文章编号:1009-038X(2001)05-0526-05

Web文献库模糊优化查询的实现

李同英[1], 林 意[1], 殷开成[2]
(1.江南大学 信息工程学院, 江苏 无锡 214036;2.淮阴工学院, 江苏 淮安 223001)

摘 要: 由于目前大型的Internet结点日益数据库化,同时,传统的数据库系统向Internet平台转移,本文针对一种Web数据库访问模型,提出一种基于Web的数据库模糊优化查询法,以适应网上普通用户信息查询需求,能较有效地改善网上信息查询查准率低的状况.

关键词: Web数据库; 模糊匹配; 模糊优化查询
中图分类号: TP 311.134.1 **文献标识码**: A

Fuzzy Optimization Query for Web-based Database

LI Tong-Ying[1], LIN Yi[1], YIN Kai-cheng[2]
(1. School of Information and Control Engineering, Southern Yangtze University, Wuxi 214036, China; 2. Huaiyin Institute of Technology, Huaian 223001, China)

Abstract: In this paper, a Web-based database accessing model with hierarchy structure was introduced and a method of fuzzy optimization query was given.

Key words: Web-based database; fuzzy match; fuzzy optimization query

Internet是一个大型、自治式的分布式系统,网络速度的提高,使得大量声音、图象、文本等大型媒体信息能高效地在网络上传输,进而使Internet网络的使用范围越来越广.为了适应其可用信息和所用信息规模的爆炸式的增长,大型的Internet结点正日益发展为数据库系统,同时,传统的数据库系统也向Internet平台转移,浏览器/Web服务器/数据库服务器分布式结构逐渐形成.一方面Web从1991年出现经过短短几年已发展成为一个全球化信息空间,网上信息的飞速增长吸引了更多的网民,用户的范围进一步扩大,不再仅限于专业人员,在Web上进行信息查询已成为人们日常生活的一部分,上网成为人们获取信息的新方式;另一方面,网络信息查询的查准率是一个主要问题,面对网络信息查询需求的增长和目前存在的主要问题,提出构建Web信息数据库,将模糊优化算法植入查询中,提高信息查准率,提供给用户更高质量的信息服务.

1 Web数据库访问模型及其体系结构实现

NT上Web数据库的访问的方法大致有以下3种:

1) 公共网关接口CGI(Common Gateway Interface):这是传统的方式,但CGI技术有很多缺点,如不易开发,更改成本高,功能有限,不易调试和检错,不具备事务处理的功能且很耗费服务器资源.

2) Internet数据库连接器IDC(Internet

收稿日期:2001-05-17; **修订日期**:2001-09-08.
作者简介:李同英(1969-),女,江苏淮安人,计算机应用专业硕士研究生,助理工程师.

Database Connector)：IDC 是集成在 Internet Server API(ISAPI)的应用.但是，由于 IDC 技术在同一时刻，只有一个实例在运行，要求能运行在安全的多进程，多个请求同时到达，每一个函数在争用同一文件或同一数据块的内容时，必须多加小心，而涉及多进程的代码是很困难的.目前，ISAPI 还不具备跨平台的功能，只限于 NT 平台.

3) 先进数据库连接器 ADC(Advanced Database Connector)：ADC 提供一个数据处理“Advanced Database Control”的 ActiveX Control，以访问 ODBC 的数据库.ADC 与以上两种方案最大的不同点在于：ADC 的数据查询操作是在用户端的浏览器上执行的.但 ADC 要将服务器端数据库中的可高达数千笔的记录，先下载到用户端，所以只适合一些特别频繁的数据库查询操作.

鉴于上述 Internet 上访问数据库几种方案的利弊，拟通过 ActiveX 数据对象(ActiveX Data Object，缩写为 ADO)访问，与 ASP 配合使用，采用完整的站点数据库访问的解决方案. Microsoft 公司 1997 年推出的功能强大的 Web 脚本编写开发工具 ASP (Active Server Page)，嵌入 VBScript 及 JavaScript 语言脚本，ASP 在站点的 Web 服务器上解释此脚本，分布执行，可产生动态、交互、高效的站点服务器应用程序.当程序在服务器而不是在客户端执行时，Web 服务器将完成产生浏览器的 HTML(Hypertext Markup Language)的所有工作.

由于 ASP 在服务器上运行，所以 ASP 的源程序代码不会传到用户的浏览器，可保护源程序不会外漏.此外，ASP 也是面向对象的，而且还可自己制作 ActiveX 服务器组件来扩充功能，可使用 Visual Basic，Java，Visual C++，COBOL 等程序语言来实现.

ASP 可以让开发者在 VBScript 或 JavaScript 中编写代码并允许把 ActiveX 控件直接集成到 Web 服务器内部.而 ActiveX 控件可用 VB/VC 等编写，使用户能自由选择最方便的方法来编写应用程序.由于 ASP 的 ActiveX 是在服务器端，客户端可使用任意的浏览器，非常灵活.

服务器端操作系统使用 Windows NT，以其 Internet Information Server(IIS)构建 Web 服务器平台，这样就构成了 Internet/Intranet 数据库应用系统比较理想的构架，这种构架的特点是程序、数据库及组件集中于服务器端，而客户端只要有 IE 或 Netscape 较新版本的浏览器即可.在此构架上，利用 ADO 组件的数据库连接功能可轻松实现对数据库的存取.

ADO 是一个 ASP 内置的 ActiveX 服务器组件(ActiveX Server Component).ADO 通过在 Web 服务器上设定 ODBC，可建立与多种数据库，如：SQL Server，Oracle，Informix，Access，VFP 等的连接.可以把它与 ASP 结合起来，建立提供数据库信息的网页内容，在网页画面执行 SQL 命令，用户在浏览器页面中输入/更新或删除 Web 服务器信息，动态地请求信息，由服务器对 Web 数据库进行相应的操作.

综上，ASP 的基本工作原理是，当用户通过 ASP 文件访问数据库时，WebServer 如果响应该 HTTP 请求，则调用 ASP 引擎，通过 ADO 组件与 ODBC 数据源连接，当数据库服务器的得到请求后，对该请求进行合法性验证，然后执行 SQL 指令，并将执行结果动态地生成一个标准的 HTML 页面返回给 Web 服务器，以响应用户的请求(如图 1 所示).

上述数据模型，其体系结构为典型的使用系统组织的层次体系结构模型如图 2 所示：

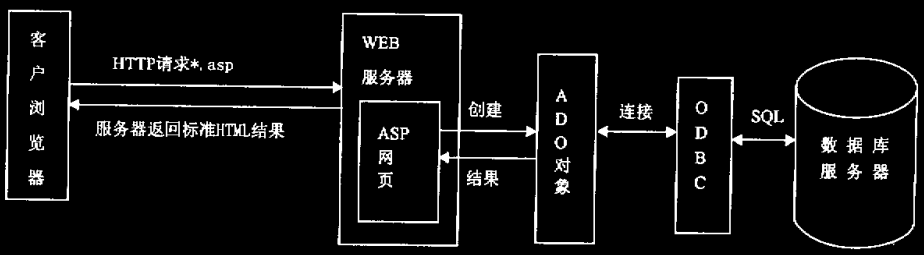

图 1　ASP 工作原理图

Fig.1　Working principle of Active Server Page

图 2 体系结构模型

Fig.2 The model of the structural system

2 Web 数据库的模糊优化查询

构建了上述 Web 数据库后，将模糊优化算法植入数据库查询中.

2.1 非一致性模糊匹配算法

一致性匹配即用户的查询提问与数据库记录的严格匹配，当用户提问与数据库表中记录一致时，即数据库表中有符合用户查询的记录，结果记录返回给用户；当用户查询提问与数据库记录不一致时，用户只能得到"无此记录"的结果信息. 非一致性模糊匹配查询弥补一致性匹配的不足，在数据库中没有与用户严格匹配的记录时，按数据库表中记录与用户提问的接近程度，为用户提供一个或一组最接近用户查询要求的数据库表中记录.

不妨设数据库表中有 n 个记录 $D=\{Y_1, Y_2, \cdots, Y_n\}$，每个记录有 m 个描述项描述其特性 $P=\{P_1, P_2, \cdots, P_m\}$，$D$ 到特征集 P 的模糊关系可用矩阵 $\mathbf{R}=(r_{ij})_{n\times m}$ 表示，而 r_{ij} 可由模糊统计得到.

用户根据特征集 P 提出查询要求. 设用户的查询词字记为 $A=\bigcup_{k=1}^{L} A_k$，其中 A_k（第 k 个查询分句）是特征集 P 上的模糊集合，A_k 对于 P 的隶属函数 $A_k=\{V_{1k}, V_{2k}, \cdots, V_{mk}\}$，$k=1,2,\cdots,L$，此 V_{jk} 表示 A_k 对特征属性 P_j 的占有程度. 确定优选阀值增量 Q，$0.5<Q<1$.

求查询解向量. 设矩阵 $S=(V_{jk})$ 为 $m\times 1$ 阶查询矩阵，称

$$\beta_{ik}=\frac{\sum_{j=1}^{m}\min(r_{ij}, V_{jk})}{\sum_{j=1}^{m}\max(r_{ij}, V_{jk})}$$

为数据库记录 Y_i 对查询提问的相近度，亦是 Y_i 与 A_k 按定义所给的贴近度，表示记录 Y_i 对查询提问 A_k 的符合程度.

令 $M=\max \beta_{ik}$，$m=\min \beta_{ik}$

其中 $1\leqslant i\leqslant n$，$1\leqslant k\leqslant L$

记 $\delta=m+0.618(M-m)$ 为优选阀值，

$\eta=m+0.618(M-m)Q$ 为扩选阀值，

变换 β_{ik} 得：

$$Q_{ik}=\frac{\beta_{ik}-m}{M-m}\cdot\frac{\sqrt{5}+1}{2}$$

$0\leqslant Q_{ik}\leqslant(\sqrt{5}+1)/2$，分为 4 种情况：

1) $1\leqslant Q_{ik}\leqslant(\sqrt{5}+1)/2$，即 $\delta\leqslant\beta_{i\,k}\leqslant M$，$Y_i$ 为 A_k 的优解；$Q_{ik}=1$ 时，Y_i 为 A_k 的极优解，记 $t_{ik}=(Y_i, Q_{i\,k})=Y_i\,Q_{ik}$

2) $Q\leqslant Q_{ik}\leqslant 1$，亦即 $\eta\leqslant\beta_{i\,k}\leqslant\delta$，$Y_i$ 为 A_k 的扩解，$t_{ik}=(Y_i, Q_{ik})=Q_{ik}Y_i$

3) $0.5\leqslant Q_{ik}\leqslant Q$，

即 $m+(\sqrt{5}-1)(M-m)/4\leqslant\beta_{ik}\leqslant\eta$，$Q_{ik}$ 为 A_k 的候解，$t_{ik}=Q_{ik}Y_i$

4) $0\leqslant Q_{ik}\leqslant Q$，

亦即 $m\leqslant\beta_{i\,k}\leqslant m+(\sqrt{5}-1)(M-m)/4$，记 $t_{ik}=(Y_i, Q_{ik})=\theta$

优解和扩解均为查询解，而候解作为参考备用的候补解，$T=(t_{ik})$ 为查询预解矩阵，进一步可得查询解矩阵，记 $T^*=(t_{ik}{}^*)$.

返回解向量给用户. 据 T^* 将其各列相加（θ 算作 0），设 $t_{ik}{}^*$ 对应的数为 μ_{ik}，令

$$W_i=\sum_{k=1}^{L}\mu_{ik}\qquad i=1,2,\cdots,n$$

取 $\zeta=2Q\times 0.618$，则 A 对 D 的查询向量为：

$$\left\{Y_1\frac{W_1}{\zeta}, Y_2\frac{W_2}{\zeta}, \cdots, Y_n\frac{W_n}{\zeta}\right\}$$

按 W_k 的大小顺序重排，得 A 对 D 的有序查询向量：

$$\left\{Y_{r1}\frac{W_{r1}}{\zeta}, Y_{r2}\frac{W_{r2}}{\zeta}, \cdots, Y_{rn}\frac{W_{rn}}{\zeta}\right\}$$

其中 $W_{r1}\geqslant W_{r2}\geqslant\cdots\geqslant W_{rn}$，此查询解不仅限于传统的一致性查询解，它提供给用户有灵活使用余地的最优和相近查询解，与一致性匹配查询相比有更高的查准率，同时又兼顾了查全率.

2.2 模糊分类

我们把上述非一致性模糊匹配算法用在无锡轻工大学（现江南大学）学报管理系统中的稿件数据库查询中，发现这样的中小型数据库有很高的查准率、查全率和查询速度，效果很好，但当此用法用到无锡轻工大学图书馆馆藏数据库或其它大型文献数据库的书刊目录查询时，与其它查询算法一样会遇到查询速度问题. 为了解决此问题，可先把所有文献模糊分类，将大型文献库分成几个中小型类子库，再查询.

设文献库 F，有 $Y_1, Y_2, \cdots, Y_n$ 这 n 个文献组成，每个文献有 m 个特征 $P=\{P_1, P_2, \cdots, P_m\}$，如文献名称、作者、关键词等，由 m 个描述项 $d_1, d_2,$

$\cdots, d_m$ 表示,因此每一个文献都可用一个 m 维向量 $Y_k=(\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{km})$ 表征.

设要将 F 分为 k 类:$F_1, F_2, \cdots, F_k$,每一类 F_i 可看成论域 $F=(Y_1, Y_2, \cdots, Y_n)$ 上的一个模糊子集,由隶属函数 $\rho_i(Y)$ 来表征,假定描述项 d_j 在模糊子类 F_i 中出现的概率为 P_{ij},则得矩阵:

$$a=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ P_{k1} & P_{k2} & \cdots & P_{km} \end{bmatrix}$$

其中行表示类,列对应描述项,文献 Y_h 对于类 F_i 隶属函数为:

$$\rho_i(Y_h)=\frac{\sum_{j=1}^{m}\alpha_{hj}P_{ij}}{\sum_{j=1}^{m}\alpha_{hj}}$$

其中 $i=1,2,\cdots,k$, $h=1,2,\cdots,n$,则 $Y_1, Y_2, \cdots, Y_n$ 隶属于 F_i 的函数矩阵为

$$b=\begin{bmatrix} \rho_1(Y_1) & \rho_2(Y_1) & \cdots & \rho_k(Y_1) \\ \rho_1(Y_2) & \rho_2(Y_2) & \cdots & \rho_k(Y_2) \\ \cdots & \cdots & \cdots & \cdots \\ \rho_1(Y_n) & \rho_2(Y_n) & \cdots & \rho_k(Y_n) \end{bmatrix}_{n\times k}$$

如果文献 Y 隶属于类 $F_i \cap F_j$ 隶属函数为 $\rho_{ij}(Y)$,则有 $n\times k(k-1)/2$ 矩阵:

$$c=\begin{bmatrix} \rho_{12}(Y_1) & \rho_{13}(Y_1) & \cdots & \rho_{k-1k}(Y_1) \\ \rho_{12}(Y_2) & \rho_{13}(Y_2) & \cdots & \rho_{k-1k}(Y_2) \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{12}(Y_n) & \rho_{13}(Y_n) & \cdots & \rho_{k-1}k(Y_n) \end{bmatrix}$$

其中 $\rho_{ij}(Y_h)=\min\{\rho_i(Y_h), \rho_j(Y_h)\}$, $i=1,2,\cdots,k-1$, $h=1,2,\cdots,n$, $j=2,3\cdots,k$.

借助于矩阵 c 可确定模糊分类的阀值 E,E 介于 0 和 1 之间.E 取得越大,分类精度越高,一个文献属于多个子类的可能性越小;反之,E 取得越小,则分类越粗糙,一个文献同时属于多个子类的可能性就越大,所以 E 要选择适中,$E<\min\{\max\rho_{ij}(Y_h)\}$,使得每个文献至少应分到一类中,则得 λ 截集:

$$F_i=\{Y\mid\rho_i(Y)\geqslant E\}(i=1,2,\cdots,k)$$

于是即可将文献库 $F=\{Y_1, Y_2, \cdots, Y_n\}$ 的一个分类 $F_1, F_2, \cdots, F_k$,将对 $F=\{Y_1, Y_2, \cdots, Y_n\}$ 的匹配转化为对 $F_i(i=1,2,\cdots,k)$ 的匹配,匹配范围减小,查询效率随之显著提高.

模糊分类的手段有完全人工、机器自动和人工与计算机相结合 3 种方式.机器自动方式通过计算机软件完成,此软件有初步的切词、分词和语义理解功能,自动地对文献进行特征标识提取,再据此特征标识利用上述模糊分类算法由计算机自动完成文献的分类.

2.3　模糊分类和模糊匹配相结合的模糊优化查询

利用上述模糊分类方法将大型文献库自动分成数个类库后,用户输入所查文献的描述项,计算机据此用模糊分类法判断所查文献隶属的类库,再在隶属库中模糊匹配,返回所需文献,这样既提高了查询速度和查准率,又充分考虑了查全率,同时由于采用模糊分类和模糊匹配,而不是一致性匹配,不要用户选择类库,对用户查询输入要求降低,因此与一致性精确查询相比更加友好,真正面向一般用户.

3　实验

学报稿件的篇名一般很少两三个字,有的篇名达到二十个字左右,就是著者本人时间长了也很难一字不漏按原顺序输入查询,而且术语很多,对一般用户很难想象能精确输入篇名,查看得到自己需要的文章.我们将上述模糊算法初步植入篇名查询中,给予满意的解决.

用户只要在输入框将篇名有关的词字填入(如图 3),与过去文献检索相比,不用指明各词字间的逻辑关系(and, or, not),按下提交按钮即可得到最接近所输入查询要求的稿件如图 4 所示.

若该稿件为用户所需,点击浏览即超链接得到稿件全文,可浏览和下载;若用户不满意该稿件文章,即非用户所需,按下一篇得到与查询要求次相近的稿件,或回退修改查询输入或重新输入查询要求.

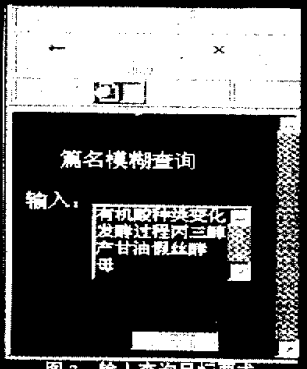

图 3　输入查询目标要求

Fig.3　The input interface

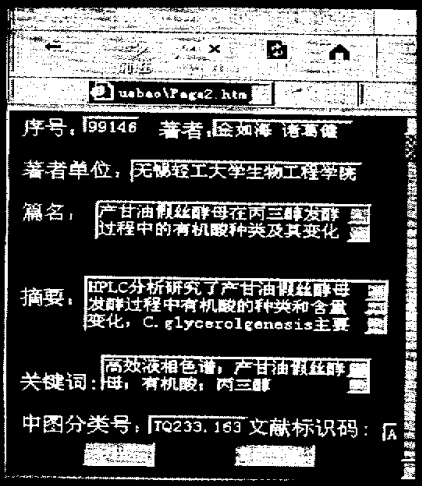

图4 查询结果

Fig. 4 The result query

由于模糊算法的植入，匹配过程是非一致性模糊匹配，不是一致性精确匹配，按照人们的思维过程，用户可边想边输入，模棱两可的字词都输入进去(如醇酯、有机酸无机酸、标点、符号、空格、字母等)，在不淹没查询目标要求的情况下，均可帮用户查询到与查询目标最接近的或相近的文章.

从上述实验结果看模糊优化查询不仅降低了用户输入要求，对用户更加友好，一般 Web 用户都可得到满意结果，而且有一定的抗干扰噪声能力以及一定的容错性.

4 结 语

当今是信息爆炸的时代，Web 为人们提供了一个全球化信息空间，针对目前大型的 Internet 结点日益数据库化，和传统的数据库向 Internet 平移，提出基于 Web 的数据库模糊优化查询法适应网上一般用户信息查询需求，改善网上信息查询查准率低的状况，此算法可用于网站搜索引擎中，另外此法因其一定的抗干扰噪声能力和容错性还可用于基于 Internet 的多通道用户界面如目标选择等方面.

参考文献：

[1] LAWRENCE S, GILES C L. Accessibility and distribution of information on the Web[J]. Nature, 1999, 400:107～109.
[2] 刘增良主编. 模糊技术与应用选编[M]. 北京：航空航天大学出版社，1997.
[3] 肖位枢编著. 模糊数学基础及应用[M]. 北京：航空工业出版社，1992.
[4] 董士海等. 基于 Internet 的多通道用户界面[J]. 计算机学报，2000(12)：1270～1275.
[5] 张德. Internet 上的数据库联合查询优化[J]. 计算机学报，2000(2)：171～176.

(责任编辑：朱 明)

(上接第525页)

类的 get Attribute(id)方法获得标记“〈产品〉”的“id”属性值，程序把该“id”同从 DBMS 中检索到的零件编号“Num”相比较，若相等，则进一步对产品子树进行遍历. 或通过 java.ms.xml.om.Element 类的 getText()方法返回“〈产品〉”标记的所包含的全部非标记信息，即所查找的汽车零件的详细信息.

4 小 结

传统的数据库技术只有与飞速发展的网络技术相适应，才能得以进一步的发展和应用. XML 技术特别适合于半结构化数据资源的组织和发布. XML 中的文档类型定义 DTD，使人们可以根据实际需要构造所需的标记，不仅方便用户开发基于 Internet 的应用系统，而且结合 Java 技术，可方便地构建一个与用户平台无关的统一而简单的交流方式. 本系统基于 JSP 技术，把 XML 技术和 DBMS 技术结合起来，充分发挥传统关系型数据库在信息检索方面的优越性，实现了信息的描述、信息的内容和结构与系统的查询真正地分开. 提高了整个系统的网上查询的效率.

参考文献：

[1] CHRISTOPHER A. Linux Web 编程[M]. 北京：电子工业出版社，1999.
[2] HORSTMAN S. Java 2 核心技术[M]. 北京：机械工业出版社，1999.
[3] 黄理. 用 JSP 轻松开发 Web 网站[M]. 北京：希望电子出版社，2000.
[4] HAROLD Rusty 著. XML 实用大全[M]. 北京：中国水利电力出版社，2000.
[5] 怀石工作室编著. XML 完全手册[M]. 北京：中国水利出版社，2000.
[6] 武苍林著. JDBC 在数据库中的应用[J]. 计算机应用，1998，18(10)：35～36.

(责任编辑：朱 明)