Web信息检索技术最新进展

陈定权

（中国科学院文献情报中心 北京 100080）

【摘要】 Web信息检索是一个集人工智能、数据挖掘、自然语言处理、数据库技术、计算机网络等于一体的综合技术。尽管搜索引擎的出现在一定程度上缓解了用户对信息需求难以满足的矛盾，但是它还没有达到令人满意的程度。本文对Web信息检索技术作了一个比较全面的论述，尤其从超链分析的角度，对网络链接的重要性做了深入探讨并指出了它的发展方向，对这方面的理论研究和实际开发具有一定的参考价值。

【关键词】 搜索引擎 链接分析 Web挖掘 万维网 【分类号】 G354.4

An Review of Web Information Retrieval

Chen Dingquan

(The Documentation and Information Center of CAS, Beijing 100080,China)

【Abstract】 Web information retrieval is integrated with artificial intelligence, data mining, natural language processing, computer network, database technology, etc. search engine resolves partly users' information needs, but it can't satisfy users' needs. This paper offers a review of the current search engine design. Especially from the view of Web link analysis, how to apply link structure into search engine is popular research topic.

【Keywords】 Search engine Link analysis Web mining World Wide Web

1 引 言

随着互联网飞速发展，互联网上的信息呈爆炸式增长。如何在互联网上查找所需信息一直是热门的研究课题，搜索引擎也因此应运而生。从最早的目录主题型的搜索引擎发展到检索型的搜索引擎、元搜索引擎、分布式搜索引擎，现在发展到应用数据挖掘、人工智能等技术的智能搜索引擎，从而使得用户可以更好地利用网络信息资源。

目前许多搜索引擎一般都使用传统信息检索算法和技术。然而传统的信息检索算法主要是从相对少量和同构的文献集合（如新闻、书目等）发展过来的。然而，Web上的信息具有巨量的、异构的、非结构或半结构的、动态的、分布的等特点，对传统的信息检索技术提出了挑战。本文就网络信息检索的相关技术作一个综述和对最新技术作一个展望。

2 搜索引擎

在具体讨论搜索引擎的各项技术之前，让我们先回顾一下搜索引擎是如何工作的。典型的搜索引擎由以下几个部分组成[1,2]：爬行器（或称为机器人、蜘蛛等）、索引生成器、查询检索器等三大模块。其中爬行器主要完成信息获取工作，以期为将来的服务提供数据，索引生成器是通过分析获取的网页，排除HTML等语言的标志符号，将出现的字或者词（排除停用词）抽取出来，并记录每个字词的出现网址及相应位置，最后将结果存入索引数据库。检索模块首先分析用户检索时给出的提问式，再访问搜索引擎已经建立的索引，并通过一定的匹配算法，获得相应的检索结果。一般要对结果按照相关程度将结果有序地返回给用户（有时还可以返回它认为与用户检索词相关的特征词），用户可以将认为是相关的文档（和相关特征词）反馈给服务器，服务器根据用户的反馈重新构造检索式、修改检索词权值重新检索[3]，重新将结果返回给用户，这个过程可以重复直到用户满意或放弃为止。下面几部分分别介绍一下信息采集、网页存储系统、索引、结果排序、以及Web link的各项应用。

收稿日期：2001－12－08

3 信息采集

在爬行器部分，主要考虑以下几个问题：

· 下载什么网页[1]？在大多数情况下，爬行器不可能下载所有网页，只能下载其中一部分的。这样，如何下载比较“重要”的网页就是一个很现实的问题。判断一个网页是否重要的依据是什么？目前有以下3种：①兴趣驱动；②流行性驱动；③位置驱动。在具体爬行过程中，有两种模式，其一是当下载了K页后自动停止，另一种是下载了K个“重要”页面后停止。爬行器要负责等待访问的URL堆栈，如何对其中的URL按照某种原则决定其优先级也影响爬行器的效率，使得爬行器爬行最少的页面获得最多的“重要”页面。

· 怎样更新网页？当网页被下载后，怎样去探测网页已经被更新以及刷新文档库。它以一定的频率对全部网页进行刷新或对网页的重新访问频率与该网页自身更新的频率相适应，但不是成正比。例如某网站一天更新5次，那么爬行器是否也更新5次呢？这就涉及到最优更新频率问题。

· 怎样减少爬行器对Web服务器的负担？当很多爬行器在工作时，对Web服务器而言是一个不小的负担，消耗了服务器资源。如何较少服务器负担，避免阻塞也是设计爬行器应该考率的问题。例如，有的搜索引擎与网站达成协议，只有在网站服务器端放置特殊标记文件，爬行器才采集；有的网站服务器按照爬行器的要求建立索引文件，爬行器只采集这个索引文件即可。

· 爬行器怎样并行来工作？由于网页数量的庞大，许多爬行器在多台机器上工作，并行下载网页，从而使得在最短的时间内下载更多的网页。很显然，这些并行工作的爬行器必须协同工作，以便使得不同的爬行器不会重复访问。并行工作处理的好坏，直接影响爬行器的效率。

4 Web网页存储

网页数据库用来管理所有采集的网页，它应该是一个扩展性能好的存储系统，有两个接口：面向爬行器的和面向索引器的。一个优秀的网页存储体系具备以下几点[1]：

· 良好的扩展性：为了适应网页爆炸式增长，必须能够无缝链接分布的网页数据库。

· 双重访问方式：应该能够提供随机访问和流方式访问。随机方式可以快速某一指定的网页，用于响应用户的检索查询，而流方式可以访问整个网页数据库或其中的一部分，用于索引和分析模块之中。

· 大量数据的快速更新：Web变化相当快，网页数据库需要处理快速变化的网页，避免更新过程与检索网页相互冲突。

· 管理过时的网页：很多网页可能从它的站点中删掉，这样必须有一种机制能够自动监测和管理过时的网页，或者删掉或者保留。如Google就将过时网页当成"网页快照"保留起来[1]。

在技术上要考虑以下问题[1]：

· 网页的存储分配：是采用哈希(Hash)方式还是统一分配(Uniform distribution)· 物理网页的组织方法：对于单个网页，可能存在三种可能的操作：网页添加/插入、高速流方式访问、随机页面访问。组织方式的不同很大程度上决定了这些操作的性能好坏。目前有Hash－based、Log－structured、Hashed－log等组织方式。

· 更新策略：更新一般由爬行器来完成。爬行器是定期采集还是不停地采集？定期采集是部分爬行还是全部？当是部分爬行时，只是爬行部分网页用来更新，全部爬行时，是将采集结果完全替换已经存在的网页。更新也有两种方式：1)将采集的网页直接存入数据库；2)先存入到另外地方，用另外的程序来单独更新。

5 索 引

在搜索引擎中一般有以下几类索引：内容索引(content index)和结构索引(structure index 或链索引(link index)[1]。这些索引在建立时候涉及到索引的结构、索引的可扩展性和分布特点、索引生成的并行化等技术问题。

· Link索引 为了对超链创建索引，被抓取的网页看成是有节点和边的有向图。其中节点为页面，边为超链。超链结构索引必须是可扩展和高效的。最通用的结构信息是相邻信息。例如网页p，获取p所指向的网页以及指向p的网页。如何处理这些大规模的有向图是一个非常棘手的问题。

· 文本索引 即使超链索引所带来的功能多么有效，但文本索引始终是索搜引擎在判断一篇文档是否与查询相关的主要方法。英文主要是基于词的索引，汉语有基于词和基于字的两种情况。为了处理汉字与英文的混合查询，如"甲A"等关键词，也有学者将中文英文统一处理[3]。这方面有比较成熟的算法，在此不作详细讨论。

· 其它索引 有些搜索引擎允许对某一特定领域或特定站点进行检索，此时需要建立站点索引(site index)。该站点索引将某一域名映射为属于那个域名的网页。还可以建立其它类似的索引(如地域、语种等)用来改善或增加搜索引擎的功能。

索引文件通常占有相当大的空间。如何存储大容量索引并且要求存取快捷方便，也是一个热点问题。有许多研究者提出了基于压缩的二或三级索引的方案[6,7]。在保证响应速度的前提下，该方案可以大大减少索引文件的空间。

6 结果排序与超链分析

查询检索器根据用户提交的查询式，按照某种策略，将它认为是与用户查询式相关的文档根据相关度从大到小返回给用户。经典的布尔模型、概率模型、向量空间模型是基于特征词来进行相关度判断。许多网页因为含有检索词，但是离用户真正的相关度相差太远，有的不含有检索词，但也是与用户相关的。再者，用于表达文档的那些特征词也不能完全代表文档。后来许多学者发现，在网页之间的相互链接中包含了丰富的信息[8]。根据这个思想，人们试图从那些链接信息来挖掘一些可以改善搜索引擎的技术。如Google[4]、百度[9]等搜索引擎已经部分采用了这项技术，尽管不是很成熟，但已经显示出Link对检索性能具有很大的影响。

· PageRank[10] PageRank的基本思想是：一个页面被多次引用，即很多页面有指向它的链接，则这个页面很重要；一个页面尽管没有被多次引用，但被一个重要页面引用，则这个页面也可能很重要；一个页面的重要性被均匀的分布并传递到它所引用的页面。

B(i)代表指向页面i的页面集合。N(i)表示页面i中指向其它页面的超链数目。R(i)表示为页面i的相关度。在实际的Google中采用如下公式：

$$R(i)=(1-c)+c\cdot\sum_{j\in B(i)}\frac{R(j)}{N(j)}$$ (Page&Brin认为C的最佳值为0.85)。

Page&Brin就根据这个原理，与关键词检索以及其它基于文本的技术一起来提高查询质量。例如链接的标记文字(anchor text)可以认为是对链宿页面的概括。因为网页数量之巨大，不可能对全部的页面进行PageRank分析，所以实际的工作过程如下：先用基于关键词的搜索得到一个集合，取前面N个。然后对这N个页面应用PageRank算法，得到最终的排序结果。这个算法只对In－degree trees、Out－degree trees、Complete bipartite Graph有效，而对Bipartite、General Graphs无效。具体可以参考[11]。所以PageRank还需要进一步优化结合其它的技术一起才能很好工作。

· HITS[12~14] HITS(Hypertext Induced Topic Search)最早是在1999年由Kleinberg提出。与PageRank不同的是它依赖于查询式。HITS认为页面的重要性依赖于正在查询的查询式；每页有两个级别

(ranking):权威级别(依赖于指向它的页面)、中心级别(依赖于它指向别人的页面)。工作过程如下:1)用基于文本的搜索引擎得到某一查询的结果结合R(称为Root Set);2)将R所指向的页面集合以及其它指向R的页面集合包含进来形成集合S;3)将所有页面的权威级别、中心级别全部置初始值为1,再按照如下算法计算:

Repeat until convergence(收敛)

For all Page iin S, $a_i=\sum_{j\in B(i)} h_j$; $h_i=\sum_{j\in F(i)} a_j$;

Normalize: $\sum_i a_i^2=1$; $\sum_i h_i^2=1$

End

其中B(i)代表指向页面i的所有页面集合;F(i)代表页面i所指向的页面集合。最后根据权威级别的大小返回给用户。Meghabghab[11]在只考虑链接因素条件下,对In-degree trees、Out-degree trees、Complete bipartite Graph、Bipartite、General Graphs几种网络拓扑图的性能做了一个大概的评估。其它类似的算法还有SALSA、pSALSA、PHITS等,基本原理与HITS相同,具体可以参考文献[15]。

· 其它Link Analysis应用

⚓ Results Cluster[16]

目前搜索引擎的结果还不是令人满意。与基于词或短语的文本聚类算法不同,有学者使用超链分析来对结果进行聚类。它是基于Co-citation和Coupling分析来过滤无关文档,将质量高的文档进行聚类,提供给用户进行浏览和访问。例如用户检索"Jaguar",将结果聚类如下:Jaguar Car、Jaguar Club、Magazine on Jaguar car、Jaguar Game等等,从而方便用户浏览。

⚓ Query By Examples[8,17]

根据实例查找,也可以称之为找相关网页。根据用户需要查找的某一实例,例如一个网页,找出与之相关类似的网页。在Google和Netscape中支持这个服务。传统的信息检索技术是采用文本相似度,而在Web环境中,可以充分挖掘链接结构来实现。基本思想是:如果网页A指向网页B和C,则B和C可能相关。Kleingerg声称将HITS算法稍加修改也可以用来实现实例查找。Dean&Henzinger提出了两种算法:Companion算法、Co-citation算法。基于链接分析的算法总体上优于基于文本相似度算法。

⚓ Mirrored Hosts[8]

网页路径是URL地址的一部分。例如:URL:http://google.com/about.html。www.google.com就是主机地址,/about.html是路径。两台主机H2和H1是镜像网站,当且仅当H2中的文档,在H1中具有相同路径的相似文档,反之亦然。镜像网站具有相似的超链结构。通过超链分析可以检测出近似的镜像网站,从而可以节省索引空间和存储空间。

⚓ Geographic Scope[8]

给定的Web网页是局部地区的还是全国的抑或是全世界范围内的人对其感兴趣?通过对超链分析可以得出该网页的兴趣覆盖范围。这个信息可以帮助搜索引擎根据用户所在地区来裁减检索结果。

⚓ Identifying Communities[18]

在网络上有许多在线的由某些有共同兴趣的人们创建、维护、使用的网页,这些网页组成了一个虚拟的社会团体。例如数据库、Java兴趣小组等。根据中心网页和权威网页之间的相互关系可以找出这些虚拟的社会团体。

⚓ Categorization & Resource Compilation. 许多搜索引擎按照层次型的分类系统将网页分类。传统方法是用人工来进行分类、编辑。基于分类样本的文档自动分类问题目前在信息检索领域正在研究。Chakrabarti等人在传统的技术的基础上加入超链信息分析,通过试验验证加入超链分析可以提高分类的精度。通过该项技术可以在一定程度上实现某一特定分类体系对资源自动分类。

⚓ Web Impact Factor[20~22]

Web影响因子是从期刊影响因子发展过来。它的基本原理是:越多网页通过链接指向某一站点或区域,它就越有影响力。但是这些链接也需要进行分析:有些链接是导航;链源网页的重要性也影响链宿页面的重要性;可能一个链源指向一个站点的几个网页等。目前WEB影响因子能否应用到实际当中还有待进一步研究。

7 总 结

Web数据挖掘,可以分为三类[23~26]:内容挖掘、结构挖掘、用户访问模式挖掘。其中内容挖掘有基于文本和基于多媒体;结构挖掘主要是从Web组织结构和链接关系中推导信息、知识;用户访问模式挖掘主要是想从用户的访问日志中挖掘用户的访问模式。过去的搜索引擎主要是从网页中挖掘内容信息,现在正在转向充分利用结构信息来挖掘信息,利用用户访问模式来实现个性化服务,改善服务模式。如何有效利用结构信息,目前正处于一个初步试验发展阶段,如何抓住机会开发自己的智能型、知识型搜索引擎是我们应该共同面临的课题。

参考文献:

[1] Arvind Arasu, Junghoo Cho, et al. Searching the Web. ACM Transactions on Internet Technology, 2001,1(1):2-43

[2] Maristella Agosti and Massimo Melucci. Information Retrieval on the Web. ESSIR'2000, pp242-285. Spinger-Verlag Berlin Heidelberg 2000

[3] Zhixiang Chen, Xiannong Meng, Richard H. Fowler. FEATURES: Real-Time Adaptive Feature and Document Learning for Web Search. J. of The American Society for Information and Technology. 2001,52(8):655-665

[4] Google搜索引擎. http://www.google.com

[5] 陈华辉.一个中英文全文搜索引擎的设计与实现.计算机应用研究.2001,18(3):131-135

[6] 刘祖斌,王永成,刘椿年.中文全文检索系统中的压缩模型和模式匹配技术.中文信息学报,2000,14(4):42-46

[7] 余海燕,张仲义.基于单汉字索引的全文检索系统的优化研究,中文信息学报,2001,15(4):14-19

[8] Monika R. henzinger. Hyperlink Analysis for the WEB. IEEE Internet Co mputing, 2001,(1):45-50

[9] 百度全球中文搜索引擎. http://www.baidu.com.cn/

[10] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertexual Web Search Engine. Available from http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

[11] Georage Meghabghab. Google's Web Page Ranking Applied to Different Topological Web Graph Structures. Journal of the American Society for Information Science and Technology (JASIS), 2001,52 (9): 736-747

(下转第58页)

进而提示出其它相关的页面或链接。

④ 通知代理，通知代理会按照监督代理所给出消息（如："页面内容已更新"）通知界面代理或信息库更新信息内容，以便用户浏览最新的信息。

⑤ 监督代理。在智能代理已经搜索到一定的信息之后，监督代理会跟踪该信息所属网络页面的变化，只要该网页内容发生变化，监督代理即会通过通知代理发出消息。

⑥ 用户信息库和档案管理代理。前者存放用户自己的档案信息（如：专业领域，以前搜索过页面）和该用户目录下的信息内容，其实质是一个数据库。后者是对该数据进行管理的代理。

⑦ 网络探测代理。当在网上进行某项搜索应用时，网络探测代理（Webspider Agent）就会以该搜索应用为中心，向网络四周幅射出去，探测出与搜索应用相关的页面或衔接，经过处理再次生成搜索应用规则，传递给搜索代理进行二次搜索。

⑧ 描述模式代理。描述模式是由描述子构成的。描述模式代理根据外来请求或消息生成描述子，再由该代理生成描述语言（DL Description Languages）传递给规则应用库，由规则生成代理生成规则或应用，推动智能代理的进一步操作。

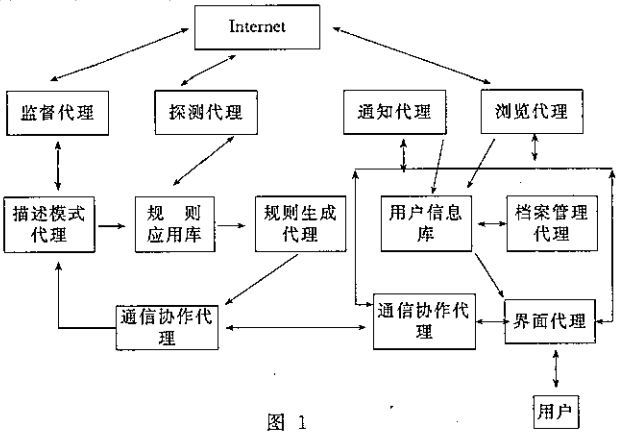

图 1

⑨ 规则应用库和规则生成代理，主要是根据描述语言生成规则和应用。（见图1）

（3）智能代理的运作过程

① 界面代理收到用户的信息需求之后，经过初步分析生成描述子，通过通信协作代理，传递给描述模式代理，生成描述语言，再由规则应用库和规则生成代理生成一定的搜索规则或应用（包括搜索范围限定，搜索主题词，搜索引擎的选定，以及信息过滤规则等），传递给搜索代理。搜索代理进行信息搜索，过滤后传递给用户信息库，由档案管理代理进行分类，查准率排序，相关率排序。最后将内容传递给界面代理，生成浏览页面，呈现在用户面前。

② 监督代理，网络探测代理及通知代理的运作过程信息用户提出了需求并浏览过信息之后，用户信息库将会存放一个由档案管理代理生成的标记（Dirty Flag），表示该信息内容已被浏览。该标记将传送给规则库（通过描述模式代理），作一个完全拷贝。监督代理根据规则生成代理实时地跟踪作过标记的信息所在Internet页面内容的变化，而网络探测代理则跟踪是否有相关内容的新网页（站）出现。如果这两个条件同时或满足其中一个的时候，描述模式会生成新的描述语言，（如：新的内容或页面出现），发往规则应用库，以规则生成代理生成新的规则或应用，发给通知代理。通知代理收到该消息之后通知用户信息库更新相关信息内容。上述的操作过程是在没有用任何干预之下自动完成。通知代理同样会将通知界面代理。页面该内容已改变。当用户浏览时会向用户提醒——有新的信息内容有待浏览，并适时调查显示内容和方式，给信息用户呈现最新的信息。

以上三种技术是现阶段实施网上主动性信息服务的主要技术，它们的成熟与完善还有待时日。随着信息技术的发展，必将有更多有关主动信息服务的技术出现。

参考文献：

[1] 盛小平. 试论虚拟图书馆的信息共享管理. 图书馆杂志，1999，(11)

[2] 蔡 巍. PUSH技术简介. 中国信息导报，1999，(3)

[3] http://www.puinfo.com/zhuantiwenzhang/ebusiness/business

（上接第41页）

[12] Jon M. Kleinberg. Hubs, Authorities, and Communities. ACM Computing Survey, 1999,31(4): 1—3

[13] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 1999,46(5), 604—632

[14] George Meghabghab. Discovering authorities and hubs in different topological Web graph structures. Information Processing and Management, 2002,38(1): 111—140

[15] Allan Borodin, Gareth O. Roberts, et al. Finding Authorities and Hubs From Link Structures on the World Wide Web. WWW10, May, 2001, Hong Kong. 415—429. Available from http://www.acm.org

[16] Yitong Wang and Masaru Kitsuregawa. Link Based Clustering of Web Search Results. Web Age Information and Management (WAIM'2001), 225—236. Springer—Verlag Berlin Heidelberg 2001

[17] Jeffrey Dean, Monika R. Henzinger. Finding Related pages in the World Wide Web. Computer Networks, 1999, 31: 1467—1479

[18] David Gibson, Jon Kleinberg, et al. Inferring Web Communities from Link Topology. Hypertext'98, Pittsburgh PA USA. 225—233. Available from http://www.acm.org

[19] Soumen Chakrabarti, Byron Dom, et al. Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks and ISDN system. 1998, 30: 65—74

[20] 黄 奇，李 伟. 基于链接的学术性WWW网络资源评价与分类方法. 情报学报，2001，20(2)：186—192

[21] Peter Ingwersen. The Calculation of Web Impact Factors. J. of Documentation, 1998,54(2):236—243

[22] Mike Thelwall. Web Impact Factors and Search Coverage. J. of Documentation, 2000,56(2):185—189

[23] 阳小华. Web站点的超链结构挖掘. 计算机工程与应用，2001，(8)：64—65

[24] 邓 英，李 明. Web数据挖掘技术及工具研究. 计算机工程与应用，2001，(20)：64—65

[25] 王继成，潘金贵，张福炎. Web文本挖掘技术研究. 计算机研究与发展. 2000,37(5):513—520

[26] 韩家伟，孟小峰，王 静，等. Web挖掘研究. 计算机研究与发展. 2001,38(4):405—414