生日。

Google在Pagerank、动态摘要、网页快照、DailyRefresh、多文档格式支持、地图股票词典寻人等集成搜索、多语言支持、用户界面等功能上的革新,象Altavista一样,再一次永远改变了搜索引擎的定义。

在2000年中以前,Google虽然以搜索准确性备受赞誉,但因为数据库不如其它搜索引擎大,缺乏高级搜索语法,所以推广并不快。直到2000年中数据库升级后,又借被Yahoo选作搜索引擎的东风,才一飞冲天。

4 搜索引擎存在的问题

搜索引擎在飞速发展的同时也存在着很多缺陷,需要进一步改进和完善,笔者对这些问题进行了归纳,如下:

(1)网络信息质量控制欠缺。任何人只要具备相应的条件就可以把任何信息送到网上,800C这些信息不经任何质量控制就被搜索引擎标引,未经质量控制的信息必然会影响搜索结果的质量。

(2)大量占用昂贵的网络带宽和CPU资源。由于搜索引擎必须将大量资源站点的内容传送到搜索站点本地,然后进行分析索引,这样大规模的资源文件的传送和出路无疑会增加网络传输的负担,使网络变得更加拥塞,此外也大量占用了被搜索站点和搜索站点本身的CPU资源,致使用户的访问不能得到系统及时的响应。

(3)覆盖面有限。《科学》杂志最近一份研究报告表明,即使功能最完善的搜索引擎,也只能找到Web上大约三分之一的网页。

(4)索引数据库更新困难,提供的信息滞后。搜索引擎一般都有庞大的索引数据库,其更新速度总是落后于时刻在更新的因特网信息的更新速度。并且索引库越大,其更新周期越长,索引失效问题越突出。许多搜索引擎必须通过人工方式对信息进行二次处理,这也是造成信息滞后的一个重要原因。

(5)搜索引擎之间各行其事,缺乏合作。目前很多搜索引擎都出现对同一个资源站点进行分析、索引的情况。这种重复造成很大的资源浪费。

(6)搜索速度不理想。为了提高效率,人们开始倾向于开发较小的专用搜索引擎,通过集中地执行特定任务,专用的搜索引擎在其运行领域中会表现出更大的灵活性。

(7)误检率低,漏检率高。原因有很多:a.虽然搜索引擎能检索到大量信息,但是与全部因特网信息相比,仅是沧海之一粟;b.现在搜索引擎主要是通过Robot等软件将网页全部或部分内容下载到自建索引库中,下载的页面许多是无用或暂时信息;c.搜索引擎一般不会遗漏较重要的网站,但由于对网站的描述较为简单,不能深入网站的内部标引。要解决误查和漏检问题,最根本的途径是搜索引擎具有认知能力和推理能力。目前人工智能搜索引擎还处于研究开发阶段;d.用户检索机制不完善;e.信息分类不规范。

(8)搜索引擎的功能尚待完善。a.搜索引擎的发展程度参差不齐;b.目前还没有任何一个网络检索工具可在检索功能上与传统的计算机化检索工具相媲美,其功能还有很大的发展余地。

(9)检索结果重现性差。现行Web搜索引擎由于其检索技术存在的问题和不足,使得同一检索策略试用不同搜索引擎的检索结果各不相同;甚至同一搜索引擎在不同时间检索时所得检索结果也完全不相同。需要同时试用多个搜索引擎才能得到相对全面的检索结果。

(下转第152页)

//其他的方法实现

●跟踪日志 Before 通知方面 LogAdvice 代码描述:

```
public class LogAdvice implements MethodBeforeAdvice{
    public void before(Method m, Object[] args, Object target)
    throws Throwable
    {
        Logger auditor = Logger.getLogger(target.getClass()); //
        生成 Log4j 实例
        auditor.debug("将要执行" + method.getName()); //跟踪
        记录业务 Bean 的动作
    }
}
```

在 springconfig.xml 文件中,定义切入点,通过声明方式,把跟踪日志和所要应用的业务 Bean 联系起来,具体描述省略。

当 Action Bean 或应用程序执行 BookManager,日志关注点和业务 Bean BookManager 被编织起来,当它的公共函数 saveBook 被执行的时候,它们被跟踪并记下日志。

业务 bean 执行序列图如图 2 所示。

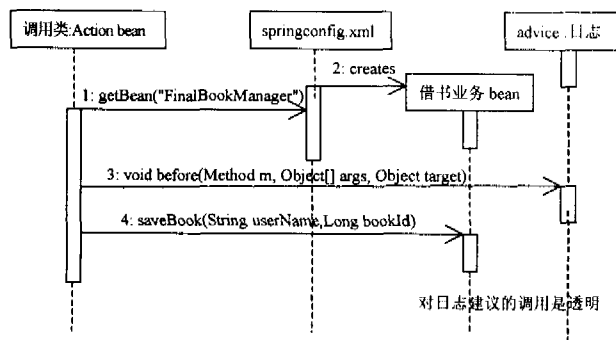


图 2 应用日志记录通知的借书业务 bean 的序列图

日志记录通知 LogAdvice 作为 MethodBeforeAdvice 子类,它将拦截访问借书业务 bean(目标对象)的方法。在借

书业务 bean 执行之前,日志记录通知 LogAdvice 在 before 方法内记录借书业务 bean 所作的动作,然后借书 Bean 再完成所要做的动作。

如果 OOP 实现借书业务 bean,业务 bean 要实现借书业务功能,还要实现跟踪日志的功能。而采用 AOP 来实现,则借书业务 bean 只实现借书的本职功能,跟踪日志的功能交由专门处理日志的方面来完成,这样就使得业务 Bean 和跟踪日志功能解耦。

3 结束语

AOP 技术简化了 J2EE 应用系统的开发,减少了实现横切关注点的重复代码,节约了时间,增加了开发效率,应用系统变得可测试、容易维护。设计师再也不必陷入设计不足或者过度设计的两难境地。Spring 框架提供的 AOP 实现推动了 AOP 技术在 J2EE 应用系统中的使用。

参考文献:

- [1] Sharwood S. A new aspect to programming[EB/OL]. http://www.builderau.com.au/architect/0_39024564_39183763.00.htm, 2005-04-08.
- [2] Johnson R, Hoeller J. Expert One-on-One J2EE Development without EJB[M]. Indianapolis, Indiana: Wiley Publishing, Inc, 2004.
- [3] O'Regan G. Introduction to Aspect-Oriented Programming[EB/OL]. <http://www.onjava.com/pub/a/onjava/2004/01/14/aop.html>, 2004-01-14.
- [4] Walls C, Breidenbach R. Spring in Action[M]. Greenwich, CT: Manning Publications Co, 2005.
- [5] Miles R. An Introduction to Aspect-Oriented Programming with the Spring Framework, Part 1-2[EB/OL]. <http://www.onjava.com/pub/a/onjava/2004/07/14/springaop.html>, 2004-07-14.

(上接第 149 页)

(10)缺乏检索专业信息的能力。通常用的搜索引擎,一是不以专业划分检索范围,二是特定专业的检索工具应该试用与之相应的标引和检索语言,而这是国际互联网检索工具难以做到的。因而利用网罗检索工具检索专业的网络信息效果不可能太理想。

(11)检索过程的重复性。现有的网络信息检索需要用户自行组织检索过程,单个用户的结果不能被其他相同需要的用户共享,这也是一大缺陷。

(12)搜索引擎的知识产权问题。信息社会中,产权问题无时不在,无处不有。搜索引擎涉及的知识产权问题也受到了学界的关注。

5 结束语

随着 WWW 上数据量的不断扩大,信息内容的不断丰富,人们对搜索引擎的要求也不断提高,这也促进了搜

索引擎的进一步发展。虽然搜索引擎已经有二十几年的历史了,但不能否认其仍然处于研究开发阶段。因为有很多问题还需要解决。同时也说明搜索引擎也是一个非常具有挖掘潜力的技术。

参考文献:

- [1] 雷鸣,王建勇,赵江华,等. 第三代搜索引擎与天网二期[J]. 北京大学学报(自然科学版), 2001, 37(9): 735-740.
- [2] 胡冉. 关于搜索引擎的几个理论问题的综述[J]. 晋图学刊, 2003, 74(2): 74-77.
- [3] 许晋军, 苏新宁. 信息搜索引擎综述[J]. 计算机系统应用, 1999(4): 22-24.
- [4] 北京大学天网搜索引擎[EB/OL]. <http://e.pku.edu.cn>, 2005-05-06.
- [5] 朱俊卿. 搜索引擎 Google 研究[J]. 广州大学学报(综合版), 2001, 15(11): 7-10.