●张 辉 隋 佳

基于 Z39.50 的元搜索引擎优化策略

摘 要 元搜索引擎具有扩大查询信息覆盖面、提高查全率,保证检索结果权威性、可靠性,以及操作简便的优点。基于 Z39.50 的元搜索引擎可实现检索、优化策略,具有屏蔽具体转换过程、形成统一操作模式,提高检索效率、质量及系统安全性等优点。图 2。参考文献 4。

关键词 元搜索引擎 Z39.50 优化策略 机检

分类号 G354.4

ABSTRACT In this paper, the authors introduce advantages of meta search engines, especially the advantages of the meta search engines based on Z39.50, such as search optimization, process filtering, unified platform and high search efficiency. 2 figs. 4 refs.

KEY WORDS Meta search engine. Z39.50. Optimization strategy. Computer retrieval.

CLASS NUMBER G354.4

根据专家的预测,目前主要搜索引擎返回的相关结果的比率不足 45%,而且由于所采用的机制、算法与适用范围等不同,导致统一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。元搜索引擎的出现,在一定程度上解决了这一问题。

1 元搜索引擎及其评价

元搜索引擎(Meta Search Engine)也称索引搜索引擎、集成搜索引擎。元搜索引擎将多个独立搜索引擎看成一个整体*,为用户提供一个统一的界面,用户只需提交一次检索请求,由元搜索引擎负责处理后提交给多个预先选定的独立搜索引擎,并将所有查询结果集中起来,以统一的格式呈现到用户面前。目前比较有代表性的元搜索引擎有:Dataware, Ixquick, Metor, C4, InforZoid,万维搜索等等。

独立搜索引擎一般由 3 部分构成:网络蜘蛛、索引和搜索引擎软件。独立搜索引擎在工作时一般采用集中的方式,它们用网络蜘蛛遍历因特网的信息,并对文档按照一定规则建立索引,然后根据用户的需求检索索引库并将检索结果提交给用户。集中的工作方式有信息覆盖面窄、搜索范围有限、更新维护困难等局限。与独立搜索引擎的工作方式不同,元搜索引擎采用分布式的信息检索策略**,分布式检索可以检索分布在因特网上不同位置的资源,可以在一定程度上扩大信息的覆盖率,提高检索结果的准确性。

具体来说,元搜索引擎主要由 3 部分组成(如图 1 所示):①请求提交代理:它可以实现用户个性化的检索请求设置,包括调用哪些搜索引擎、检索时间的显示、结果数量的限制等等;②检索接口代理:它负责将用户的检索请求转化为独立搜索引擎能够识别的格式,然后把检索请求发放给选定的独立搜索引擎;③结果显示代理:元搜索引擎将独立搜索引擎返回的结果通过各种算法进行去重、合并、排序等以统一的格式呈现给用户。

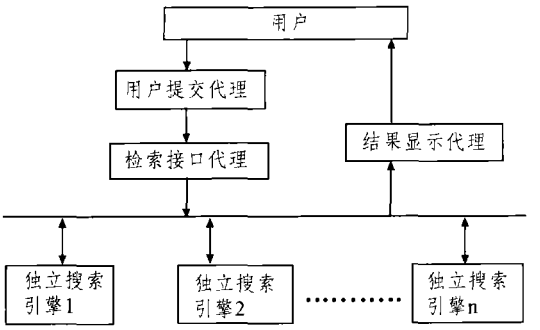

图 1 元搜索引擎工作原理

元搜索引擎的优点有三:

第一,独立搜索引擎由于本身的技术和因特网资源的动态性的限制,每个独立搜索引擎只能覆盖网上部分的信息资源。元搜索引擎的出现充分发挥了各个独立搜索引擎在某个搜索领域的功能,扩大了用户查询信息的覆盖面,提高了信息的查全率。

第二,元搜索引擎在进行检索时只能选择几个(一般不超过 16 个)独立搜索引擎同时进行检索。所以元搜索引擎一般都调用它自己认为比较优秀的独立搜索引擎,这在一定程度上保证了检索结果的权威性和可靠性。

第三,元搜索引擎与独立搜索引擎相比,省去了网络蜘蛛搜集网页和建立索引库的工作。使用元搜索引擎无需在服务器端建立大容量的存储设备,在普通的服务器上运行即

* 我们将普通搜索引擎视为独立搜索引擎,以区别于元搜索引擎。

** 分布式信息检索是指由检索代理程序将检索任务同时提交给网络上的多个主机,由位于这些主机上的检索程序分别独立检索并将结果返回到检索代理程序。

可。另外工作人员不需要花很大精力去维护庞大的数据库，而把精力放在算法的选择和结果的优化上，可以提高检索结果的准确性。

元搜索引擎的局限主要有：

第一，检索指令转化的局限。由于元搜索引擎工作时需要同时调用几个独立搜索引擎，每个独立搜索引擎都有自己的检索机制和查询语言，元搜索引擎在接受用户的检索请求之后，需要将用户的检索请求转换成与独立搜索引擎相适应的检索指令。由于系统要同时适应不同的检索策略，必然会牺牲某些搜索引擎的特殊性能，从整体上降低元搜索引擎整体性能。

第二，查询结果处理的局限。元搜索引擎在对返回的结果进行处理时，并不是把所有的检索结果都罗列出来，它需要对这些结果进行去重、合并、排序等处理，这会大大延长用户的等待时间。为了提高系统的效率，大多数元搜索引擎只返回每个独立搜索引擎的前几个检索结果，如此一来会大大缩减信息的覆盖面。

2 元搜索引擎搜索结果的优化策略

从元搜索引擎的结构中我们可以看出，元搜索引擎的技术重心在于查询前的处理（检索请求机制和检索接口代理）和结果的集成。目前有很多关于结果的优化策略。有人提出了通过自动采集用户的兴趣，然后对用户的检索结果进行过滤，以提高用户搜索精度的方法；也有人指出通过采用Agent优化技术来提高用户查询的精度。具体来说，有人从考虑各个独立搜索引擎所给出的相关度，从而消除各个数据源本身带来的偏差这个角度提出了一种新的基于概率模型的排序优化方法，利用贝叶斯规则，结合各独立搜索引擎组成系统平均执行性能的信息，推导出一种新的相关度计算公式，较好地解决了结果融合中相关度规范化和均衡化的问题。采用基于概率的检索结果的排序方法，最大的优点是元搜索引擎的搜索覆盖率增大。但是这种方法对系统的响应时间并没有太大变化，并不能提高系统的响应速度。而且不同的人对相关度有不同的测定方法，即使是同一个文件，不同的人对它的相关度赋值也不一样。所以采用这种方法很难进行准确测定。

有人提出一种基于Agent的元搜索引擎结果优化技术，考虑到Agent具有能够进行高级问题求解，可随环境变化修改自己的目标、学习知识并提高能力等智能特性。通过Agent的逐步学习，了解用户兴趣之所在，并以此为依据对元搜索引擎的检索结果进行过滤、合成和排序。基于Agent的元搜索引擎结果优化技术旨在通过建立兴趣模型对检索结果进行优化，从而针对不同用户提供更具个性化的信息。

这种方法的优点是：元搜索引擎通过不断学习来掌握用户的兴趣和喜好，检索结果能更好地满足用户的个性化需求。

该系统在建立用户兴趣模型的时候有两种方式：一是通过用户主动提出自己的兴趣爱好来建立模型；二是通过日志文件，观察用户所访问的页面并从其中挖掘相关信息从而建立模型。前者是通过人机交互的方式，对用户提出一系列问题，根据所选择的答案，启发式地转到下一个问题，这样根据问答结果对用户的兴趣进行综合评价和聚类。由于目前的人工智能技术还不是很完善，根据这种方式建立的兴趣模型很难精确地把握用户的兴趣和喜好。后者是从大量的历史数据中提取信息，一般需要的时间较长，而且这些数据都是历史数据，不能更好地反映用户当前的需求。由于用户的兴趣广泛而且易变，如果对用户的信息不能及时更新，也会影响检索的结果。

3 基于Z39.50的元搜索引擎的优化策略

3.1 Z39.50协议及其原理

Z39.50标准由美国国家标准化组织（NISO）于1998年公布，其全称是“信息检索（Z39.50）应用服务定义与协议描述”。它是一种基于ISO/OSI参考模型的应用层协议，当时提出这个标准主要是解决书目信息检索系统之间的通讯问题。Z39.50协议是计算机系统之间相互联系的一系列标准，是在网络上传输数据的高层协议，它不涉及数据库的名称和具体结构，也不考虑数据库的具体实现，独立于任何特定类型的信息或特定类型的数据库系统，能适用于不同数据源、不同数据格式之间的数据交换，便于实现信息的分布式检索。

Z39.50协议中主要是定义了一系列应用协议数据单元APDU，客户端和服务器进行交互联系都通过APDU。不同APDU单元完成不同的功能，主要由创建请求，创建响应；查询请求，查询响应；提交请求，提交响应三组APDU组成。一般是客户端发出请求APDU单元。服务器端返回响应数据单元。通过Z39.50协议进行信息检索的基本步骤有以下三步：首先，源端通过发送创建请求APDU来申请与目的端连接，服务器端收到后将响应一个APDU，客户端收到响应的APDU就可以判断目的端是否同意连接并得到一些基本信息。然后客户端在查询请求APDU里包含查询的具体内容，形式上采用逆波兰表达式，在接收到目的端的查询响应APDU后可以得知命中记录个数。如果存在命中记录，则在服务器端自动生成一个结果集。最后在提交请求APDU里写明要提交的记录范围，我们便可以在提交响应APDU里获得与查询相对应的详细记录信息。

3.2 基于Z39.50的元搜索引擎的运行机制的设想

近年来，Z39.50在因特网上也有了广阔的应用前景。因特网为Z39.50提供了一个较好的网络介质，使网上服务能够真正地做到跨地域甚至全球化，达到最广泛的空间。由于Z39.50标准本身并不规定服务器端的信息组织方式，除了支持书目信息的检索，理论上它可用于检索各种类型的数据资源，这就大大拓宽了用户检索的范围。另外，Z39.50的应用也可以在一定程度上解决因特网上信息的无序与难以检索的问题。借助于Z39 50，用户可以通过一个统一的接口程序同时检索多个本地或远程数据库，并对从多个数据库中获得的检索结果进行合并、去重和排序，用户只需要提交一次检索请求，就可以对多个资源服务器进行检索，而不需要逐个进入不同的检索服务界面，大大提高了检索效率。

元搜索引擎调用多个独立搜索引擎时，由于不同的搜索引擎所提供的用户界面风格、检索机制、检索语言等不同，元搜索引擎需要将用户的检索请求转换成独立搜索引擎识别

的检索语言，这样会大大降低元搜索引擎的检索效率，而且也会影响检索结果的数量和质量。

Z39.50在网络上是传输数据的高层协议，它为不同数据库提供了统一的接口标准。根据这个标准构成的检索系统，能够使用户对因特网上的异构系统进行检索，还可以在异构网络环境之间实现数据交换。

笔者提出了将Z39.50协议和元搜索引擎相结合的分布式的体系结构，以实现元搜索引擎的优化。我们在服务器端安装Z39.50网关，并且独立搜索引擎的索引数据库要具备Z39.50协议访问接口。具体来说，基于Z39.50的元搜索引擎的结构如图2所示。

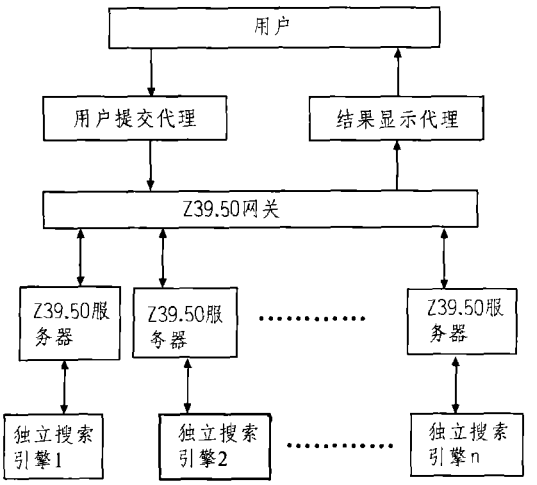

图2 基于Z39.50的元搜索引擎的结构

基于Z39.50的元搜索引擎的工作过程是：

（1）用户提出检索请求，元搜索引擎代理对检索请求进行预处理。

（2）元搜索引擎代理将用户的检索请求通过Z39.50网关转换为Z39.50支持的标准格式。

（3）Z39.50网关将转换后的检索请求同时递交给多个用户选定的Z39.50服务器。

（4）Z39.50服务器调用独立搜索引擎的索引数据库，提取符合检索要求的结果并返回检索结果。

（5）由Z39.50网关把检索结果处理后转化为HTML格式，并进行结果的处理。由元搜索引擎把检索结果呈现给用户。

3.3 基于Z39.50的元搜索引擎的评价

基于Z39.50的元搜索引擎，用户不直接和Z39.50服务器交互，而是通过一个应用服务器来访问Z39.50服务器。应用服务器由Web服务器和Z39.50网关组成。这种结构的优点是：

第一，运用Z390.50协议可以屏蔽各个独立搜索引擎的不同的检索语言、检索策略、文件格式和操作平台，它还规定了异构系统之间传递检索命令和数据的标准方法，以及自我编码和解码的统一机制。它通过屏蔽具体的转换过程，形成了统一、有效的操作范式。

第二，由于服务器端安装了Z39.50网关，元搜索引擎不需要将客户的检索请求一一转换成与独立搜索引擎检索型策略相适应的检索式。只要独立搜索引擎的索引库具备Z39.50协议访问接口，元搜索引擎就可以调用不同索引库的信息。这样可以大大提高系统的检索效率和检索质量。

第三，应用Z39.50协议的元搜索引擎可以对某一区域或者多个区域的Z39.50服务器进行检索，检索的内容不局限于文本文档，还可以是图像、声音和多媒体信息，这就会扩大用户的检索范围，提高用户的满意度。

第四，应用Z39.50协议可以提高系统的安全性。用户通过元搜索引擎进行检索必须要通过一个中间件才能完成对Z39.50服务器的检索，这显然要比普通方式安全得多。

Z39.50协议是一个比较复杂的协议，在应用服务器端需要解决实现Z39.50网关与Z39.50服务器的转化问题。

总之，元搜索引擎是继独立搜索引擎之后在信息检索方面的又一个研究热点。它以比较成熟的搜索引擎技术为基础，并进行了扩展和综合，对网络资源的发现和检索提供了更多机会。虽然元搜索引擎技术在许多方面还不是很成熟，但是随着新的信息检索技术的发展与应用，它必将成为网络信息检索的有效工具。

参考文献

1 徐宝文，张卫丰. 搜索引擎与信息获取技术. 北京：清华大学出版社，2003
2 郭少友. Web环境下分布式信息检索模式. 情报科学，2003(6)
3 丁峰，马范媛. 基于Z39.50的分布式WWW信息检索. 计算机工程，2001(2)
4 文坤梅，卢正鼎，邓曦，陈莉. 元搜索引擎中检索结果排序的优化方法. 华中科技大学学报，2003(3)

张 辉 山东大学管理学院副教授。通信地址：山东济南。邮编250100。

隋 佳 山东大学管理学院硕士研究生。通信地址同上。

（来稿时间：2005-07-28）