个人搜索引擎 GONIA-OFFICE 98 的设计与实现

魏星 丁伟 高毓航[1]

（东南大学 计算机系，210096 南京）

【摘要】本文介绍了一个面向 Win98 平台的个人 Office 文档和 HTML 文档用搜索引擎系统的设计与实现，该系统基于原有面向 NT 平台的内部网搜索引擎 GONIA-OFFICE NT 实现，通过加入“用户标识”实现有关的访问控制，以保证用户个人信息的安全。系统的所有功能均基于 WEB 实现。

【关键词】信息定位；个人搜索引擎；OFFICE；访问控制

中图分类号：TP393

Designation and Realization of an Individual Search Engine GONIA-OFFICE 98

Wei Xing, Ding Wei, Gao Yuhang

(Southeast University, Computer Science Dept., 210096 Nanjing, P.R.China)

【Abstract】 In this paper, a personal search engine GONIA-OFFICE 98 is introduced, which works with individual OFFICE and HTML documents on Win98 platform by improving an existing Intranet search engine system's code on MS-NT named GONIA-OFFICE NT. 'User identifier' is used for access control, which is the major difference GONIA-OFFICE 98 from GONIA OFFICE NT and the key way to keep user's personal information secured. All the implementation work is based on WEB for convienent use.

【Key words】 Information locating, individual search engine, OFFICE, access control

GONIA 是一个由 CERNET 华东（北）地区网络中心开发的具有分布式体系结构的通用中英文搜索引擎。它通过计算文档之间相关度（RSV）值的搜索引擎内核，采用低语义依赖强度的中文特征（features）产生方式和索引建立方式，支持自然语言查询和用户反馈功能。Gonia 通过正规化的数据收集处理接口实现数据信息源的编码特征屏蔽，并以此实现通用性，使其能面向所有数字化的信息资源。该系统内核运行于 UNIX 平台上。

为了使该系统能在跨平台的内部网 WINDOWS－NT 文件服务器上的 MS－Office 系列文档和超文本

[1] **作者简介**：魏星，硕士研究生，主要研究方向为信息发现技术。

丁伟，工学博士，东南大学计算机系副教授、硕导，主要研究方向包括网络管理、网络安全、网络体系结构、开放分布式处理等。

高毓航，硕士研究生，主要研究方向为网络安全。

定稿日期：2000-08-01

文档中使用，实现对该类文件的定位，我们没有采用将搜索引擎内核向 NT 平台移植的方法，而是仅在在 NT 端实现了一个收集代理程序，它的主要功能是将指定目录下的有关文档滤去控制和格式信息，转换成 Gonia 内核能够处理的标准化格式，再将其发往安装了 Gonia 内核的 UNIX 平台。以此构成的系统称为 Gonia-officeNT，它是分布式通用搜索引擎 Gonia 中的一个子系统。采用这种方法的优点在于开发工作量小，便于管理。更详细的介绍参见 CERNET 99 论文集中《内部网搜索引擎 GONIA－OFFICE 的设计与实现》一文。

同样的实现思路转向 Windows98 平台，可以实现该平台上个人文档的定位，这个系统称为 Gonia-office98，它与 Gonia-officeNT 一道，构成 Gonia-Office。Gonia-officeNT 系统代码的主要部分均可在 Gonia-office98 中使用，但随之而来的问题主要集中在信息的安全方面，此时的用户最重要的要求是要保证搜索引擎运行不能以任何方式泄露文档的内容。由于 Gonia 内核采用了不以任何形式保存文档原文的工作机制，因此 Gonia-office98 需要解决的问题主要在访问控制方面。下面本文将详细介绍 Gonia-office98 的设计和实现方法。

1. 系统设计

1.1 访问控制政策

如上所述，局域网环境下的个人搜索 Gonia-office98 实现的关键在于访问控制。访问控制的实质是对资源使用的限制。在计算机系统中设立安全机制的最初目的就是为了控制用户对系统资源的访问。一般来讲，访问控制政策有三种：

（1）自主访问控制政策

自主访问控制的含义是由客体来自主地确定各个主体对它的直接访问权限（又称访问模式）。这种方法能够控制主体对客体的直接访问，但不能控制主体对客体的间接访问（利用访问的传递性，即 A 可访问 B，B 可访问 C，于是 A 可访问 C）。

（2）强制访问控制政策

由一个授权机构为主体和客体分别定义固定的访问属性，且这些访问权限不能通过用户来修改。例如将数据分成绝密、机密、秘密和一般等几类。用户的访问权限也类似定义，即拥有相应权限的用户可以访问对应安全级别的数据，从而避免了自主访问控制方法中出现的访问传递问题。

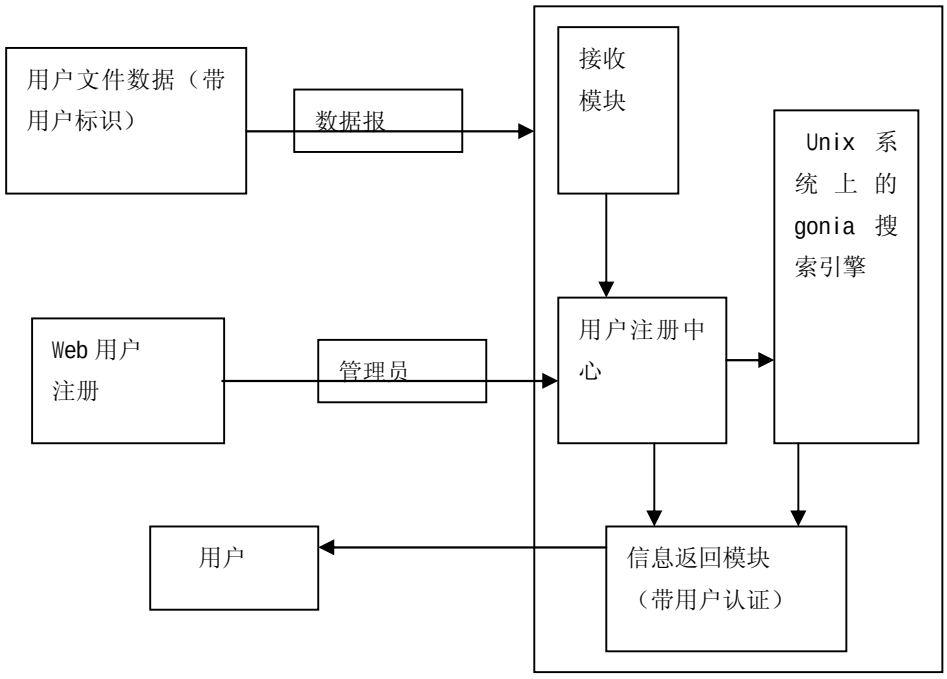
（3）基于角色的访问控制政策

这种方法是对自主控制和强制控制机制的改进，它基于用户在系统中所起的作用来规定其访问权限，这个作用（即角色 role）可被定义为与一个特定活动相关联的一组动作和责任。例如担任系统管理员的用户便有维护系统文件的责任和权限，而并不管这个用户是谁。

本系统采用的访问控制政策是基于强制访问和角色访问的结合。首先，用户的权限不能自己修改，更不能由其他用户修改，而是由管理员统一定义的，每个用户可在自己权限范围内进行查询；其次，对用户的权限分配基于角色。强制访问控制机制和基于角色的访问控制机制可有多种实现方法，本系统采用了适用于一个集中式方法：由一个安全管理员负责。

访问控制的实现可通过访问控制表（Access Control List - ACL）、容量控制（Capabilities）和授权关系（ Authorization relations）。本系统采用了容量控制，这是一种稀疏矩阵表示法，以主体为索引。每个主体对应有一个 CL（Capabilities List），指出对各个客体的访问权限。这种方法便于主体，不利于客体访问权限的维护，而本系统中客体的访问权限是确定的。

结合访问控制，系统总体结构如下图：



图一 系统总体结构图

1.2 公用数据和私有数据

计算机系统中所有可控制的资源均可抽象为客体（object）；对客体实施动作的实体称为主体（subject）；主体对客体所实施的动作需要得到授权；这些授权对于主体可表为访问权限，对于客体可表为访问模式。

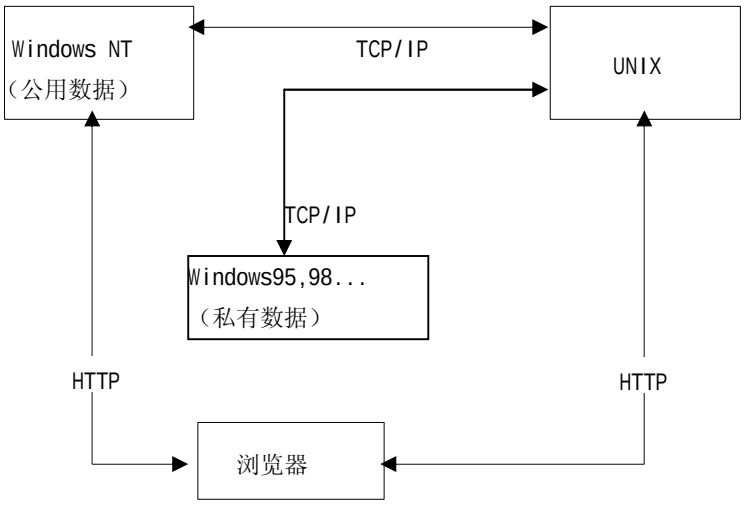
本系统中客体即为所有数据，从总体上概括地划分，并以数据所有者的角度来区别，数据可简单地分为两种：公用数据和私有数据。公用数据是指有两个或两个以上的用户需要并有权访问的数据，一般都要求在 Intranet 上任何一点都能够访问到它；私有数据则相反，它的所有权属于某个特定的用户，并且只允许数据的所有者对它进行操作，操作点在某台特定的计算机（一般是数据所有者个人使用的 PC）上则更为合适。显然，只有公有数据是需要并允许共享的。

由于以上两种数据的存在，并且由于它们在权限和需求上的差异性，在 Intranet 中，一般都要为两种数据分配两个不同的空间，即公用数据空间和私有数据空间。公用数据空间可从服务器的硬盘空间中分出一块，在网上共享出去，允许用户共享；私有数据空间则建立在各用户的个人 PC 上。如尚未公开的论文、报告等等都是私有数据。私有数据空间由各用户个人管理。Gonia-office98 要实现的访问控制实际上是系统运行过程中对私有数据空间中的数据在离开该空间情况下的保护,即在传输过程中和在 UNIX 系统中的保护。

结合公用数据和私有数据的划分，Gonia-office 系统的通讯示意图见图二。

1.3 用户标识

公用数据（Gonia-officeNT 产生的数据）和私有数据（Gonia-office98 产生的数据）以用户标识作为区分分别入库。用户标识是 Gonia-office98 客户端安装时配置的一个面向用户的特定字符串，在整个系统范围内唯一，公有数据不带用户标识，因此用户标识是 Gonia-office 用于区别私有数据和公有数据的唯一依据，也是 Gonia-office98 实现访问控制的最重要和最基本的手段。用户通过浏览器使用搜索引擎以实现信息定位。公用数据不需鉴别，所有用户均可使用，而对私有数据的定位则需通过用户鉴别后方能进行，用户只能查找与自己的标识匹配，即权限范围内的数据。



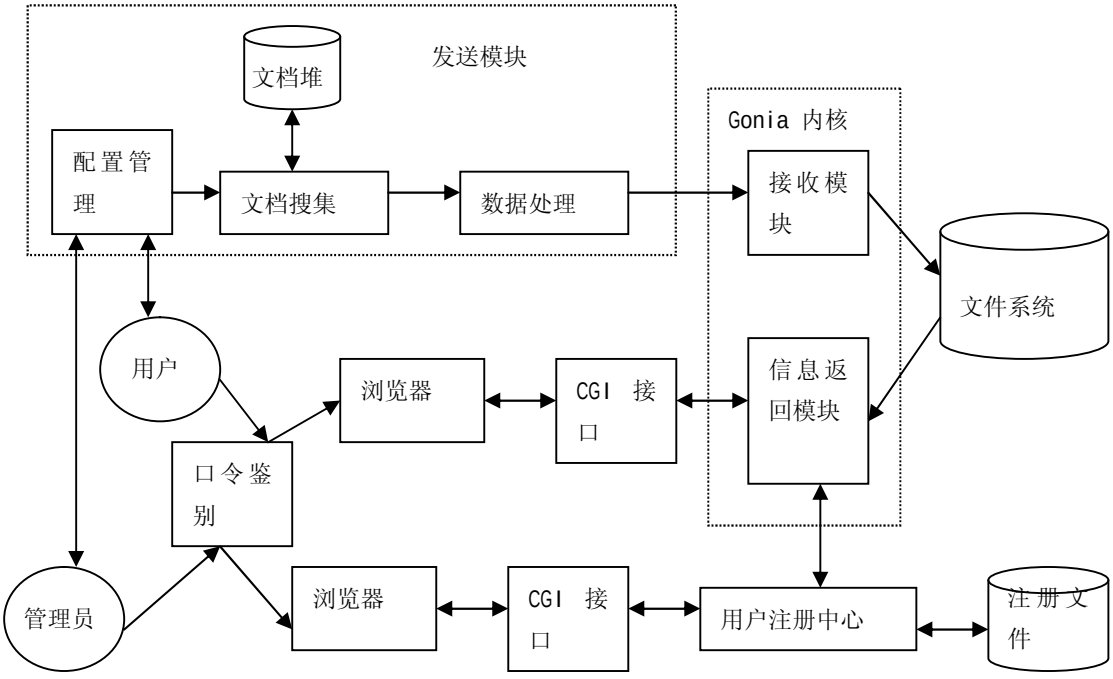
图二　系统通讯图

2 系统实现

2.1 工作原理

下图（图三）为系统的基本工作原理。可分步骤描述如下：

1) 管理员通过口令鉴别确定管理员身份后为在用户注册中心为用户注册
2) 用户得到标识，并完成相应配置
3) 发送模块按照指定配置搜集数据并发送，用户可通过修改配置定制发送数据的要求
4) 接收模块收到数据，存入 Gonia 的文件系统
5) 用户在口令鉴别后通过浏览器发出查询要求
6) 信息返回模块根据确认用户身份后，到用户注册中心获取用户标识，根据该标识进行允许范围内的查询，并返回结果



图三 系统工作原理

2.2 发送模块

发送模块功能包括搜集和发送数据，发送的数据包含用户标识信息。关于所发送数据的位置等信息都是系统的可更改配置，由于私有数据都是用户本机上的文档类信息，因此，可将数据鉴别所需的用户标识信息作为配置文件配在用户本机上，在数据发送时读取并标识该数据。对于公用数据则可加特殊标识以区分。

管理员和用户分别可修改公用数据和私有数据的配置，以此来定制特殊的发送需求。考虑到用户不可能将自己的数据标识为其它用户的信息，所以该用户标识也是允许发送的用户修改的

发送模块采用 VC 和 VB 编写，主体程序采用原系统，主要添加了标识项。另外，为了进一步提高安全性以及方便用户，将原系统支持的绝对路径改为相对路径，也即，用户指定要发送的路径之后，该路径便不再显示，只显示其下的相对路径。

2.3 用户注册中心

用户注册中心功能包括用户登记及查询。登记是管理员添加新用户的操作，查询是为信息返回模块调用以确定用户 ID 的。为了安全地确认用户，需要管理员的人工干预。用户注册即是标识的分配，用户得到该标识后即可在用户端建立配置文件，同时注册中心保存 CL 表，记录每个用户的权限及对应标识。

WWW 服务器一般能够提供较好的访问控制功能，本系统即采用了 Apache 的 WWW 服务器来做用户鉴别和基于角色的访问控制。

新用户注册时，先与管理员联系，管理员经过人工确认后，将给用户添加入合适的组中（即分配角色），然后到用户注册中心为用户注册。注册中心先查看 CL 表，然后按一定规则为该用户分配标识。当用户进行提交检索要求时，信息返回模块也要先到用户注册中心查询 CL 表，以得到该用户标识。

该部分程序采用 C 语言编写，在 UNIX 上工作，与管理员的交互采用了 CGI 接口通过浏览器进行。

2.4 信息返回模块

信息返回模块功能包括用户鉴别及根据鉴别结果的限制搜索。首先根据 WWW 服务器的访问控制进行基于口令的用户鉴别，得到用户名，而后该模块在用户发出查询请求后根据用户的查询要求以用户名为匹配标识向注册中心查询该用户的标识信息，从而可根据得到的用户标识进行限制搜索。在 Gonia 的文件系统中，查得所有含该标识的数据，返回给用户即可。

本模块涉及 Gonia 内核，采用 C 语言编写，在 UNIX 上工作。与用户的交互部分也是采用 CGI 接口通过浏览器进行。

3 结论

本系统采用的开发环境为 Windows 系统下的 VC++、VB 和 Unix 下的 gcc，目前运行在华东北地区网络中心内部管理网 Intranet 上，提供公用数据和私有数据的信息查询服务，当然还包括用户注册。对于公用数据，任何人在 Intranet 的任何地方都可以进行查询；对于私有数据，用户在本机发送后，通过浏览器确认该用户身份后即可查询。

Gonia-office98 充分利用了 Gonia-office 系统原有的功能，并利用了 WWW 服务器本身的安全鉴别功能，以较小的代价获得了一个性能稳定并有较高价值的实用软件系统。

进一步的工作将集中在系统的安全方面，在其安全水平达到一个新的水准后，将可以在更加开放的环境中运行。

【参考文献】

1 高毓航 丁伟。 “内部网搜索引擎 GONIA－OFFICE 的设计与实现”。 CERNET 的研究与发展，1999，第四卷：448