

元搜索引擎研究

A Survey of Meta Search Engine

张卫丰 徐宝文 周晓宇 李 东 许 蕾
(东南大学计算机科学与工程系 南京 210096)

Abstract With the explosive increase of the network information, it is more and more difficult for people to look up information. The occurrence of the Web search engines overcomes this problem in some degree. However, because different search engines use different mechanisms, scope and algorithms, the repetition of the search results for the same query is no more than 34%. If wish to get relatively full-scale, accurate search results, multi-search engines should be used and the meta search engines occur. In this paper, the meta search engines are surveyed. At first, the history, the principles and the elements of the meta search engines are discussed. Then, the related criteria of the meta search engines are analyzed and several typical meta search engines are compared. Finally, on this base, the trend of the meta search engine is introduced.

Keywords Internet, WWW, Search engine, Meta search engine

1 引言

从早期的 ARPANet, 到目前的 Internet, 互联网的技术与应用都在超常规飞速发展。互联网上的信息也越来越多, 因此迫切要求一种信息检索工具。1991年, XWAIS 提供了一个界面友好的信息搜索系统, 但是这个系统要求很特殊的文件格式。在同一年出现了另外一个信息检索系统 GOPHER, GOPHER 一时成为最为流行的检索工具。由于 GOPHER 仍基于字符界面, 一般的互联网用户还是很少使用它。真正让互联网普及的转机出现在 1993 年。当年美国国家计算机安全协会 NCSA 推出第一个基于 HTML 语言的可以浏览图形的浏览器 Mosaic, 普通用户因而可以方便地使用互联网。次年, 美国网景公司推出免费浏览器 Netscape, 从而使得以 HTML 为格式的信息迅速膨胀。差不多同时 Yahoo 公司推出基于目录的信息检索服务, 然而这种服务还不完全具备通常意义上搜索引擎的功能。是年春天, M. Mauldin 将 J. Leavitt 的“网络蜘蛛”(spider)程序接入到索引程序中^[1-2]并研制成功了 Lycos, 这是第一个真正意义上的搜索引擎。

根据专家的评测, 目前主要搜索引擎返回的相关结果的比率不足 45%, 而且由于所采用机制、算法与适用范围等的不同, 导致同一个检索请求在不同搜索

引擎中的查询结果的重复率不足 34%。因此, 要想获得一个比较全面、准确的结果, 就必须反复调用多个搜索引擎^[3,4]。元搜索引擎的出现, 在一定程度上解决了这些问题。为了与元搜索引擎区别起见, 在下文中所指的独立搜索引擎均为基于因特网的搜索引擎。元搜索引擎通过调用其它独立搜索引擎来完成搜索功能。下面将首先探讨元搜索引擎的基本结构, 然后对元搜索引擎进行分类, 指出元搜索引擎与独立搜索引擎的差别, 接着分析元搜索引擎的主要指标并对几个典型的元搜索引擎进行了比较, 在此基础上讨论元搜索引擎所面临的问题以及它的发展趋势, 最后对我们在元搜索引擎研究方面的工作做了简要介绍。

2 基本构成

为了更好地说明元搜索引擎的构成, 先简单地介绍一下独立搜索引擎的工作原理和基本构成。独立搜索引擎根据用户的查询请求, 按照一定的算法从索引数据库中查找对应的信息返回给用户。为了保护用户查找信息的精度和新鲜度, 搜索引擎需要建立并维护一个庞大的索引数据库。独立搜索引擎中索引数据库中的信息是通过一种叫做网络蜘蛛(spider)的程序从互联网中采集得到的网页, 所以, 一般独立搜索引擎主要由网络蜘蛛、索引与搜索引擎软件等部分组成。

张卫丰 博士研究生, 主要从事程序设计语言、软件体系结构、网络语言等方面的研究, 徐宝文 教授, 博士生导师, 主要从事程序设计语言、软件工程、并行程序设计等方向的教学与科研工作。

与独立搜索引擎相比,元搜索引擎不需要维护庞大的索引数据库,也不需要网络蜘蛛去采集网页。具体说来,元搜索引擎主要由三部分组成(如图1所示),即:请求提交代理、检索接口代理、结果显示代理。“请求提交代理”负责实现用户“个性化”的检索设置要求,包括调用哪些搜索引擎、检索时间限制、结果数量限制等。“检索接口代理”负责将用户的检索请求“翻译”成满足不同搜索引擎“本地化”要求的格式。“结果显示代理”负责所有源搜索引擎检索结果的去重、合并、输出处理等。

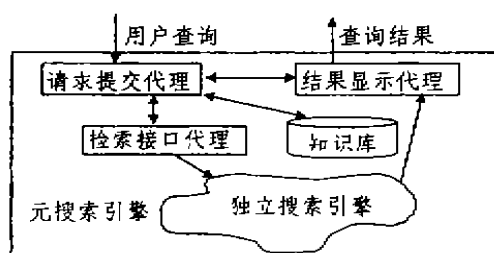


图1 元搜索引擎原理图

· 请求提交代理 负责将用户的请求分发给独立搜索引擎。

一般的元搜索引擎设定了它所调用的独立搜索引擎,比如 widewaysearch。有些元搜索引擎让用户自己选择所用的搜索引擎,还有一种通过分析用户的兴趣和网络的实际情况来选择搜索引擎,这有利于提高用户的查询的准确度和对用户的响应速度。有关如何智能化地选择搜索引擎的相关研究工作正在进行之中。文[14]提出通过遗传算法来智能化地选择搜索引擎。

· 检索接口代理 由于不同的搜索引擎所支持的查询方式不同,比如有些搜索引擎支持 stemming 方式,即便是同一种方式,也有不同的表达方法,所以必须将元搜索引擎中的查询请求映射到对应的搜索引擎中,而且使得语义信息不丢失。

· 结果显示代理 元搜索引擎的结果一般有网页标题、内容摘要、所指向网页 URL、相关度、信息返回时间、所采用的引擎标志等。这些搜索结果是多个独立搜索引擎的并集。元搜索引擎的结果应该具有多种排序方式以满足不同用户的需要。元搜索引擎常用的排序方式有:相关度排序、时间排序、域名分类排序、搜索引擎排序等^[5]。

元搜索引擎把用户的查询串分配给几个指定的独立搜索引擎,再将各独立搜索引擎所得结果分级排序,删去重复内容,然后给出查询结果,也就是说元搜索引擎是建立于独立搜索引擎之上的搜索引擎。

3 分类

元搜索引擎有多种分类方式,如可以根据用户应用模式、调用独立搜索引擎的方式等进行分类,我们根据图1中请求提交代理、检索接口代理和结果显示代理的复杂程度,将元搜索引擎可以分为简单元搜索引擎和复杂元搜索引擎;根据请求提交代理、检索接口代理和结果显示代理所在位置的不同,又将复杂元搜索引擎分为桌面型元搜索引擎和基于 Web 的元搜索引擎。

· 简单元搜索引擎 它实际上不能称为搜索引擎。它只是给用户提供一个搜索引擎列表,用户可以选择所用的搜索引擎。用户输入查询请求,然后直接以 CGI 方式调用对应的搜索引擎。由于是由用户选择搜索引擎,而且查询请求只能发送到一个搜索引擎,所以请求提交代理和检索接口代理的设计就非常简单,可以直接设计在静态网页中。简单元搜索引擎不进行搜索结果的处理,这就省缺了图1中的结果集成代理。其实将简单元搜索引擎叫做搜索引擎列表更恰当。简单元搜索引擎设计上的简单性,成为众多网站采用的提供搜索服务的方法,如 Freeality (<http://www.freeality.com>) 提供的搜索服务。

· 桌面型元搜索引擎 它以程序的方式提供给用户。它运行在用户的机器上,用户的查询请求直接由用户端分发给它所调用的搜索引擎,然后对返回的搜索结果进行集成后以一定的方式显示。对桌面型元搜索引擎来说,图1中的请求提交代理、检索接口代理和结果显示代理都在客户端。典型的桌面型元搜索引擎有 BeeLine (TransCom 公司) 和搜索至尊 (世纪伙伴软件工作室) 等。

· 基于 Web 的元搜索引擎 它以 Web 方式为用户提供元搜索服务,请求提交代理、检索接口代理和结果显示代理都存放在元搜索引擎所在的服务器端。这种方式中,用户的元查询请求经过服务器端的请求提交代理和检索接口代理将查询请求分发到它所调用的独立搜索引擎,这些独立搜索引擎返回的搜索结果由服务器端的结果显示代理处理后再返回给用户。由于 Web 易用性的特点,基于 Web 的元搜索引擎使用得更为广泛。

为了描述方便,在下文中如不作特别说明,元搜索引擎是指基于 Web 的元搜索引擎。国外具有代表性的元搜索引擎有 Dataware (<http://queryserver.dataware.com/general.html>)、Ixquick (<http://www.ixquick.com>)、Metor (<http://www.meteor.com>)、C4 (<http://www.c4.com>)、InfoZoid (<http://www.infozoid.com>)、Search.com (<http://www.search.com>)、

Dogpile (<http://www.dogpile.com>), MetaCrawler (<http://www.go2net.com/search.html>) 和 ProFusion (<http://www.profusion.com>) 等。国内这几年也相继出现了几个中文元搜索引擎,代表性的有万维搜索 (www.widewaysearch.com)。

4 与独立搜索引擎的比较

从图1中元搜索引擎的结构中,可以看出,元搜索引擎的技术重心在于查询前的处理(检索请求提交机制和检索接口代理)和结果的集成。元搜索引擎可以灵活地选择所要采用的独立搜索引擎。它一般都是选择那些比较典型的、性能优异的独立搜索引擎。这种强强联合的结果保证了搜索结果的权威性和可靠性。它还可以充分发挥各个独立搜索引擎在某个搜索领域的功能,弥补独立搜索引擎信息覆盖面上的局限性。总的来说,元搜索引擎与独立搜索引擎相比,具有如下主要优点。

- 信息的覆盖面 元搜索引擎一般缺省调用它自己认为比较好的几个搜索引擎,而且大多数的元搜索引擎都提供给用户在一定范围内选择搜索引擎的功能。有些元搜索引擎还以频道的方式为用户提供专业搜索引擎的分类。这样用户可以根据自己的喜好和要查询的内容选择相应的搜索引擎。

- 搜索结果的权威性和可靠性 独立搜索引擎索引数据库的更新需要一定的周期,而且搜集的信息也各有一定的侧重,元搜索引擎调用多个独立搜索引擎获取搜索结果,这种方式首先保证了信息的互补性,其次与独立搜索引擎相比,提高了信息的新鲜度。如果同样的搜索结果在多个独立搜索引擎中同时出现,那么说明这个搜索结果比较重要。这样避免了有些独立搜索引擎人工干预有些搜索排名的缺点,使得搜索结果的排序更加公正。有些元搜索引擎还检查搜索结果链接的存在性,这样可以保证用户得到的元搜索结果的可靠性^[5]。

- 易维护性 是针对元搜索引擎的管理者而言的。元搜索引擎省去了独立搜索引擎中收集和存储网页、建立和存储索引的工作。它将它所调用的搜索引擎看成一个可以独立完成一定功能的实体,本身不需要去维护,只需知道它们的调用接口。元搜索引擎的查询精度在很大程度上在于它所调用的搜索引擎的查询精度。所以,元搜索引擎可以把主要精力放在搜索引擎的选择、查询请求的优化和搜索结果的优化上。一般的元搜索引擎都提供了对应的优化机制。

如果要尽快查询到一个独特的术语或某个课题的概述,用元搜索引擎。当用其他独立搜索引擎查询而得不到所需文件时,可改用元搜索引擎。即元搜索引擎主

要用于提高搜索的广度。对其它搜索引擎不是很熟悉的时候,也可以使用元搜索引擎作为通向其它搜索引擎的门户。

5 主要指标及其分析

作为一种搜索引擎,元搜索引擎也有搜索引擎的一些基本指标,如响应速度、准确率等。但是元搜索引擎个体差异很大,很难进行精确的比较^[4]。下面给出元搜索引擎的几个主要指标并对其中的一些指标进行比较。

- 选择独立搜索引擎的策略 有些元搜索引擎固定地调用几个独立搜索引擎,用户不能修改。有些元搜索引擎的高级特性中让用户选择调用哪些搜索引擎。这种方式对于那些对独立搜索引擎比较了解的用户来说,是可取的,而对于不了解的用户来说,可能选择不到适合自己查询的搜索引擎。独立搜索引擎的各种技术飞速发展,性能也随之不断地提高。元搜索引擎只能选择几个(一般不超过16个)搜索引擎同时进行检索,因为选择的搜索引擎越多,固然得到的搜索结果更全面,但是结果的集成将花费大量的时间。元搜索引擎如果一直固定地调用几个搜索引擎,将可能跟不上搜索引擎的发展潮流。文[14]中提出了通过遗传算法实现独立搜索引擎的自动调度的方法是解决这种矛盾的有效途径。

- 覆盖网络资源的广度 元搜索引擎由于不需要建立自己的索引,避免了大量信息的存储和处理。一般的元搜索引擎尽量覆盖多种网络资源。有些元搜索引擎还支持更加专业的搜索,比如Mp3,各种专业的论文查找、健康医药查找等搜索引擎^[4]。

- 是否提供足够的检索选项 包括是否提供高级检索服务,是否可以限定最长检索时间,是否可以设置每个搜索引擎返回的结果数量,是否可以设置每页显示的结果数目,是否可以设置标题大小(搜索引擎可以从title标记中显示的最大字符数)和摘要大小(搜索引擎所显示的结果中摘要的最大字符数),是否提供显示选项(用户可以通过它来设置结果的其它显示方式,如只显示标题、按照时间排序等)等。检索选项越多,用户使用的时候就越灵活。但是由于元搜索引擎的检索特性向它所调用的独立搜索引擎检索特性转换所具有的复杂性,许多元搜索引擎不提供复杂的检索特性。大多数元搜索引擎提供通用的布尔检索。而对于如高级布尔检索、短语检索、自然语言检索等高级特性则只有少数几个元搜索引擎能够提供,如Dataware和Ixquick等。

- 对搜索结果的处理能力 对独立搜索引擎返回的搜索结果的处理是元搜索引擎的又一重要技术^[16]。

它包括结果的处理和结果的显示。有些元搜索引擎提供多种显示结果的方式,如有些元搜索引擎提供方式让用户按照时间、按照搜索引擎、按照相关度等来排序,如国内第一个元搜索引擎--万维搜索。有些元搜索引擎提供了让用户定制搜索结果的聚类方式,如按照域名聚类、按照主题分类等。

• 相关度指标 每个搜索引擎开发商为了将最满意的结果放到越前面,不遗余力地创建出各种相关度指标体系,从检索词的位置/频率到链接和流行度^[8]等等。虽然没有一种方法是完美的,但都有创新和独到之处。面对这些众多的相关度评价指标,按照怎样的方式对从独立搜索返回的结果进行一致性的排序是元搜索引擎结果处理部分面临的主要问题。元搜索引擎的结果排序有多种方法^[9],有根据搜索结果在源搜索引擎

中的位置进行排序的方法,有根据搜索结果的摘要信息进行排序的方法^[10],还有的干脆获取这些网页,然后按照位置/频率法对搜索结果进行一致性排序^[3]。Ixquick 在肯定各个独立搜索引擎所用的相关度指标的基础上,通过统计搜索结果记录被多少个独立搜索引擎所青睐,作为元搜索结果相关度评价指标,简称“星星体系”(如果在一个搜索引擎的前几条记录中出现,就得一个星。得到的星越多,则该记录越重要)。

利用以上指标,我们对几个典型的元搜索引擎进行分析比较,得到如表 1 所示的结果。前四个元搜索引擎在国外非常流行,获得较高的评价。万维搜索引擎作为国内第一个元搜索引擎,与国外的元搜索引擎相比,在一些指标方面还存在一定的差距。

表 1 元搜索引擎比较

搜索引擎 指标	Ixquick	Metor	C4	InfoZoid	万维搜索
可供调用搜索引擎的个数	14	10	16	15	11
缺省调用的搜索引擎个数及分类	14 分成四类: Web、New、Mp3、图片	10 提供 24 个分类频道	16 提供 11 个类别的分类检索	15 提供 Web、usenet、ftp 搜索	11
查询方式	支持 and、or、not 等查询命令 将查询请求翻译到能够理解用户查询的引擎 支持各种形式的统配符 支持自然语言	支持完全匹配,部分匹配和短语搜索 可以设置最大结果等待时间	提问式查询和关键词查询 提供个性化搜索服务	支持完全匹配,部分匹配和短语搜索	支持完全匹配,部分匹配 可以设置最大结果等待时间
结果显示方式	分页显示 结果统计 网页标题 内容摘要 URL 来源及在对应搜索引擎中的位置	网页标题 相关度 内容摘要 还提供对应链接的检查、分析和浏览功能	结果统计 分页显示 网页标题 内容摘要 URL 信息来源(图形化表示)	网页标题、内容摘要的显示可选	提供可选的分页显示功能 结果统计 网页标题 内容摘要 URL 相关度 信息返回时间 搜索引擎标记 结果来源
结果排序方式	相关度结合“星星”评价指标	相关度结合“星星”评价指标	相关度	相关度,搜索引擎	时间、搜索引擎、相关度、域名

6 问题和发展趋势

元搜索引擎的出现可以大大提高用户查询信息的覆盖面,但是元搜索引擎返回的结果只能满足“最低常用分母”——即它不能提高搜索结果的准确性。应当说,元搜索引擎在查准率上不一定强于其他独立搜索引擎,但是由于能够在多个搜索引擎中搜索,必然能够

为搜索提供更多的机会。具体说来,元搜索引擎的局限性主要体现在以下几个方面:

• 调用搜索引擎上的局限性 元搜索引擎大多只调用如 Google、AltaVista、Infoseek 等几个主要的搜索引擎。即便有些元搜索引擎提供了让用户选择搜索引擎的选项,但是只能在其提供的几个搜索引擎中进行选择。

· 元搜索引擎查询请求上的局限性 元搜索引擎在查询语法转换上的局限性限制了许多高级的查询语法,因为元搜索引擎的查询语法要考虑到它所调用的所有搜索引擎的查询语法,这种复杂性使得大多数元搜索引擎只能提供一些简单的查询语法。

· 查询结果数量上的局限性 由于网上信息的急速膨胀,搜索引擎搜集的信息越来越多,一个查询往往能够返回成千上万的搜索结果。元搜索引擎需要对多个搜索引擎的返回结果进行处理,考虑到系统运行的效率,大多数元搜索引擎只取每个搜索引擎返回的前几个的搜索结果。这样必定影响了信息的覆盖面。有些元搜索引擎提供了选项,以便设置从每个搜索引擎返回结果的数量。

· 查询结果排序上的局限性 由于不同的搜索引擎在收集信息数量、收集信息的范围、采用的索引方法、使用的相似度评价等方面的巨大差异,再加上元搜索引擎的设计者不能获取这些搜索引擎的好多技术细节,对不同搜索引擎返回的结果不能进行很好的排序。

针对以上问题,有人提出了智能化元搜索引擎,以提高元搜索引擎对用户的适应能力,提高用户的满意度。

文[15]中提出了通过自动采集用户的兴趣,然后对用户的搜索结果进行过滤以提高用户搜索精度的方法。Pinkerton 通过实验指出^[7],用户使用搜索引擎平均输入的关键词的个数为 1.5 个。这将使得用户得到成千上万看似相关的结果。Chia Hui Chang 在文[1]中指出通过对搜索结果建立聚集的方法和记录用户查询的模式来提高对用户的精度。通过对搜索结果中抽取重要的词汇以及利用用户兴趣的反馈来修正查询,对搜索结果建立目录层次空间来缩小用户的搜索空间。

由于人们各自感兴趣的领域不同,各自对词意的理解也不尽相同,不同的用户对同一个检索请求得到的检索结果有不同的评价。智能化元搜索引擎应该对不同的用户在相同的检索请求下有不同的检索结果,即对用户具有自适应能力,这可以通过两种方案实现,其一、需要系统在检索请求提交数据库之前智能化地调整查询表达式和查询域,即查询预处理;其二、在查询结果返回时候,智能化地对搜索结果进行处理后再返回给用户。

6.1 查询预处理

将元查询请求映射到对应搜索引擎的查询请求,关键要保证用户的查询信息不缺损。文[2]中提到一种用户查询请求在不同信息源之间映射的方法,但是有些情况下用户的查询意愿肯定会有所缺失^[7]。这样查询的结果将很难符合用户的要求。这可以通过两种方法得到弥补。

其一,在映射过程中,根据元搜索引擎所调用引擎的特点,通过加强查询的语意来实现。比如用户在元搜索引擎中查询的时候,选择了 stemming(词干提取)选项,但是对应的搜索引擎不支持 stemming 选项。如果直接将元搜索请求提交给该搜索引擎的话,将导致用户意愿的缺损。这样搜索得到的结果可能没有用户所要求的,对于这种情况,有些元搜索引擎在元搜索请求向该搜索引擎请求的转换过程中,首先对元搜索请求进行 stemming 处理,然后进行 Extracting 处理。这样就扩充了用户查询请求,使得从该搜索引擎返回的结果符合用户的意愿。这实际上是元搜索引擎的搜索前处理能力。

其二,利用元搜索引擎的搜索后处理能力,对于从独立搜索引擎返回的结果进行再排序。文[16]中对元搜索引擎的结果生成技术进行了比较,并提出了几种改进的元搜索结果集成技术。这种方式主要在于过滤掉搜索结果中大量的冗余信息,以提高用户查询的精度。例如,用户在利用元搜索引擎查询的时候,明确地选择了不用 stemming 技术,但是好多独立搜索引擎都缺省采用了 stemming 技术。这样从这些独立搜索引擎返回的结果中包含了很多用户不需要的信息,这样将大大降低查询的精度。有些元搜索引擎为了提高用户查询的精度,往往根据元查询请求中的信息对从独立搜索引擎返回的搜索结果进行过滤。

关于搜索查询的精确表达,INM 公司(<http://www.ineural.com/entrance.html>)提出了基于概念的搜索引擎,它主要利用语言与认知学、句子与句子及句子内部的消歧等技术对用户查询请求中概念的精确表达和理解来获得跟用户紧密相关的搜索结果。

6.2 搜索结果集成

对独立搜索引擎所返回的搜索结果不能很好集成的主要原因在于这些搜索引擎是相互独立的,元搜索引擎不能获知它们的技术细节。虽然有些搜索引擎定期公布自己的一些技术细节,也有一些专门的站点^[8]来评测一些著名的搜索引擎,元搜索引擎的设计者可以获取这些著名搜索引擎的一些技术细节,但是元搜索引擎不仅仅调用这几个著名的搜索引擎,而且这些搜索引擎只是公布其中技术细节的一小部分而已,它们所采用的技术也在不断发展变化之中。元搜索引擎为了提高对用户的服务性能,经常需要调用一些专业的搜索引擎,而这些搜索引擎的技术细节很难得到。

元搜索引擎与其寻找这些搜索引擎的技术细节,不如通过一种科学的统计学的方法来检测它们所采用的某些技术^[4]。对于结果排序来说,要检测出这些搜索引擎所用的相似度函数,然后给出一个统一相似度函数对所有的搜索结果进行重新评价。检测出来的技术

细节虽然不可能很精确,但是可以大概反映该搜索引擎的技术状况。

为了使得得到的元搜索结果更加紧凑,可以对搜索结果按照一定的方式聚类^[1]。这样用户可以根据感兴趣的类别找到对应的搜索结果。随着网络速度的不断提高,元搜索引擎可以首先获取搜索结果所指向的页面,然后按照自己的相似度评价指标对搜索结果进行排序后再返回给用户^[1]。

元搜索引擎的核心就是搜索前预处理和对搜索结果的集成。未来的元搜索引擎应该在如下方面有所突破:

- 自然语言,精度更高 自然语言的输入将更加方便用户的使用,更易于用户与搜索引擎的交互,自然语言更能贴切地表达用户的查询需求,从而有利于提高查询的精度。

- 多种语言搜索 元搜索引擎可以充分利用在搜索预处理方面的优势,按照语言的不同将用户的查询请求发送到对应语言的独立搜索引擎,以提供多语言的搜索功能。

- 善解人意,学习个人喜好 元搜索引擎通过不断地学习来掌握用户的喜好。通过对用户的搜索习惯、用户兴趣的掌握,达到自动选择合适的独立搜索引擎的效果,从而改进搜索结果。

- 多通道输入 元搜索引擎应该利用其在搜索预处理上的能力,对于用户输入的查询,如声音、图像、视频等,实现查询的转换功能,从而得到预想的结果。

- 个性化和本地化 未来的元搜索引擎应该充分发挥搜索预处理和搜索结果集成方面的能力,考虑人的性别、年龄、地域等方面的差别,给出个性化的搜索结果。

7 我们的工作

近年来,我们在对 Web 技术作了初步研究的基础上^[3~12],又尝试将数据挖掘技术、Agent 技术和遗传算法等应用于智能 Web 搜索引擎的研究中,并取得了初步成果^[13~20]。

我们通过数据挖掘技术来提取用户的兴趣,然后根据用户的兴趣来过滤搜索引擎所返回的结果^[17,19]。这使得用户所得到的搜索结果善解人意,可以满足用户的个人喜好。

用户的反馈信息对于提高搜索引擎的精度非常重要。我们通过建立用户对搜索结果的反馈机制^[15],使得搜索引擎不断学习用户的兴趣信息,以此来提高搜索引擎对用户的自适应能力。

元搜索引擎通过调用其它独立搜索引擎来实现搜索,而选择哪些独立的搜索引擎是个非常棘手的问题。

遗传算法是通过模拟生物的进化过程来实现问题优化的一种方法,我们通过遗传算法实现了对独立搜索引擎的合理调度^[14]。

我们将来的工作将主要集中在元搜索引擎搜索请求的预处理和搜索结果的后处理上。通过对搜索引擎的机器学习,来达到对这些搜索引擎更进一步的掌握。通过对用户搜索习惯、用户兴趣的学习,达到改进搜索结果的目的。我们设计的 WIRPL 语言^[9]可以方便元搜索引擎对独立搜索引擎的处理。

结束语 本文介绍了搜索引擎和元搜索引擎的发展历史,讨论了元搜索引擎的基本工作原理并对元搜索引擎进行了分类,分析比较了元搜索引擎与独立搜索引擎的优点,讨论了元搜索引擎的几个关键指标并对几个典型的元搜索引擎的具体指标进行了比较。在此基础上,我们分析了元搜索引擎面临的问题和将来的发展趋势,我们还介绍了我们在搜索引擎方面所做的工作,我国对于元搜索引擎已经有了初步的研究,并且出现了一些元搜索引擎。但是与国外的元搜索引擎相比,在很多指标上还存在很大的差距,我们将对元搜索引擎进行进一步的研究。未来的元搜索引擎应该覆盖面更广、搜索速度更快、搜索精度更高和能够满足用户个性化的要求。

感谢英国 De Montfort 大学 Hongji Yang、Staffordshire 大学 Kecheng Liu、台湾东海大学 William C. Chu 以及香港大学 William Song 等教授与我们在 Web 搜索引擎研究方面的友好合作。

参考文献

- 1 Chang Chia Hur, et al. Customizable Multi Engine Search Tool with Clustering. Sixth International World Wide Web Conference. Available at: <http://www6.ntt-labs.com/hypernews/get/paper53.html>
- 2 Chang C, Garcia-Molina H. Mind your vocabulary: Query Mapping Across Heterogeneous Information sources. In: Proc SIGMOD99. Philadelphia, PA, June 1999. 335 ~ 346
- 3 Huang Lienlung, et al. ADMIRE: An Adaptive Data Model for Meta Search Engines. Available at: <http://www9.org/w9edrom/165/165.html>
- 4 Selberg E, Etzioni O. Multi-Engine Search And Comparison Using The MetaCrawler. In: Proc. of the Fourth World Wide Web Conference '95, Boston USA, Dec. 1995
- 5 Steve L, et al. the NECI meta search engine, In: Seventh Intl. World Wide Web Conf. Brisbane, Australia, Elsevier Science
- 6 Daniel D D. Integrating Heterogeneous WWW Search Engines. May 1995. Ftp: //132.239.54.5/savvy/report.ps.gz

(下转第 35 页)

每个记录 100 字节, $P=500$ 记录, $M=4$, 各结点机可用内存容量为 IM , 由于当 $N < 16$ 时系统的主机内存容量不足, 因此部分数据需要在磁盘中存储, 而 IO 的代价相对网上代价和内存访问代价要大得多, 此时预留显然增加了这一部分的开销, 均分以后使得数据集变小, 又可以并行处理, 从而降低 IO 代价, 此时宜采用均分法。当 $N > 16$ 以后, 由于系统总主存容量大于数据量, 此时又呈现出图 1 中讨论的情况, 在一段范围内, 预留数据将比较好。

另外, 通过这两幅图我们还注意到网络速度的提高对网络并行数据库有着很重要的作用, 一旦网络速度足够快, 或者在 DSM 结构下, 预留数据就没有必要了。所以我们认为, 预留数据的数据分配方案是在目前网络速度不够快的情况下, 利用网络并行环境设计并行数据库时, 减少数据分配时数据偏斜的一种权宜之计。实际操作中, 我们应该根据网络速度、总内存容量、以及系统中最多可用结点数等因素, 合理地确定采用预留数据法还是采用均分数据法, 在预留数据法中, 再根据相应的计算, 确定参与运算的结点数。

结论 通过上述讨论我们提出了一种在网络并行环境下数据均衡分配的方案。该方案考虑到了目前网络速度与内存存取速度之间的差异带来的数据偏斜问

题, 提出一种预留数据的分配方案, 并经过进一步的探讨, 说明在一定的特定条件下, 预留数据分配方案确实能收到较好的性能。同时, 文中还讨论了在网络并行环境的规模、网络速度、总内存容量变化的情况下, 如何合理选择数据分配方案的问题。

参考文献

- 1 廖朝晖, 张鹏, 王珊. 并行数据库系统: 目标、体系结构及研究方向. 计算机科学, 1994, 21(6): 23~28
- 2 Hua K A, Su J X W. Dynamic Load Balancing in Very Large Share-Nothing Hypercube Database Computers. IEEE Trans. On Computer, 1993, 42(12): 1425~1439
- 3 金树东, 冯玉才, 王元珍. 并行数据库系统的数据重组研究. 小型微型计算机系统, 1998, 19(3): 28~33
- 4 杨全胜, 徐宏炳, 王能斌. 网络并行集群中的多路合并算法研究. 计算机工程与科学, 1997, 19(A1): 94~97
- 5 Wen Zhaofang. Multiway Merging in Parallel. IEEE Trans. Parallel and Distributed Sysrem, 1996, 7(1)
- 6 孙家昶, 张林波, 等. 网络并行计算与分布式编程环境. 北京: 科学出版社, 1996
- 7 陈数清. 并行计算机的现状与发展趋势. 中国计算机学会、当代计算机体系结构、操作系统的发展学术研讨会, 1996

(上接第 41 页)

- 7 Pinkerton B. Finding What People Want. Experiences with the WebCrawler. In: Second Intl. WWW Conf. '94, July 1994, Chicago, USA, Oct. 1994. <http://info.webcrawler.com/bp/WWW94.html>
- 8 <http://www.searchenginewatch.com/webmasters/rank.html>
- 9 张卫丰, 徐宝文. Web 搜索引擎框架研究. 计算机研究发展, 2000, 37(3): 376~378
- 10 张卫丰, 徐宝文, 周晓宇. Web 页面中的计数器研究. 小型微型计算机, 2000, 21(10): 1096~1099
- 11 张卫丰, 徐宝文, 周晓宇. Web 页面中元素间交互技术研究. 计算机工程, 2000, 26(8): 62~64
- 12 张卫丰, 徐宝文, 许蕾. Web 页面安全性技术初探. 计算机工程与应用, 已录用
- 13 张卫丰, 徐宝文, 许蕾. 利用 Agent 个性化搜索结果. 小型微型计算机, 已录用
- 14 张卫丰, 徐宝文, 周晓宇, 管宇, 许蕾. 基于遗传算法的搜

索引擎调度, 待发表

- 15 张卫丰, 徐宝文, 周晓宇, 许蕾. 带反馈自适应 Web 搜索引擎研究, 待发表
- 16 张卫丰, 徐宝文, 周晓宇, 许蕾, 李东. 元搜索引擎结果生成技术研究, 待发表
- 17 张卫丰, 徐宝文, 周晓宇, 李慎之, 许蕾. 数据挖掘技术在 Web 预取中的应用研究, 待发表
- 18 Zhang Weifeng, Xu Baowen, Yang Hongji, Chu William C. A Genetic Algorithm Based General Search Engine. In: Proc. of IEEE MSE'2000
- 19 Zhang Weifeng, Xu Baowen, Chu William C, Yang Hongji. Data Mining Algorithms for Web Pre-Fetching. In: Proc. of The Workshop on the World Wide Web Semantics (WebSem'2000)
- 20 Xu Baowen, Zhang Weifeng, Chu William C, Yang Hongji. Application of Data Mining in Web Pre-Fetching. In: Proc. of IEEE MSE2000