

内部网搜索引擎 GONIA—OFFICE 的设计与实现

高毓航 丁伟

(东南大学 计算机系, 210096 南京)

【摘要】 本文介绍了一个面向内部网的专用搜索引擎—GONIA-OFFICE 系统的设计与实现, 该系统利用已有的 UNIX 平台上的搜索引擎内核 GONIA, 实现对 NT 局域网文件服务器上的资源定位。本系统可以处理 NT 文件服务器上的 OFFICE 系列文档与 HTML 等文件资源, 从 WEB 向用户提供信息查询服务, 同时与控制台操作一起提供灵活方便的系统管理。

【关键词】 内部网、信息定位, 搜索引擎, NT, OFFICE

1

1. 引言

由于 Intranet 文件服务器上的文档的存放位置完全由人为方式确定, 随着时间的迁移, 由于忘记路径而无法找到以前存放的文件是经常发生的事, 能否把已有的搜索引擎技术应用在这样的环境呢? 本文以 NT 服务器和 OFFICE 系列文档处理为背景, 利用华东北地区网络中心已有的搜索引擎内核 GONIA, 进行了这方面的研究和开发工作。

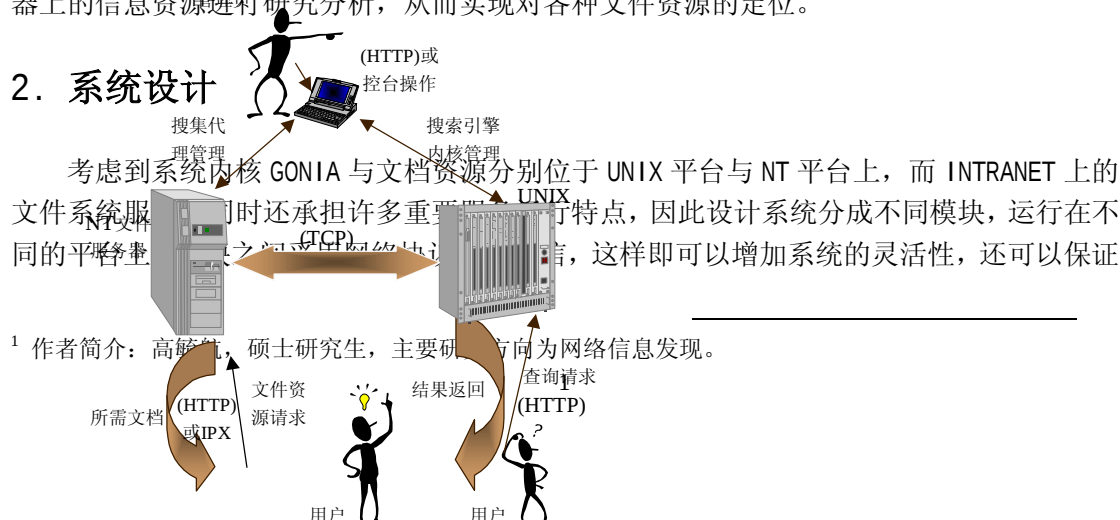
搜索引擎技术是在 Internet 上从海量的页面资源中建立用户与其感兴趣的页面之间的最有效的连接的一种信息发现方法。同其他信息查询工具一样, 搜索引擎把内容纷繁复杂, 形式多样的海量信息内容按照一定的形式组织起来, 通过在数据集查找与用户需求属性相关的元素, 为用户方便快捷地查找各种有用信息提供了一种很有效的途径。用户通过输入某主题的关键词就可以得到按照与查询关键词的相关程度 (RSV) 的大小进行排序的页面地址条目, 并以一种标准格式在用户本地的浏览器上显示出来。目前在 Internet 上的搜索引擎大多是以类似“蜘蛛”(ROBOT) 的搜索器自动运行搜集网页。

与 Internet 中的搜索引擎不同, 局域网文件服务器上的信息资源是以目录树的结构进行组织, 其类型复杂多样, 而且活动性各异。因此在面向 Intranet 的引擎数据搜集处理端需要更灵活复杂的数据发现、收集处理及代理管理功能。

GONIA 是一个由 CERNET 华东 (北) 地区网络中心开发的有关通用搜索引擎, 它通过计算文档之间相关度 (RSV) 值的分布式搜索引擎内核。它采用了低语意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。用层次化的方法实现数据信息源的特征屏蔽, 提供一个标准化的数据收集处理接口。

本文所提出的 GONIA-OFFICE 系统则利用了 GONIA 的索引功能, 并对局域网络文件服务器上的信息资源进行研究分析, 从而实现对各种文件资源的定位。

2. 系统设计



系统运行不影响到文件服务器的其他服务。如图 1 所示, 搜集代理及相关管理模块运行在 NT 上; GONIA 内核与用户查询代理部分则运行在 UNIX 平台上, 通过 WEB 的 CGI 程序向用户提供查询请求接收与结果返回功能, 二者之间通过 TCP/IP 协议相互通信。用户可以直接点击返回结果中的资源地址链接, 或通过文件共享的方式按相应地址进一步查看所需信息资源。系统管理员可通过 WEB 浏览器或直接从控台上分别管理搜索引擎内核与搜集代理。

再仔细地研究系统工作过程与总体结构，可以从下图看出系统的总体组织结构。

图 2 系统总体总体组织结构

由于应用环境的差异,本系统的用户查询服务模块除了与一般搜索引擎相类似的查询关键词的提交与结果返回的功能,还设计提供了多种查询精度选择与按资源类型进行结果范围界定,例如使用户能选择采用准确方式来查找有关“信息发现”的所有 POWERPOINT 文档。

对上面的系统设计进行具体的实现分析,其重点在格式过滤与代理管理两个部分。格式过滤部分的主要功能模块的实现主要根据系统的运行环境特点来决定;而代理管理的实现主要考虑其配置的灵活性和易管理性。

能完全滤去多种含复杂格式的文档的格式信息，并且保有简单方便的扩充接口，以便将来能轻易地增加系统信息资源的涵盖面到 NT 平台上的其他类型文档。因此在设计格式过滤器时采用了集中性与分散性相结合的模块组织结构，如图 3 所示。分散性体现在各类不同类型的文档都单独提供元格式过滤器，而且元格式过滤器都是以插件的形式存在；集中性体现在模块对接收到各种不同类型的文档先统一判断类型，而且通过统一的 API 接口来调用各对应元格式过滤器完成格式过滤功能，各元过滤器处理后得到的数据采用统一中间格式送交下面的文本数据处理与封装模块进行进一步的处理。当系统将来要扩充到对 NT 其他格式的文档进行定位时，所需的系统维护工作仅限于编写新的元格式过滤器和在类型判断与分配单元中增加对新文档类型的指向，因此系统维护工作量很小。

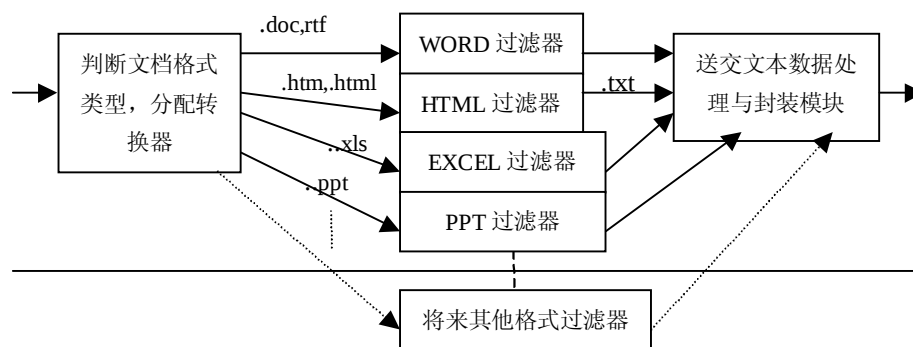


图 3 文档格式过滤器结构框图

搜集代理的管理是用来对搜集代理的运行方式与系统参数进行灵活的管理。由于系统代理模块的平台分布性，常常需要能动态配置通信参数。而且文件服务器上的各不同目录的活动性也是不同的，有的可能完全用来存档，还有的每天都在更新，因此各自的更新周期大不一样，需要搜集的目录结构与其资源类型也千差万别。所以需要一套灵活、方便的动态管理与配置功能来保证搜集代理的运转。

NT 平台上的搜集代理配置管理提供了对搜集代理正常运转所用的各参数的管理功能。从运行形式上，分为两方面实现：一部分提供在 WEB 上从远端管理配置信息的功能；另一部分则是在本地的应用程序系统中提供方便、有效的各种参数配置管理。这两部分的核心功能是一致的，即提供合理、全面的对待搜索目录、免搜索目录和基本确省参数的有关信息管理。搜索目录和免搜索目录管理一起界定了从起始目录开始的文档的搜集范围及一些用于自动定时搜索的参数，如目录的更新频率与最近更新时间等。

4. 关键技术

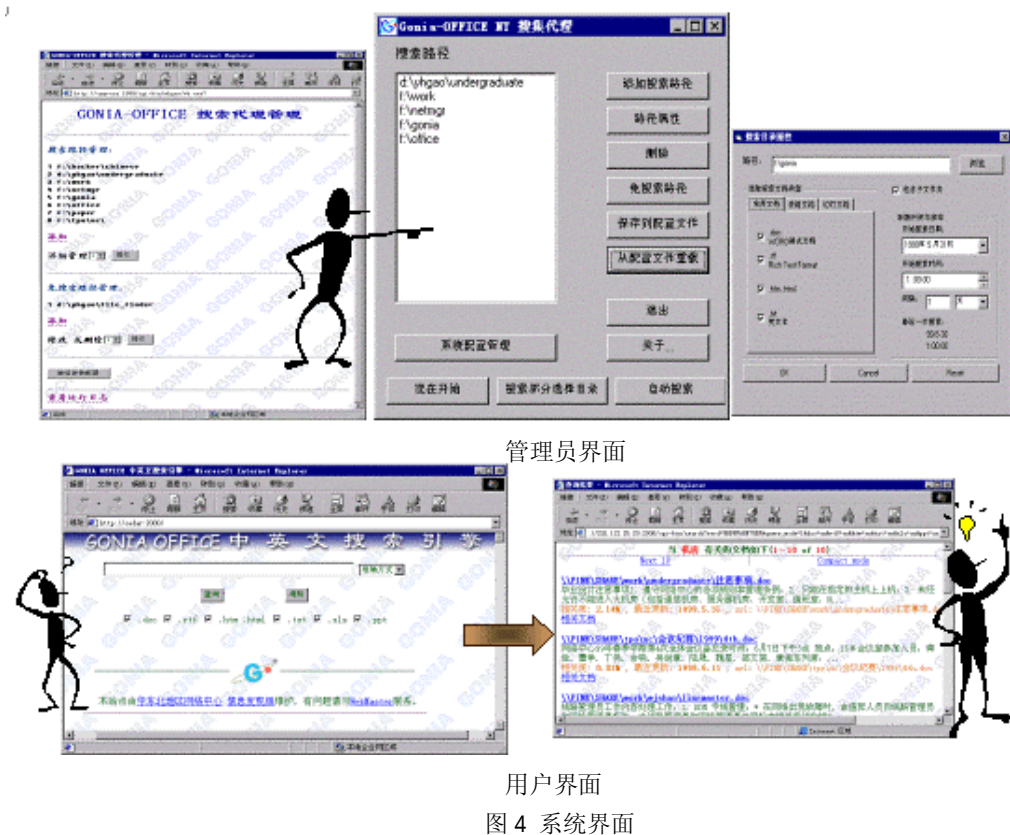
在 Internet 上，所需的信息大多源于用 HTML 编写的相互链接的网页，其语法公开、较简单。但是，在 Intranet 文件服务器上的情况就不一样了。在主要以 NT 为平台的文件服务器上以目录树的结构存放的存档文档类型多为 OFFICE 套件系列，其格式复杂、不公开，因此本系统在数据搜集时文件发现算法采用了广度优先算法，格式过滤器主要用 Automation 技术实现。

Automation 的前身是 OLE（动态对象链接与嵌入），支持在用 C++或 Visual Basic 或是用其他描述语言写的应用程序之间进行通信。现在是包括 Microsoft Word, Microsoft Access 和 Excel 在内的大多数 Microsoft 应用程序的编程标准。这里，就通过 OFFICE 的

Automation 接口来直接调用各应用程序自带的转换功能。使用 Automation 技术比单纯地人工分析出各类型的各不同版本的文档存储格式要更方便快捷,易于实现,减少大量的工作量。

5. 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, 目前运行在华东地区网络中心内部管理网 Intranet 上, 提供信息查询服务, 有三种查询方式供选择: “精确方式”、“准确方式”和“模糊方式”, 支持用户反馈与多种类型范围界定, 如 WORD、EXCEL、HTML、POWERPOINT 等。一般提交查询请求 1-2 秒后, 即可得到满意结果。搜集代理支持即时搜索与定时自动搜索方式, 可从应用程序控台或 WEB 管理。系统管理界面与查询服务界面分别见下图所示。



本系统的进一步完善重点是实现增量索引和即时搜索功能。增量索引是指当收集代理进行自动定时搜索时, 只把在上次搜索以后被修改过的文档送交内核进行索引分析。即时搜索则是即时地发现与收集文档信息资源, 并使内核立即把相应的文档分析、入库。

今后的工作除了进一步完善本系统外, 主要扩展方向是提供面向 WINDOWS 9X 系统的更具个性特色的个人桌面搜索引擎。由于个人搜索引擎与局域网中的搜索引擎在运行环境与面向的资源是相类似的, 因此可以将缩小局域网中的搜索引擎规模, 并按各种不同用户的需求增加更个性化的个人管理与查询功能, 把搜索引擎技术应用于个人的信息资料定位管理。

【参考文献】

[1] Peter Schauble. "Multimedia Information Retrieval, Content-Based Information Retrieval from Large Text and Audio Database". Kluwer Academic Publishers, 1997.

- [2] Microsoft Corporation, 《Visual Basic 6.0 中文版程序员指南》, 北京希望电脑公司出品, 1998.
- [3] Microsoft Corporation, 《Visual Basic 6.0 中文版语言参考手册》, 北京希望电脑公司出品, 1998.
- [4] David J.Kruglinski, Scot Wingo, George Shepherd, 《Programming Visual C++ 6.0 技术内幕 (第五版)》 [美], 北京希望电子出版社, 1999.4.
- [5] 贾耀炜、杨华中编著, 《HTML 语言与主页设计》, 人民出版社出版发行, 1998.
- [6] 蔡奇玉, 连振汉, 张志强著, 《CGI 编程指南》, 机械工业出版社, 1997.

Designation and Realization of an Intranet Search Engine

GONIA-OFFICE

Gao Yuhang, Ding Wei

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

【Abstract】 In this paper, an Intranet search engine named GONIA-OFFICE is introduced, which focuses on resources from NT file system server with the base of an existing search engine core called GONIA. At present, the system can deal with OFFICE documents and HTML pages etc. From WEB, both normal search services for users and easy configuration management of gather agent are provided. This, together with flexible management on console, caters the system for practical use.

【Keywords】 Intranet, information locating, search engine, NT, OFFICE