基于网络信息检索的研究

马福晶[1],葛润霞[2]
(1.山东工商学院,山东 烟台 264005;
2.山东师范大学 信息科学与工程学院,山东 济南 250014)

摘 要:在信息时代,面对日益庞大的信息资源,需要通过一种科学手段来获取自己需要的准确及时的信息,这种科学的手段就是检索,信息检索。信息检索就是只从任何文献集合中查出所需信息的活动、过程和方法。通过论述信息检索的工作原理和其在网络环境下的作用,对比分析了基于网络的信息检索几种类型的特点,对高速而有效的信息检索系统的核心技术搜索引擎技术进行了分析,指出随之带来的亟待解决的快速有效获取信息的问题和搜索引擎技术符合时代要求的发展方向。

关键词:信息检索;联机检索;搜索引擎
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2008)08-0111-03

Research on Information Retrieval Based on Network

MA Fu-jing[1], GE Run-xia[2]
(1. Shandong Institute of Business and Technology, Yantai 264005, China;
2. School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: During inforamtion times, starve for a scientific method to get exact and timely information from progressively increasing resouses. This should be information retrieval, which look for activity, process and method of information from any documents . Via discussing the Conception, the principle and the significance of information retrieval, analyze and compare types of different network environment, and then analyze the core technology of information retrieval with high speeding and in effect, which is different types of search engine system, put forword the development trend of search engine system and the new problem that haven't solved.

Key words: information retrieval; on-line retrieval; search engine

1 信息检索原理和作用

信息检索包括信息的存储和检索两个方面[1]。其中存储是为了检索,而检索又必须先进行存储。信息的存储过程实际上是对信息进行整序的过程[2],信息的检索过程则是将信息特征标识与检索提问标识进行匹配的过程,也就是对大量的分散无序的信息依据一定的方法和规则,进行收集、加工、组织、存储,建成各种各样的检索系统,通过使用统一的检索语言和名称规范,将与用户所需的检索课题要求相匹配的内容从检索系统中检出。

信息检索的原理是"相符性比较"和"匹配运算"[3]。即首先必须对广泛、大量、分散、无序的信息进行搜集、记录、组织、存储,以建成各种检索系统(如手工检索工具、计算机检索的数据库与搜索引擎)。用户根据检索课题的需要,将需求转变为系统所能识别的检索式,再与检索系统中表征信息资源特征的标识进行逐一的相符性匹配与比较,两者完全一致或部分一致时即为命中信息。这就是信息检索的一般构成和原理,其中的统一检索语言和名称规范是存储和检索人员所必须共同遵守的。

信息检索是获取科学知识的最佳捷径。掌握了信息检索的方法和技能,就能够掌握获取文献的方法,提高信息意识和信息观念,最快捷、最有效地获取自己最需要的信息,并利用这些信息顺利完成自己的工作任务。熟练地掌握文献检索的方法是缩短科研时间、提高工作效率的重要途径。科学研究是一种探索未知的活动,信息检索可以使科学研究避免重复。

2 网络信息检索类型

根据网络的地域范围分类,可分为局域网信息检

收稿日期:2007-11-12
基金项目:山东省自然科学基金(2006ZRB01001)
作者简介:马福晶(1979-),女,山东烟台人,助教,硕士,研究方向为信息管理。

索和万维网信息检索两种类型[4]。

局域网中信息检索以传统的联机光盘检索系统为主,指把单用户系统发展成多用户的局域网系统,通过网络(一指局域网,如图书馆网、校园网等)连接多个用户终端,用服务器管理多组光盘数据库及其检索系统。它可以连接到许多用户终端,网上用户可以分时共享光盘数据库的信息。联机光盘检索系统由若干台微机、光盘驱动器、光盘服务器、光盘数据库、检索系统软件、管理系统软件等构成。

随着网络技术的发展,具有全球性的分布结构、开放性的因特网为计算机检索提供了广阔的发展平台。这种检索方式可同时使用网上多个主机、甚于所有主机的某种资源而并不需要用户预先知道它们的具体地址。这就极大地拓宽了检索的空间和信息量,包括各种文献信息资源及其指向的网络页面。其中这类信息检索技术也因技术不同分为:基于 Web 的数据库检索;基于 Web 的分类浏览方式和链接嵌套方式;基于 Web 的搜索引擎方式。

以上综合讨论两种检索类型的特点和功能,现对它们在服务的主要方面进行比较,见表 1。Web 版联机检索从信息量的存储和数据更新及通讯方式等方面具有更强的优势,传统的联机数据库将更多的以 Web 版方式放到 Internet 网上。

表 1 两种检索类型的比较

	传统的联机检索	Web 版联机检索
检索界面	需要熟悉	直观简洁
检索方式	命令检索	超文本链接方式
开发系统结构	C/S 结构	B/S 结构
数据库类型	单一	多样化
被攻击指数	较低	较高
查准率	较高	较低,冗余大

3 搜索引擎技术

3.1 搜索引擎工作原理

搜索引擎具有对网络资源进行采集、标引并提供检索的功能,其基本结构如图 1 所示。

(1)数据采集模块:搜索、采集和标引网页。有人工采集和自动采集两种方式。人工采集由专门信息人员跟踪和选择有用的网页,并按规范方式进行分类标引。自动采集则是通过软自动采集器来完成的。网页自动标引借鉴了文献标引过程中的这样一种观点:即文献的主要内容可以用一些关键句的集合来表达(如摘要);关键句包含了最能反映文献主题的重要词汇;而词汇在文献中使用的次数,即词频则反映了词汇的重要程度。基于这一观点,网页自动标引是建立在词频统计基础之上的。目前几乎所有重要的搜索引擎都采用全文索引方式,分析网页的所有词汇,并依据词频、词汇在网页中出现的位置等确认词汇的权重,由此来选择标引词。

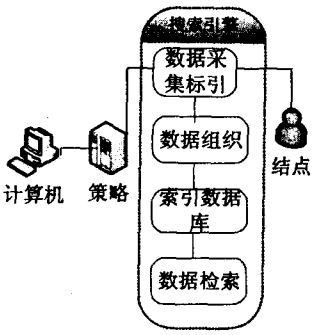

图 1 搜索引擎工作原理

(2)数据组织模块:通过数据库管理系统来组织所采集的网页信息,建立相应的索引数据库。索引数据库中的一条记录对应于一个网页,记录的内容包括网页标题、关键词,网页摘要及 URL 等信息。

(3)数据检索模块:根据用户检索要求,从索引数据库中检索出符合用户需要的网页。

此外,还有一种被称为"页面搜索器"的检索工具,工作原理类似于"Push"浏览器(http://www.netmind.com)。用户只要把自己感兴趣的页面地址输入"页面搜索器"中,并提供自己的电子邮件地址,"页面搜索器"就会定期检索。一旦发现相关页面的地址,"页面搜索器"就会自动将结果送入用户邮件地址。搜索的文件类型可以包括"http","FTP","Gopher"等。通过"页面搜索器"可以跟踪站点以及其内容的变化,以便得到最新的信息。

3.2 搜索引擎分类

按照信息搜集方法和服务提供方式的不同,主流搜索引擎系统可以分为三大类,这三类系统根据不同的检索建立原理分别解决了不少信息检索的问题,结合前面提到的 6 个标准总结其解决的问题:

(1)目录式搜索引擎:以人工方式或半自动方式搜集信息,由编辑员查看信息之后,人工形成信息摘要,并将信息置于事先确定的分类框架中。信息大多面向网站,提供目录浏览服务和直接检索服务。该类搜索引擎因为加入了人的智能,所以信息准确、导航质量高,缺点是需要人工介入、维护量大、信息量少、信息更新不及时。这类搜索引擎的代表是:Yahoo, LookSmart, Open Directory, Go Guide 等。

(2)机器人搜索引擎:由一个称为蜘蛛(Spider)的机器人程序以某种策略自动地在互联网中搜集和发现信息,由索引器为搜集到的信息建立索引,由检索器根

据用户的查询输入检索索引库,并将查询结果返回给用户。服务方式是面向网页的全文检索服务。该类搜索引擎的优点是信息量大、更新及时、毋需人工干预,缺点是返回信息过多,有很多无关信息,用户必须从结果中进行筛选。这类搜索引擎的代表是:AltaVista, Northern Light, Excite, Infoseek, Inktomi, FAST, Lycos, Google;国内代表为:"天网"、悠游、OpenFind等。

(3)元搜索引擎:这类搜索引擎没有自己的数据,而是将用户的查询请求同时向多个搜索引擎递交,将返回的结果进行重复排除、重新排序等处理后,作为自己的结果返回给用户。服务方式为面向网页的全文检索。这类搜索引擎的优点是返回结果的信息量更大、更全,缺点是不能够充分使用所使用搜索引擎的功能,用户需要做更多的筛选。这类搜索引擎的代表是WebCrawler, InfoMarket等。

除上述三大类引擎外,还有以下几种非主流形式:集合式搜索引擎;门户搜索引擎;免费链接列表(Free For All Links,简称FFA)[5]。

3.3 搜索引擎方式的新问题及发展趋势

(1)注意提高信息查询结果的精度,提高检索的有效性[6~8]。

用户在搜索引擎上进行信息查询时,并不十分关注返回结果的多少,而是看结果是否和自己的需求吻合。对于一个查询,传统的搜索引擎动辄返回几十万、几百万篇文档,用户不得不在结果中筛选。解决查询结果过多的现象目前出现了几种方法:一是通过各种方法获得用户没有在查询语句中表达出来的真正用途,包括使用智能代理跟踪用户检索行为,分析用户模型;使用相关度反馈机制,使用户告诉搜索引擎哪些文档和自己的需求相关(及其相关的程度),哪些不相关,通过多次交互逐步求精。二是用正文分类(Text Categorization)技术将结果分类,使用可视化技术显示分类结构,用户可以只浏览自己感兴趣的类别。三是进行站点类聚或内容类聚,减少信息的总量。

(2)基于智能代理的信息过滤和个性化服务。

搜索引擎搜索出来的内容真正被用户使用的,可能只有最前面很少的一部分,而用户真正感兴趣的内容却不能被找到。该问题的解决方案可以采用个性化服务技术[9]。通过收集和分析用户信息来学习用户的兴趣和行为,从而实现主动推荐的目的。对于搜索引擎返回的结果和它们本身所代表的原文档分别进行分类,结果发现误差为20%[10]。对于追求速度的搜索引擎来说,这是可以忍受的。但是由于搜索结果包含的信息较少,这对于个性化推荐是不利的,并且用户在使用系统时等待太长的时间是不合适的。文献[10]提出的后缀树聚类算法能够把搜索结果分离为具有较好内聚性的聚类,且该算法具有线性时间复杂度。文中将基于内容过滤的个性化推荐技术和后缀树算法相结合,实现了一个基于搜索结果的个性化推荐系统。利用领域分类模型上的概率分布表达了用户的兴趣模型,对搜索结果的聚类结果进行内容过滤,给出个性化推荐的文档。

(3)采用分布式体系结构提高系统规模和性能。

搜索引擎的实现可以采用集中式体系结构和分布式体系结构,两种方法各有千秋。但当系统规模到达一定程度(如网页数达到亿级)时,必然要采用某种分布式方法,以提高系统性能。搜索引擎的各个组成部分,除了用户接口之外,都可以进行分布:搜索器可以在多台机器上相互合作、相互分工进行信息发现,以提高信息发现和更新速度;索引器可以将索引分布在不同的机器上,以减小索引对机器的要求;检索器可以在不同的机器上进行文档的并行检索,以提高检索的速度和性能。

(4)重视交叉语言检索的研究和开发。

交叉语言信息检索是指用户用母语提交查询,搜索引擎在多种语言的数据库中进行信息检索,返回能够回答用户问题的所有语言的文档。如果再加上机器翻译,返回结果可以用母语显示。该技术目前还处于初步研究阶段,主要的困难在于语言之间在表达方式和语义对应上的不确定性。但对于经济全球化、互联网跨越国界的今天,无疑具有很重要的意义。

4 结束语

网络信息检索范围宽、信息量大、信息检索的时效性强,但是处理的信息类型繁杂而载体形式多样,所以搜索引擎的研究应符合时代要求,发展智能化、个性化和高效化,这是其发展方向和其亟待解决的问题。

参考文献:

[1] 赵玉玲,滕　飞.试论信息检索途径的多样性[J].重庆图情研究,2007(1):40-41.

[2] 乔振林.试论网络环境下的信息检索和服务[J].成功教育,2007(8):166-167.

[3] 张　帆.信息存储与检索[M].北京:高等教育出版社,2003.

[4] 焦玉英.信息检索进展[M].北京:科学出版社,2003.

[5] 崔　旸,张兆南.搜索引擎的工作机理及发展前景[J].科技资讯,2007(6):122-123.

[6] 郑京华.提高搜索引擎检索准确率的策略[J].科技情报开发与经济,2007(17):54-56.

(下转第117页)

步骤 4　对整个粒子群目前最优位置进行 P_i 排序,前 k 个作为精英集团 Ω_k;

步骤 5　每个粒子从精英集团 Ω_k 中,随机选取 P_r 作为 P_g;

步骤 6　如果适应度变坏,P_g 代入速度迭代方程,重新计算速度,否则,速度不变;

步骤 7　检查速度各个分量是否在[V_{min}, V_{max}]范围内,如果大于 V_{max} 设为 V_{max};如果小于 V_{min} 设为 V_{min};

步骤 8　按位置更新方程,更新粒子位置;

步骤 9　检查各个分量是否在[X_{min}, X_{max}] 范围内,如果超出,在[X_{min}, X_{max}] 内随机取一个值,设为该分量;

步骤 10　如果未达到预先设定的最大代数或未达到足够好的函数值,则返回步骤 2。

2.5　粒子群算法的应用

由于粒子群算法出色的性能,目前已广泛应用于函数优化、神经网络训练、模糊系统控制等众多领域。下面简要介绍粒子群算法在神经网络中的应用。

当粒子群算法用于神经网络训练网络权值时,粒子就表示神经网络的一组权,粒子的纬度就是神经网络中权值的个数。一般神经网络的初始值介于 -1 和 +1 之间,训练结束后的权值也是 -1 和 +1 之间,因此,粒子的范围可以设定为 -1 和 +1 之间。惯性因子 ω 的取值既要考虑到避免陷入局部极小,又要保证收敛性,算法的初期阶段让惯性因子 ω 取较大的值,有利于跳出局部极小点,逐步调整 ω,使其递减,以保证算法的收敛性。实验结果表明,粒子群优化算法训练的神经网络收敛速度明显加快。

3　结束语

群体智能是新兴的用于寻找全局最优解的算法,已经广泛地应用于许多领域,取得很好的效果。群体智能的特点和优点是:群体中相互合作的个体是分布式的,这样更能够适应当前网络环境下的工作状态;没有中心的控制与数据,这样的系统更具有鲁棒性,不会由于某一个或者某几个个体的故障而影响整个问题的求解。可以不通过个体之间直接通信而是通过非直接通信进行合作,这样的系统具有更好的可扩充性。由于系统中个体的增加而增加的系统的通信开销在这里十分小。系统中每个个体的能力十分简单,这样每个个体的执行时间比较短,并且实现也比较简单,具有简单性。因为具有这些优点,虽说群体智能的研究还处于初级阶段,并且存在许多困难,但可预言群体智能的研究代表了以后计算机研究发展的一个重要方向。

参考文献:

[1] Dorigo M, Gambardella L M. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem [J]. IEEE Transactions on Evolutionary Computations, 1997, 1(1):53 - 66.

[2] Gambardella L M, Dorigo M. Solving Symmetric and Asymmetric TSPs by Colonies[C]//In proceedings of the IEEE International Conference on Evolutionary Computation (ICEC '96). [s.l.]: IEEE Press, 1996:622 - 627.

[3] Kennedy J, Eberhart R C. Particle swarm optimization[C]// In: Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: [s.n.], 1995:1942 - 1948.

[4] Dorigo M, Maniezzo V, Colorni A. The ant system: optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 1996, 26(1):29 - 41.

[5] Bullnheimer B, Hartl R F, Strauss C. A New Rank - based Version of the ant system: A Computational Study[R]. Vienna: Institute of Management Science, University of Vienna, 1997.

[6] Stutzle T, Hoos H H. MAX - MIN Ant System[J]. Future Generation Computer Systems, 2000, 16(8):889 - 914.

[7] Colorni A, Dorigo M, Maniezzo V, et al. Ant system for job - shop scheduling[J]. Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL), 1994, 34:39 - 53.

[8] Costa D, Hertz A. Ants can color graphs[J]. Journal of Operational Research Society, 1997, 48:295 - 305.

[9] Durbin R, Willshaw D. An Analogue Approach to the Travelling Salesman Problem Using an Elastic Net Method[J]. Nature, 1987(326):689 - 691.

(上接第 113 页)

[7] 金玉坚,刘　焱. 新型网络信息检索效果评价指标体系设计[J]. 现代情报, 2005(4):185 - 188.

[8] 李振龙. web信息检索的技术分析与发展策略研究[J]. 计算机科学, 2006, 33(4):181 - 184.

[9] 卫　琳. 基于搜索结果的个性化推荐系统研究[J]. 计算机技术与发展, 2007, 17(9):66 - 70.

[10] Zamir O, Etzioni O. Web Document Clustering: A Feasibility Demonstration[C]//In: Proc. of SIGIR'98. New York: ACM Press, 1998:46 - 54.