学校编码：10384　　分类号＿＿＿密级＿＿＿
学号：X2011230113　　UDC ＿＿＿

厦门大学

工 程 硕 士 学 位 论 文

基于搜索引擎的局域网涉密信息检测系统设计与实现

Design and Implementation of Secret-Involved Information Detection System in LAN based on Web Search Engine

潘仕杰

指 导 教 师： 史 亮 副 教 授

专 业 名 称： 软 件 工 程

论文提交日期： 2014 年 6 月

论文答辩日期： 2014 年 7 月

学位授予日期： 年 月

指 导 教 师：＿＿＿＿＿＿

答辩委员会主席：＿＿＿＿＿＿

2014 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

在经济、军事、社会信息化的今天，信息的作用和地位越来越突出，能否掌握并且控制信息资源已然成为国家稳定的决定因素。目前，人们对于信息安全问题的认识大多停留在防范来自外部的黑客攻击和病毒对内部网络的侵害，而对来自内部网络的危害重视不足，认为仅仅通过防火墙和防杀病毒软件就可以确保内网安全。但事实并非如此。根据我国国家信息安全评测认证中心的调查结果显示，信息安全问题的 80%来自内部。

随着电子政务的开展，电子政务局域网内泄密事件屡有发生。如何能够及时准确的检测出涉密信息并以此预防，成为当今电子政务发展的一块拦路石。本文建立了一种利用搜索引擎工作机制的局域网涉密信息监控系统，通过爬虫爬行局域网内 WEB 页面或者遍历 FTP 服务器，然后对页面内容或 FTP 服务器上特定文件进行处理后添加至索引库，再通过相关涉密关键字对索引库查找从而在一定程度上发现局域网泄密问题。

本文涉及到 WEB 搜索引擎的基本原理、核心技术和处理流程，并针对涉密信息对其分词词库、爬行方式做出一定优化，旨在尽可能提高其效能。其研究内容主要是借鉴搜索引擎的工作方式，检查局域网内服务器是否含有涉密信息，在一定程度上对局域网内公开信息进行审核，保证涉密信息不泄漏在局域网上，减少泄密的可能性。本文对于在电子政务局域网内防泄密方面有一定的应用价值。

关键字：局域网，涉密，搜索引擎

Abstract

Informations play an important role in the status and function of development of Economic, military and social informationization. How to master and control the information resources have been the decisive factors of national stability. Now,people awareness of information security issues still stuck in preventing hackers attackts and computer viruses but ignor the harm coming from network inside.They believe that firewall and antivirus software can ensure the network security . But, it is not the case. The results of the survey coming from CNITSEC show that about 80% information security problems come from network inside.

With the launching of e-government, e-government within the LAN leak happened frequently. How to timely and accurate of the Secret-Involved information detection and to prevent, become a stepping stone in the development of e-government.This article established a working mechanism of the search engine's LAN Secret-Involved information monitoring system,then added a specific file in the page content or FTP server after processing to the index library.Through secret keyword finding in the index database thus to find the LAN leak problem to a certain extent.

This article relates to fundamental principles of Web search engines, core technology and process, and optimized for Secret-Involved information to make certain, to maximize its effectivenes. The research audits public information of LAN in a certain extent to inspect the Secret-Involved information within the LAN referencing the workings of a search engine. It ensures that classified information does not leak on the LAN and reduce the possibility. This article has some value in research in leak-proof.

Keywords: LAN, Secret-Involved, Search Engine.

目 录

Contents

第一章 绪论

1.1 研究背景与意义

在经济、军事、社会信息化的今天，信息的作用和地位越来越突，能否掌握并且控制信息资源已然成为国家稳定的决定因素。

目前，人们对于信息安全问题的认识大多停留在防范来自外部的黑客和病毒对内部网络的侵害，而对来自内部网络的危害重视不足，认为仅仅通过防火墙和防杀病毒软件就可以确保内网安全。但事实并非如此。

根据我国国家信息安全评测认证中心的调查结果显示，信息安全问题的 80%来自内部，15%来自内部与外部的勾结，只有 5%来源自外部。这正是反映了“堡垒最容易从内部攻破”的道理。毫不夸张的说，一个能进入办公室打开涉密计算的普通员工所带来的威胁，远远比一个超一流水平的黑客更为严重。

以此，本文从上述问题出发，借鉴搜索引擎的工作方式，对局域网内 Web 等服务终端进行索引，从而检查局域网内服务器是否含有涉密信息，在一定程度上对局域网内公开信息进行审核，保证涉密信息不泄漏在局域网上，减少泄密的可能性。

1.2 网络安全现状分析

随着网络技术的发展，计算机网络安全问题越来越复杂和严重，攻击手段复杂多变，攻击途径层出不穷，使得传统的信息安全防御模式面临严重挑战。这其中主要问题表现如下：

（1） 认为信息安全建设就是利用技术手段来解决安全问题，却忽视信息安全技术集成的重要性，造成信息安全孤岛的存在；

（2） 重视技术，却轻视管理和维护；

（3） 片面强调对来自外部攻击的防御，忽视内网安全监控；

（4） 安全防御重视局部，比如各个部门都根据自己的需要进行访问控制，但缺乏整体安全考虑，因此导致整体安全和服务达不到预期要求；

（5） 存在“欠安全”和“过安全”问题

1.3 研究内容

由于目前专门针对党政军等涉密机关的密级文件进行相关安全检测防泄漏的系统发展的还远远不够完善，所以此领域有着巨大的发展前景和价值，本系统的设计内容新颖，具有以下几个方面的特色：

设计思想上，借用搜索引擎的工作方式，将传统局域网人工涉密信息检查变为智能化，利用原有局域网资源，进行密级文件的搜索与识别，省时，省事，方便，高效，真正做到提高工作效率。

实现方法上，利用现有 Lucene.Net 开源库，全新实现基于字典的正向最大长度切词，进行安全可靠的中文索引过程，面向用户简单易用的检测过程。

本课题借鉴搜索引擎的工作方式，在局域网内运用爬虫技术以及中文分词技术对 FTP 服务器和 Web 页面进行索引，从而得到使用中文分词技术的索引文件。而后通过特殊关键字搜索的方式，检查局域网内服务终端是否含有涉密信息，最终实现利用基于搜索引擎的局域网涉密检测系统进行涉密检测。

综上所述，本课题的核心部分就是中文搜索引擎的搭建，并利用搜索引擎实现相关功能。

细化来说，就是使用 Lucene.Net 作为全文本索引库，结合自定义的分词器(Analyzer，中文分词)和爬虫(Crawler，包括 Web 爬虫和 FTP 爬虫)进行索引和查询的过程。

1.4 论文结构安排

论文共分六章。

第一章，绪论。本章节主要介绍了本课题的研究背景及其意义，重点阐述了本系统的研究特色，及其相关开发环境的简介。

第二章，搜索引擎的相关技术。本章节主要介绍了搜索引擎的相关技术，主要包括网络爬虫、文本挖掘、全文检索、中文分词以及分布式的相关技术。

第三章，需求分析。本章节首先对基于搜索引擎的涉密检测系统进行了功能分析，然后针对功能性需求和非功能性需求分辨阐述。

第四章，局域网涉密信息检测系统设计。本章节首先对系统的概况进行介绍，然后针对系统功能呢个模块进行着重介绍，结合图例、代码介绍了 FTP、Web 爬虫，中文分词技术的实现，全文检索的设计，索引汇总以及索引查询等具体实现。

最后介绍了用户界面相关设计与实现。

第五章，局域网涉密信息检测系统的测试及测试结果。本章简要介绍了测试环境、测试过程以及测试结果，最后着重根据测试情况，对测试结果进行了逐项分析。

第六章，总结与展望。对本论文的主要研究内容作出总结，并指出下一步的研究方向。

第二章 系统相关技术介绍

本章为介绍系统所涉及到的开发工具和相关技术。简介了 Visual Studio 2010、C#.Net、Lucene.Net 等，然后对系统开发需要的相关技术进行介绍，最后对 B/S、C/S 模式进行简介。

2.1 Visual Studio 2010

Visual Studio（简称 VS）是 Microsoft 公司的开发工具套件系列产品。VS 是一个基本完整的开发工具集,它包括了整个软件生命周期中所需要的大部分工具，如 UML 工具、代码管控工具、集成开发环境等等。所写的目标代码适用于微软支持的所有平台。

Visual Studio 2010，代号为“Hawaii”，于 2010 年 4 月 12 日上市。Visual Studio 2010 带来.NET Framework 4.0 并且支持开发面向 Windows 7 的应用程序。除了 Microsoft SQL Server，它还将会支持 IBM DB2 和 Oracle 数据库。它将具有内置的 Microsoft Silverlight 开发支持，包含一个交互设计器。Visual Studio 2010 将会提供一些工具来使并行计算更加简单：除了.NET Framework 的本地代码并行扩展以及并行模式库（Parallel Patterns Library），Visual Studio 2010 还包含了用于调试并行程序的工具。这些新工具使并行任务以及它们的运行时堆栈可视化，可以用来可视化线程等待时间以及线程在多核心之间的移动。

2.2 C#.Net

C#是微软推出的一种基于.NET 框架的、面向对象的高级编程语言。C#由 C 语言和 C++派生而来，继承了其强大的性能，同时又以.NET 框架类库作为基础，拥有类似 Visual Basic 的快速开发能力。C#由安德斯•海尔斯伯格主持开发，微软在 2000 年发布了这种语言。

ECMA 标准列出的 C#设计目标：

- C#旨在设计成为一种“简单、现代、通用”，以及面向对象的程序设计语言
- 此种语言的实现，应提供对于以下软件工程要素的支持：强类型检查、数组维度检查、未初始化的变量引用检测、自动垃圾收集。软件必须做到强大、持久，并具有较强的编程生产力。

- 此种语言为在分布式环境中的开发提供适用的组件开发应用。
- 为使程序员容易迁移到这种语言，源代码的可移植性十分重要，尤其是对于那些已熟悉 C 和 C++的程序员而言。
- 对国际化的支持非常重要。
- C#适合为独立和嵌入式的系统编写程序，从使用复杂操作系统的大型系统到特定应用的小型系统均适用。
- 虽然 C#程序在存储和操作能力需求方面具备经济性，但此种语言并不能在性能和尺寸方面与 C 语言或汇编语言相抗衡。

2.3 Lucene.Net 简介

Lucene.Net 是一个开源的全文检索工具包，使用 C#实现。它提供了一组丰富的 API 以供开发者为.Net 应用加入全文检索功能。

Lucene.Net(http://incubator.apache.org/lucene.net/)是由其 Java 版本移植过来的。Java 版本直接叫做 Lucene。Lucene 经过 10 多年的发展，拥有大量的用户和活跃的开发团队。到目前为止，Lucene 的 C#移植有三个版本，最初的是 NLucene，然后是 Lucene.Net，当 Lucene.Net 在 2.0 版本之后开始了商业路线，又出现了 dotLucene 项目。本文中所使用的是 Apache 开源平台下的 Lucene.Net。

Lucene.Net 使用布尔模型来确定哪些文档匹配上查询词，使用向量空间模型(VSM)来对这些文档评分。一次搜索返回的结果可以有很多页，用户只看搜索结果中和查询词最相关的前面几条。所以可以使用优先队列收集最相关的 n 个文档。把搜索结果看成一个优先队列，相关度分值高的文档排在前面。

搜索引擎的灵魂是索引，是文档的最佳组织形式。搜索过程是对索引库执行折半查找。信息是有结构的，叫做文档。往 Lucene.Net 的索引库中放的是文档，然后按照词给这些文档建立索引；查询的是词，查询返回的也是文档。使用 Lucene.Net 的过程如图 2-1 所示：

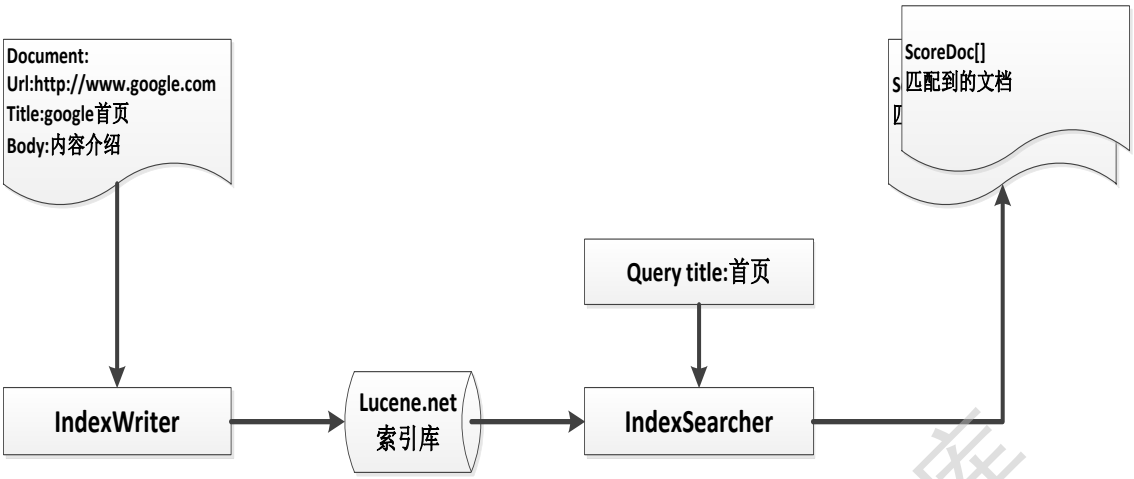

图 2-1 Lucene.Net 的索引过程

在 Lucene.Net 官方网站对其的简要说明中，给出了一个定义：“Lucene.Net 是一个用 C#编写并且运行在.Net 运行环境下的 Lucene 搜索引擎库的接口”。换言之，Lucene.Net 是 Java 版本 Lucene 的移植版本。

2.4 密级文件的特征

《中华人民共和国保守国家秘密法》自 2010 年 10 月 1 日起正式施行。其第九条规定——下列涉及国家安全和利益的事项，泄露后可能损害国家在政治、经济、国防、外交等领域的安全和利益的，应当确定为国家秘密：（一）国家事务重大决策中的秘密事项；（二）国防建设和武装力量活动中的秘密事项；（三）外交和外事活动中的秘密事项以及对外承担保密义务的秘密事项；（四）国民经济和社会发展中的秘密事项；（五）科学技术中的秘密事项；（六）维护国家安全活动和追查刑事犯罪中的秘密事项；（七）经国家保密行政管理部门确定的其他秘密事项；政党的秘密事项中符合前款规定的，属于国家秘密。

第十条规定——国家秘密的密级分为绝密、机密、秘密三级。绝密级国家秘密是最重要的国家秘密，泄露会使国家安全和利益遭受特别严重的损害；机密级国家秘密是重要的国家秘密，泄露会使国家安全和利益遭受严重的损害；秘密级国家秘密是一般的国家秘密，泄露会使国家安全和利益遭受损害。

第十七条规定——机关、单位对承载国家秘密的纸介质、光介质、电磁介质等载体（以下简称国家秘密载体）以及属于国家秘密的设备、产品，应当做出国

家秘密标志。不属于国家秘密的，不应当做出国家秘密标志。

根据《党政机关公文处理工作条例》（中办发〔2012〕14 号）文件，第三章第十条党政机关公文格式规定中，制作涉密文件的标识的格式为：1、第一行为 6 位的份数数字编码。例如：000001；2、第二行为密级+★+保密期限，例如：机密★1 年；3、第三行为紧急程度，例如：特急。4、涉密文件标识应当标注在文件、资料首页或封面左上角，如图 2-2 所示：

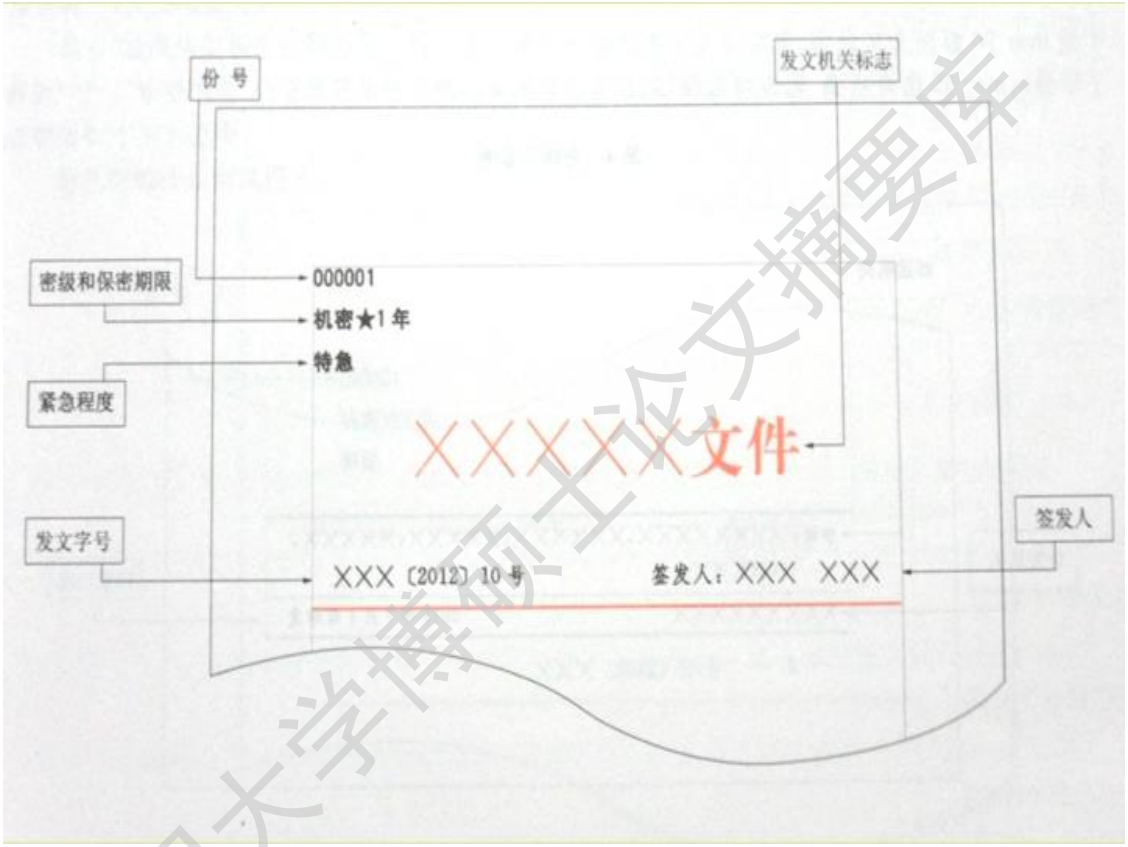

图 2-2 涉密文件标识样板

2.5 网络爬虫的相关技术

网络爬虫(Crawler)的主要作用是获取局域网上的信息。网络爬虫利用主页中的超文本链接遍历 Web，通过 URL 引用从一个 HTML 文档爬行到另一个 HTML 文档。网络爬虫收集到的信息可有多种用途，如建立索引、HTML 文件的验证、URL 链接验证、获取更新信息、站点镜像等。

网页的抓取策略可以分为深度优先、广度优先和最佳优先三种。深度优先在很多情况下会导致爬虫的陷入(trapped)问题，目前常见的是广度优先和最佳优先