

# 化学网络数据库在 WEB 中的实现

周翠松 李梦龙\* 李工兵 刘冀昆

(四川大学化学学院 成都 610064)

**摘 要** 对化学网络数据库的结构和工作原理进行了介绍, 对如何建立全面完整的化学网络数据库做了深入总结。以 MySQL 作为数据库管理系统, 选用中药数据(天然药物)为数据源, 采用 PERL 为 CGI 开发工具, 利用化学计量学算法和三维结构搜索开发了化学软件包, 建立了 WEB 中的化学网络数据库。

**关键词** 网络数据库 化学计量学 软件包 CGI 三维结构搜索

## The Realization of Chemical Web Database in Web

Zhou Cuisong, Li Menglong\*, Li Gongbing, Liu Jikun

(Department of Chemistry, Sichuan University, Chengdu 610064)

**Abstract** The structure and the principle of chemical web database were introduced and a deep summary on how to establish a comprehensive and complete chemical web database was also made in this paper. During the establishment of chemical web database, we chose MySQL as our DBMS and Chinese traditional medicine (natural medicament) as database resources, used PERL to develop the CGI and produced software packages by 3D structure research and chemometrics arithmetic.

**Key words** Web database, Chemometrics, Software package, CGI, 3D structure search

随着化学领域和 Internet 的飞快发展, 化学信息资源正以指数级的速度增长, 因此如何更好地表示、存贮、检索、传递、挖掘和利用这些信息资源已成为我们不可回避的问题。化学的信息表现为两方面<sup>[1]</sup>: 文献资料数据属于“软数据”, 从实验中取得的数据为“硬数据”; 硬数据有两种: 一种是与化学结构有关的数据如各种谱图、X 射线衍射等, 一种是非结构的化学物理参数如热化学、化学热力学、热质传递及相平衡等。化学数据库系统功能强大, 可提供各种数据<sup>[2]</sup>, 如结构式、反应式、分子式、分子量、谱图、参考文献等, 甚至可以与其它软件相连分析数据<sup>[3]</sup>。同时, 基于 WWW 的化学网络数据库系统具有用户界面统一、使用方便、费用少等优点, 它可以为 Internet/Intranet 用户提供高效、远程服务, 升级和维护很方便, 还可实现客户端零代码编程, 因此它必将取代传统的联机 and 单机版本的数据库系统。

### 1 化学网络数据库的系统结构

图 1 显示了化学网络数据库的工作流程。用户使用安装在个人计算机上的浏览器如 Internet

周翠松 女, 29 岁, 博士生, 从事化学信息学方面研究。\*联系人 E-mail: liml@scu.edu.cn

国家教育部骨干教师基金资助项目

2003-01-08 收稿

Explorer 或者是 Netscape 通过因特网登录到化学数据库所在的 WWW 服务器上, 利用服务器上的 HTML 表单文件发出数据库操作请求, 服务器接收到合法请求后调用 CGI 程序向 RDBMS 发送相应的数据库操作命令, 由 RDBMS 对化学数据库中的数据信息进行一系列操作, 并将结果回传给 CGI 程序。CGI 程序按设定好的算法对获得的信息作进一步的处理, 最后将结果交给 WWW 服务器通过因特网传送给用户的计算机, 在浏览器上显示出来。图 1 中的网络数据库是一种典型的服务器/客户机结构的数据库, 其中安装有 CGI 程序、RDBMS 软件和化学数据库的服务器 (网络数据库系统中的服务器), 而安装有浏览器的用户计算机是系统的客户机。化学网络数据库建立的关键在于化学数据的收集、组织方式设计和 CGI 程序设计这两个方面。

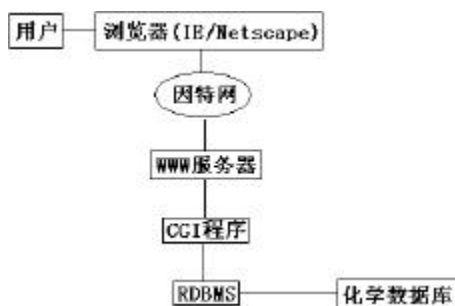


图 1 WWW 方式的化学网络数据库结构示意图

Fig.1 Sketch map of structure of chemical web databases

## 2 数据库平台构建

数据库管理系统(DBMS)是实际存储数据和系统用户之间的一个支持系统, 用户存取数据库的所有请求都要经过这个系统。目前世界上建立数据库的主流是关系型数据库管理系统(RDBMS), 有 Access、SQL server、Oracle、Sybase、DB2、MySQL 等, 大多建立在 Windows、UNIX 或 Linux 等操作平台上。从综合功能、安全性、稳定性、性价比以及易用性等方面的考虑, 笔者选用 MySQL<sup>[4]</sup>作为数据库管理系统。MySQL 可以直接从网上免费下载, 功能强大、稳定性好, 可以在多种操作平台上移植; 它支持标准的 T-SQL 和超过 1TB 的海量数据, 适合建立化学数据库; 还提供了应用程序接口, 可以方便其它应用程序开发工具通过调用这些接口对 MySQL 中的数据进行提取、分析。

## 3 化学数据库的数据源的收集及其组织方法<sup>[5]</sup>

数据源选用中药(天然药物)数据: 常用中药成分与药理手册。

字段名或可建成子数据库的字段: CAS 登录号, 英文命名, 英文别名, 中文命名, 中文别名, 主要成分分子式, 分子量, 化学结构式, 二维和三维药物分子结构, 药物分子中的药效团, 药物的物理性质, 药物的药理活性, 临床应用, 参考文献, 专利文献, 质谱、红外、紫外、核磁谱等谱图及其数据。

其中化学结构式采用二维图形和三维图形显示, 三维图形采用 VRML 虚拟现实构造语言实现。对文献和谱图及其数据单独构建数据库。并利用神经网络、小波压缩算法等先进手段对谱

图进行处理<sup>[6~9]</sup>。

## 4 化学数据库因特网接口

### 4.1 CGI 开发工具

当前流行的 CGI 开发语言和工具有 VB、Delphi、Visual C++、C、PERL、Java 和 ASP 等, 综合考虑易用性、反应速度后采用 PERL 语言编写 CGI 程序, 编出的 CGI 程序可以不作修改, 直接应用于 Windows 和 LINUX 操作系统上。但是由于 PERL 是一种解释型的语言, 不能将比较复杂的算法结合于其中, 否则整个因特网数据库的运行速度会因为这个瓶颈的影响而变得很慢, 因此对一些数据库操作、数据处理、化学理论等采取了适当的优化算法<sup>[6~9]</sup>进行描述后编成处理模块, 再结合进 CGI。

### 4.2 CGI 程序的内部结构

整个处理流程如图 2 所示。CGI(Common Gateway Interface)公共网关接口在因特网中是另一个重要组成部分, 是用户通过因特网与数据库进行交流的门户。根据数据库具有数据与应用程序相互独立的特点, 在 CGI 编程中结合化学计量学方法对从数据库中获得的数据进行分析, 得出有意义的结论并返回给用户。如果用户不需要进行处理, 此时的 CGI 程序相当于一个检索工具。

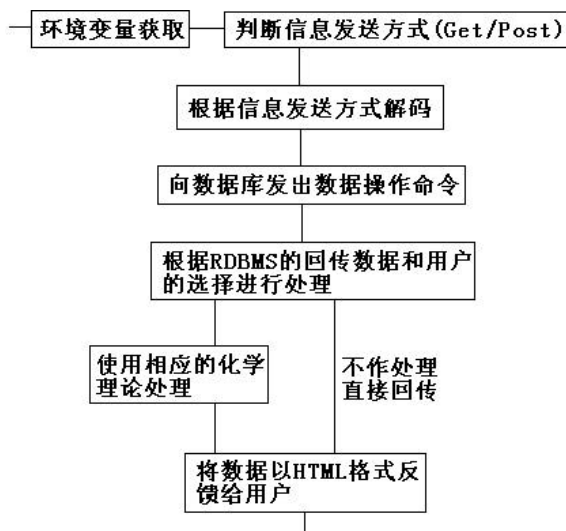


图 2 CGI 程序的主要结构示意图

Fig.2 the main sketch map of structure of CGI program

### 4.3 三维结构搜索<sup>[10]</sup>

药物设计和药物定量构效关系(QSAR)是近年来世界上的一个研究热点, 利用数据库中的药物数据进行药物分子设计研究可以更为快捷有效地发现具有特定药效的新药。其中, 三维结构搜索<sup>[10]</sup>是一种开发三维结构数据库、发现先导化合物的快捷有效的方法, 大大节约了研究的前期工作量。三维结构搜索是一个连续的多步骤过程, 包括初筛、几何查找、柔性构象搜索三个步骤。

初筛是指除去根本不可能存在和无生物活性的分子, 减少查询分子数、提高查找效率和缩

短查找时间。几何查找是指经过初筛的分子经过严格的二维结构查找,以确定原子之间的连接方式是否与查询者需要的目标分子结构特征相匹配。

构象搜索方法是三维构象搜索的主流方法,实际上是一种优化方法,命中率较高。柔性构象搜索的提出是为了避免刚性构象搜索时漏掉一些符合用户要求的三维构象;允许刚性搜索失败的分子构象按某种方法变形,再加入搜索序列中,从而获得更多的命中结构;必要时还加入范德华能量限制,以避免在柔性构象搜索过程中产生能量不合理的命中构象。周家驹等<sup>[10]</sup>报道柔性构象搜索方法有三类:穷举法、使用柔性提问结构、在查找过程中进行构象搜索。中国科学院化工冶金研究所的王亭和中国科学院计算机化学开放实验室的周家驹在他们的 3DFS 软件中采用了 POWELL 法,该优化法的特点是高效、无需求导数。

三维结构搜索算法采用遗传算法和蒙特卡罗模拟退火算法,并将 QSAR 方法、化学计量学方法、分子力学和分子动力学等方法模块化或插件化,通过软件模块接口集成于客户端软件包供 CGI 调用。但由于 CGI 是位于服务器上的一种应用程序,需要耗费服务器的资源,因此系统资源很可能因为用户的访问频率较大而被耗尽导致服务器锁死,即使运用技术水平强大的 ISAPI 程序<sup>[11]</sup>都有可能遇到这种情况。为减轻服务器的负担,笔者酌情增加或删除了 CGI 程序中的初筛和几何查找处理模块,而把它们和三维结构搜索模块集中在一起捆绑于一个客户端软件包中,挂在数据库的页面上供广大研究人员免费下载使用。

## 5 讨论

除化学网络数据库外,笔者还构建了西部植物标本网络数据库、西部天然药物网络数据库和常用化学数据资源库,并对西部天然药物网络数据库中的部分药物活性分子的化学资源展开了一系列深入研究工作。随着科技的日新月异,化学网络数据库的应用不仅提供对化学数据库进行查询和检索,还帮助用户从化学数据中提取带有结论性的有用信息,满足科研、生产和商业更多的需要,对我国化学领域的研究与应用具有很大的意义。

## 参考文献

- [1] 张俊逸. 化学通报, 1990, (12): 1~7.
- [2] 陈维明, 王 源, 朱谷 等. 计算机与应用化学, 1999, 16: 355~356.
- [3] 王奎旗, 赵 文, 闫作伟. 山东化工, 2000, 29: 13~15.
- [4] <http://www.mysql.com>
- [5] 王 源, 陈维明, 朱翠娣. 科学数据库与信息技术论文集. 北京: 科学出版社, 1998: 63~67.
- [6] 李梦龙. Internet 与化学信息导论. 北京: 化学工业出版社, 2001: 1~208.
- [7] 李梦龙, 康 彬, 戚华溢 等. 高等学校化学学报, 2002, 23: 1281~1284.
- [8] 戚华溢, 李梦龙, 康 彬 等. 四川大学学报, 2002, 39(2): 374~377.
- [9] 李梦龙, 康 彬, 戚华溢 等. 化学通报, 2002, (65): 71~73.
- [10] 王 亭, 周家驹. 化学进展, 1998, 10: 442~450.
- [11] 何 敏, 周家驹. 计算机与应用化学, 1999, 16: 280~282.