IBM Content
Manager for Multiplatforms

管理 Information Integrator for Content

版本 8 发行版 2

S152-0232-01

IBM Content
Manager for Multiplatforms

管理 Information Integrator for Content

版本 8 发行版 2

S152-0232-01

注意

在使用本资料及其支持的产品之前，请阅读第 119 页的『声明』中的信息。

第二版（2003 年 3 月）

本版本适用于 IBM Enterprise Information Portal for Multiplatforms V8 R2（产品号 5724-B43）及所有后续发行版和修订版，直到在新版本中另有声明为止。

目录

关于本指南

本指南提供对管理 Enterprise Information Portal（EIP）系统时需要了解的所有基本概念的介绍。由于 EIP 提供了几个您可以从管理客户机进行管理的组件，而且您可以通过使用 EIP 访问其它产品功能，所以本指南不是典型的系统管理指南。本文档集中于以下主题并且解释了如何执行以下操作:

- 使用 EIP 满足商务需要
- 访问和使用管理客户机
- 管理用户访问权限
- 使用 EIP 搜索多台内容服务器上的内容，包括存储在关系数据库中的结构化数据、无结构或多媒体内容或者文本文档
- 设计、实现与管理工作流

本指南的读者

本指南帮助 EIP 管理员执行以下任务:

系统管理
包括数据库、服务器和网络管理

用户管理
定义和授权对于个体和组的访问，维护访问控制表

联合搜索
定义和使用联合搜索模板从内容管理系统检索内容

信息发掘
从文档抽取信息，对文档和搜索结果进行分类

Web 搜寻
使用 IBM® Web Crawler 从 Web 搜索和导入内容

文本搜索
使用 IBM DB2® TIE 或 IBM Text Search Engine（仅 Content Manager V7.1 和更早版本）搜索文档并对其建立索引

图像搜索
使用 Content Manager V7.1（和更早版本）执行图像搜索

工作流管理
使用 EIP 工作流工具管理企业的信息工作流

管理员所需的技能

根据执行的任务，您必须了解以下内容:

- 用户访问的安全性协议
- Windows NT®、Windows® XP、Windows 2000、AIX® 或 Solaris 操作系统
- 网络管理

- 内容管理系统中内容服务器的数据模型
- 数据库管理
- 创建搜索模板时如何应用内容和搜索条件的工作原理知识。
- 信息发掘技术和工具
- 工作流设计的原则
- 希望 EIP 工作流支持的商务过程

商务分析者或进程模拟者所需的技能

商务分析者和进程模拟者还可以在本指南中找到关于如何为其企业定义与建模 EIP 工作流的概念性信息。

要使用 Enterprise Information Portal 工作流构建器，您必须具备以下条件:

- 理解企业的商务过程中所使用的人员需求、程序和数据结构。
- 对企业的商务或工作流过程作决策。

在何处找到更多信息

产品软件包包含了一组完整的信息，可以帮助您规划、安装、管理和使用系统。产品文档和支持同样可以在 Web 上得到。

产品软件包中包含的信息

产品软件包以可移植文档格式（.PDF）包含信息中心和每个出版物。

信息中心

产品软件包包含安装产品时可以安装的信息中心。关于安装信息中心的信息，请参阅 *Planning and Installing Your Content Management System*。

信息中心包含 Content Manager、Enterprise Information Portal 和 IBM Content Manager VideoCharger for Multiplatforms 的文档。基于主题的信息按照产品和任务（例如，管理）组织。除了提供的导航机制和索引之外，搜索工具也对可检索性有帮助。

PDF 出版物

您可以使用操作系统的 Adobe Acrobat Reader 联机查看 PDF 文件。如果没有安装 Acrobat Reader，则可以从以下 Adobe Web 站点下载它: www.adobe.com。

表 1 显示了 IBM Content Manager for Multiplatforms 包含的 Content Manager 出版物。

表 1. Content Manager 出版物

文件名	标题	出版物编号
install	*Planning and Installing Your Content Management System*[1]	GC27-1332-01
migrate	*Migrating to Content Manager Version 8*	SC27-1343-01
sysadmin	*《系统管理指南》*	S152-0231-01

订购 IBM Content Manager for Multiplatforms 时，您也会收到 IBM Enterprise Information Portal for Mulitplatforms。您也可以单独订购 IBM Enterprise Information Portal for

Mulitplatforms。表 2 显示了该产品包含的 Enterprise Information Portal 出版物。

表 2. Enterprise Information Portal 出版物

文件名	标题	出版物编号
apgwork	*Application Programming Guide for Windows*[1]	SC27-1347-01
ecliinst	*Installing, Configuring, and Managing the eClient*	SC27-1350-02
eipinst	*Planning and Installing IBM Information Integrator for Content*	GC27-1345-01
eipmanag	*《管理 IBM Information Integrator for Content》*	S152-0232-01
messcode	*Messages and Codes*[2]	SC27-1349-01

注意:

1. *Application Programming Guide for Windows* 包含编制 Content Manager 和 Enterprise Information Portal 应用程序的信息。
2. *Messages and Codes* 包含 Content Manager 和 Enterprise Information Portal 的消息和代码。

Web 上提供的支持

在 Web 上可得到产品支持。从以下的产品 Web 站点单击 **Support**:

www.ibm.com/software/data/cm/

www.ibm.com/software/data/eip/

文档包含在产品的软拷贝中。要访问 Web 上的产品文档，请单击产品 Web 站点上的 **Library**。

还可以从 Web 上得到基于 HTML 的文档界面，它又称为企业联机文档（EDO）。它当前包含 API 参考信息。有关访问 EDO 的信息，请转至 Enterprise Information Portal Library Web 页面。

如何发送意见

您的反馈有助于 IBM 提供高质量的信息。请将关于此出版物或其它 Content Manager 或 Enterprise Information Portal 文档的任何意见发送给我们。您可以使用以下任一方法提供意见:

- 从 Web 发送意见。访问以下位置的 “IBM 数据管理在线读者意见表（RCF）（IBM Data Management Online Reader’s Comment Form）” 页面:

  www.ibm.com/software/data/rcf

  您可以使用该页面输入和发送意见。
- 通过电子邮件把意见发送到 comments@vnet.ibm.com。请务必在意见中包含产品名称、产品的版本号以及书籍的名称和部件号（如果适用）。如果您对某特定文本发表意见，请指出文本所在的位置（例如，章节标题、表编号、页码或帮助主题标题）。

EIP V8.2 中的新内容

已对产品做了以下更改:

支持 Sun Solaris

您可以在 Solaris 系统上安装连接器、功能部件和数据库。

公共系统管理

单个客户机应用程序提供对 Content Manager 和 Enterprise Information Portal 管理的单独访问。

新的连接器

- Content Manager V8 R1 的 ICM 连接器允许您利用 Content Manager V8 强大的文档存储功能。
- 新的 C++ Extended Search V3.7 连接器运行于 AIX 上。

改善的连接器

- 支持从联合层和通过直接 Extended Search 连接进行参数文本搜索。
- 对 OnDemand 连接器的功能增强和性能改善，包括:
  - 对 OnDemand DDO 结构的修改。
  - 现在支持异步搜索

新的信息发掘服务

- 功能抽取
- 群集
- 语言标识

IBM Web Crawler

IBM Web Crawler 是允许用户在 Web 上和 Lotus Notes® 数据库中搜索和摘要信息的功能部件。

工作流增强

工作流现在在 AIX 和 Solaris 上得到了全面支持。工作流构建器、API 和 JavaBeans™ 提供了改善的工作流的功能和可用性。

信息中心

基于浏览器的信息中心包含 Content Manager、Enterprise Information Portal 和 IBM Content Manager VideoCharger for Multiplatforms™ 文档。基于主题的信息按照产品和任务（例如，管理）组织。除了提供的导航机制和索引之外，搜索工具也对可检索性有帮助。

辅助选项

辅助选项功能部件帮助身体有残疾的用户（如，行动不便或视力不佳）顺利地使用软件产品。此产品主要的辅助选项功能部件包括:

- 使用键盘代替鼠标操作所有功能部件的能力。
- 支持增强的显示特性。
- 视频和音频警告提示的选项。
- 与辅助技术兼容
- 与操作系统辅助选项功能部件兼容
- 可访问的文档格式

第 1 章 介绍 Enterprise Information Portal

许多消耗大量纸张的企业（例如保险公司和金融机构）都管理大量与商务有关的内容。许多行业都需要用于管理和访问商务信息的企业级解决方案。

*内容服务器*是一个软件系统，它存储了多媒体、商务表单、文档和相关数据、以及允许员工处理内容的元数据。当无法有效地连接到不同内容服务器时，公司可能需要浪费大量的时间和金钱来准备重复信息或进行员工培训，才能执行多次搜索。

Enterprise Information Portal（EIP）提供前沿技术，将所有企业资源都付诸工作站台式机来处理。EIP 可以通过用单个客户机连接到完全不同的内容服务器帮助您使信息和多媒体资产的价值最大化。使用 EIP 客户机，用户可以快速、并发地访问所有连接的内容服务器。用户还可以跨越内容服务器（包括 Web 或内部网）进行信息发掘或高级搜索。他们可以在您定义的商务过程中执行工作流任务。

使用 EIP，可为您的企业定制应用程序。应用程序员可以使用 EIP 样本编写台式机和基于 Web 的应用程序。

本节提供了 EIP 概述。一个有关虚构保险公司 XYZ 保险公司的方案演示了 EIP 的特征和功能。

搜索客户信息

XYZ 保险公司（XYZ）是一个大型财产和意外事故保险公司，它拥有大量的照片、索赔、保险单、调停人按语、专家报告和其它商务文档的集合。

XYZ 在 Lotus® Domino™.Doc 文件柜中保存了所有送往保险单持有者的便笺、医疗和估价电子表格。XYZ 把所有保险声明、注意事项和发票存档在 Content Manager OnDemand 服务器中，以备长期存储和快速访问。XYZ 将所有从保险单持有者那里接收到的索赔表、照片和信件保存在 Content Manager for iSeries 系统文件夹中。XYZ 将来自专家的报告保存在 DB2 通用数据库™（DB2 UDB）数据仓库中心信息目录管理器中。XYZ 还将公司的媒体资产（如高分辨率图形）存储在 Content Manager 系统中，以供广告、公共关系和新业务部门共享。另外，XYZ 还将诸如公司流程等信息保存在公司内部网上。

需求

对于索赔、客户电话和一般保险单持有者服务，不能通过来自一个服务器的内容完成其处理，因为雇员需要访问所有客户信息。为了向客户提供服务，雇员需要同时访问不同的内容服务器。XYZ 保险公司需要一种解决方案，用于连接内容服务器和公司内部网，以便搜索和检索信息。他们还希望扩充对工作流处理过程的使用。

从普通职员到索赔调停人和代理人，许多不同的雇员都需要访问文档。XYZ 必须限制对某些项的访问，同时对其它项提供无限制地访问。XYZ 还需要易用界面，从而减少培训的需要。

解决方案

XYZ 保险公司采用 EIP，因为综合搜索技术允许他们连接和搜索所有内容服务器以检索数据。现在，当 XYZ 电话服务中心代表接收到一个电话时，只要使用一个联合搜索就可以检索到所有必要的保险单持有者信息。

XYZ 保险公司还使用 EIP 信息发掘功能从公司内部网搜索和检索信息。他们还希望扩充对工作流过程的使用。

概述

EIP 是一种全面的产品；其组件联合工作，为企业提供了特有的合适的解决方案。EIP 位于多层体系结构中央，提供管理搜索的管理客户机、运行搜索的客户机和连接到完全不同的内容服务器（如 IBM Content Manager、Content Manager ImagePlus® for OS/390®、Content Manager OnDemand、Lotus Domino.Doc、DB2 通用数据库、DB2 DataJoiner® 和 DB2 数据仓库中心信息目录管理器）的客户机。您可以使用 EIP 连接器工具箱和样本为额外的内容服务器编写额外的连接器。

EIP 体系结构允许客户机应用程序在一个或多个内容服务器上运行单个搜索。要执行搜索，客户机使用由 EIP 管理员定义的搜索模板。

使用搜索模板时，客户机运行*联合搜索*，这是同时在多个内容服务器（其本机属性已映射成在搜索模板中使用的联合属性）上运行的搜索。EIP 搜索模板包含搜索条件，该条件引用每个内容服务器上映射为本机属性的联合属性。EIP 管理员创建了搜索模板。EIP 提供了访问和搜索存储在多台内容服务器上的数据的连接器。然后，内容服务器将数据对象返回给客户机。

EIP 体系结构提供了以下优点:

- 使用单一查询来访问支持电子商务事务和客户服务应用程序的多个不同的内容服务器。
- 在多个内容服务器（包括 Web）上进行信息发掘的能力。
- 用于访问多个不同种类的内容服务器上的数据的工作流过程。
- 由于客户机应用程序、索引和数据的分离性，支持与数据在任何内容服务器上的位置无关的客户机应用程序开发。

介绍 Enterprise Information Portal 组件

本节说明每个 EIP 组件并描述安装选项。

第 2 页的表 3 列出了组件和兼容的操作系统。

表 3. EIP 组件操作系统兼容性

组件	Windows	AIX	Solaris	Notes
管理数据库	是	是	是	数据库包含工作流构建器功能
管理客户机	是	否	否	客户机可以连接到安装在 Windows、AIX 或 Solaris 操作系统上的数据库。

表 3. EIP 组件操作系统兼容性 (续)

组件	Windows	AIX	Solaris	Notes
连接器	是	是	是	
信息发掘	是	是	是	
IBM Web Crawler	是	是	是	
文本搜索客户机	是	是	是	
图像搜索客户机	是	是	是	
连接器工具箱和样本	是	是	是	• Windows 版本包含编译样本客户机的源代码。AIX 上不安装任何样本客户机代码。 • 工作流样本和 API 与联合连接器样本一同安装。
查看器	是	否	否	安装 OnDemand 客户机和查看器。
信息中心	是	是	是	

管理

管理组件提供了管理数据库和管理客户机子组件。安装管理数据库时，您还将安装工作流功能部件。

管理数据库: 管理数据库是管理有关 EIP 用户和组、特权级别、密码、用户标识的信息和其它信息的 DB2 数据库。该数据库还提供工作流，并可选地提供信息发掘功能。您可以安装多个数据库。每个数据库都提供了 EIP 工作流功能。如果您有 Content Manager V8 系统，则可以与 Content Manager V8 库服务器数据库共享 EIP 管理数据库。您可以共享该数据库的原因是库服务器数据库包含 EIP 需要的所有信息。

管理客户机: 管理客户机只能安装在 Windows 工作站上。您可以安装多个客户机。如果您有 Content Manager V8 系统，则可以从同一客户机管理 EIP 和 Content Manager V8。

客户机提供了允许管理员执行以下操作的界面:

- 为联合搜索定义每个内容服务器。
- 标识内容服务器上的本机实体和属性并将它们映射为联合实体。
- 维护所有用户定义的内容服务器的本机实体和属性的库存。
- 创建搜索模板。
- 标识和管理用户和组。
- 向用户分配特权和特权集。
- 定义对搜索模板的访问权，并对用户可以对从搜索中检索到的信息执行的操作设置条件。
- 定义和管理商务工作流过程。

连接器

连接器提供了 EIP 客户机、内容服务器和管理数据库之间的通信接口。内容服务器连接器（如 Content Manager V7.1 连接器）提供了允许 EIP 登录到服务器、搜索信息以及将信息返回到管理或最终用户客户机的功能。联合连接器将管理客户机连接到管理数据库。

EIP 提供了以下连接器:

- 联合连接器将 EIP 客户机连接到管理数据库。
- DB2 通用数据库 7.1、JDBC driver 1.3（仅 Java™）、ODBC 3.0（仅 C++）和 DataJoiner 2.1.1 的关系数据库连接器。
- Content Manager V7.1 服务器的 Content Manager 连接器。
- Content Manager V8.2 服务器的 Content Manager 连接器。
- Content Manager OnDemand V7.1 的 Content Manager OnDemand 连接器。
- Content Manager for VisualInfo™ for 400® V4.3 和 V5.1。
- ImagePlus/390 Folder Application Facility V3.1 和 Image Plus/390 ODM V3.1 的 Content Manager ImagePlus for OS/390 连接器。
- Domino.Doc V3.0a 和 Desktop Enabler V3.0a 的 Lotus Domino.Doc 连接器。
- Extended Search connector for Version 3.7.
- DB2. Universal Database Visual Warehouse™ V5.2 和 DB2 Universal Database V7.2 的 Information Catalog Manager 连接器。

功能部件

EIP 有四个可选功能部件。

Information mining

Information Mining 提供语言服务，以查找内容服务器上文本文档中的隐藏信息。处理文本文档期间，元数据被创建为可以汇总、分类和搜索。WebSphere® Application Server 4.0（标准或高级版）是信息发掘先决条件。而且，您还可以群集相似文档、从文档抽取特征（例如人员和公司名称）以及确定文档的语言。

图像搜索客户机

提供在 Content Manager V7 内容服务器上访问和管理图像搜索功能所需的接口。

文本搜索客户机

提供在文本搜索服务器上访问和管理文本搜索功能所需的接口。

IBM Web Crawler

网上搜寻器是基于 Java 的内容网上搜寻器和发掘器。网上搜寻器可以在内部网、外部网或网际网、Lotus Notes 数据库中搜寻内容，也可以通过 Domino、本地文件系统和 FTP 集合搜寻内容。

网上搜寻器可以从多种类型的内容中发掘元数据和文本。例如，HTML 内容可以对 URL、标题、主体、上次修改时间和诸如作者、关键字、描述等的元标记进行发掘。用户从给定类型内容的一组预定义发掘器中进行选择。内容和／或发掘的元数据保存在本地磁盘中。

内容查看器

安装 OnDemand 查看器还将安装 OnDemand 客户机和查看从 OnDemand 服务器检索的文档所需的其它文件。

连接器工具箱和样本

EIP 提供一个连接器工具箱，它包含可用于实验和测试不同 EIP 功能的样本程序，这些功能有:

- 连接和断开连接内容服务器
- 对内容服务器执行 SQL 和其它样本查询
- 确定内容服务器 MIME 类型等等

Windows 连接器工具箱: 要在 Windows 服务器上安装连接器工具箱和样本，您必须选择“开发工作站”机器类型。然后选择“连接器工具箱和样本”组件。可以为所有连接器安装样本程序，或选择个别样本以匹配您安装的连接器。

在 Windows 服务器上，连接器工具箱样本程序以下列方式组织:

`c:\CMBROOT\SAMPLES\activex\`*xx*

`c:\CMBROOT\SAMPLES\cpp\`*xx*

`c:\CMBROOT\SAMPLES\java\`*xx*

`c:\CMBROOT\SAMPLES\jsp\`*xx*

`c:\CMBROOT\SAMPLES\server\`*xx*

其中 *xx* 是包含每个可应用连接器的样本程序的目录名称，例如 `db2`、`od`、`dl` 等等。

在 AIX 服务器上，样本程序以下列方式组织:

`/usr/lpp/cmb/samples/cpp/icm`

`/usr/lpp/cmb/samples/java/`*xx*

`/usr/lpp/cmb/samples/jsp/`*xx*

`/usr/lpp/cmb/samples/server/exit`

其中 *xx* 是子目录名，例如 `beans`、`servlets` 等等。

在 Solaris 服务器上，样本程序以下列方式组织:

`/opt/IBMcmb/samples/java/`*xx*

`/opt/IBMcmb/samples/jsp/`*xx*

`/opt/IBMcmb/samples/server/exit`

其中 *xx* 是子目录名，例如 `beans`、`servlets` 等等。

样本程序包含描述这些程序并提供使用样本代码所需的服务器设置（环境设置、内存等等）的文档。

信息中心

信息中心组件包含 Enterprise Information Portal 信息中心。信息中心是基于 Web 的、可搜索版本的 Enterprise Information Portal 库。

第 2 章 管理客户机简介

管理客户机提供了 EIP 管理数据库和 EIP 管理员之间的接口。这一部分描述了客户机为帮助管理 EIP 系统而提供的很多功能部件和功能。

您可以从位于客户机左窗格中的图标访问某些功能部件和功能，如服务器定义和用户管理。您可以通过“工具”菜单栏访问其它功能。

使用“EIP 第一步”了解系统管理客户机

*第一步*是每次安装 EIP 时附带的一个模块。*第一步*为您提供了样本数据并且填充了对象，使您不必使用实际的数据。如果要浏览服务器定义、用户和组以及其它功能部件来了解管理客户机的基础结构和外观，则使用*第一步*。

管理 EIP

作为系统管理员，您可以通过管理客户机完成以下一个或多个任务:

- 定义内容服务器
- 管理用户和组
- 管理特权和访问级别
- 创建联合搜索模板
- 创建联合实体
- 创建子域，如果管理域已启用。
- 处理工作流，如果工作流已启用。
- 在 Content Manager V7 中创建联合文本实体。

管理用户和组

您可以通过创建用户标识和特权来允许用户进行访问以搜索和处理多台内容服务器上的文档。也可以通过将适当特权定义和指定给用户来限制对存储在系统中的数据的访问权。

使用管理客户机工具

本节描述了管理客户机提供的工具。

LDAP 配置

单击此选项时，EIP 会启动一个包含四个选项卡的窗口:

- LDAP 选项卡 — 您可以启用从 LDAP 服务器导入数据源，启用 LDAP 导入和认证，或者同时启用两者。
- 服务器选项卡 — 包含定义 LDAP 服务器规范（包括主机名、用户名、参照类型等等）的字段。
- 认证选项卡 — 包含定义安全套接字层信息的字段。
- 高级选项卡 — 定义有关最大记录和服务器超时的设置。

用户映射选项

此选项允许您禁用已启用的用户映射缺省设置。

联合用户映射编辑器

联合用户编辑器显示了用户列表并给出使您可以将用户映射到特定内容服务器的选项。

搜索模板查看器

搜索模板查看器提供了有关所有搜索模板的详细信息。该查看器提供了查看搜索模板详细信息的三个选项:

- 关联映射（缺省值）— 提供了有关联合实体的详细信息和其它有关搜索模板的详细信息
- 搜索模板 — 提供了有关缺省运算符、缺省值等等的详细信息
- 显示结果 — 提供了有关显示名称、显示宽度、条件顺序等等的详细信息。

服务器目录清单查看器

显示选定一个或多个服务器的目录清单。

日志查看器

使用日志查看器查看刷新服务器目录清单之后生成的日志。日志显示了新旧目录清单之间存在差异时的消息列表。

服务 选择服务以启用工作流和／或信息发掘。

管理域 选择管理域以启用管理域。管理域启用之后无法禁用。

MIME 类型编辑器

MIME 类型编辑器列出了每个内容服务器的以下信息:

- 内容类
- 文件扩展名
- 关系数据库（RDB）列
- MIME 类型

MIME 类型编辑器中列出的内容服务器名称是缩写，与您定义新的内容服务器时出现的内容服务器名称列表相对应。**技巧**: DL 是 Content Manager V7.1 内容服务器的缩写。V4 是 Content Manager for AS/400® 内容服务器的缩写。

您可以添加到、除去或编辑 MIME 类型编辑器中的缺省信息。

MIME 到应用程序编辑器

使用 MIME 到应用程序编辑器添加到、删除或编辑五个缺省的 MIME 到应用程序关联。MIME 到应用程序编辑器中定义的值和设置会影响最终用户客户机使用的查看器。

服务器类型定义

使用此工具定义系统程序员开发的所有定制服务器。

更改 DB2 标识／密码

选择此选项修改仅用于连接的 DB2 用户标识和密码。这是与管理员用户标识完全分离的用户标识。

介绍特权

本节描述了 Enterprise Information Portal 特权。展开“权限”图标可访问这四个特权。
技巧: 因为您能从相同的客户机管理 Content Manager V8 和 EIP V8，因此客户机显示所有特权（各客户机端特权之总和）。

特权

管理客户机提供预定义的特权、特权组和特权集。特权给予系统用户（例如管理员和客户机最终用户）以某种方式操作某个对象的权限。

特权 EIP 提供多种特权。特权就是以特定方式操作某一对象的权限。例如，如果您有客户机最终用户，则可以向他们分配 ItemAdd 和 ItemDelete 特权，这就使他们有权限对内容服务器添加或删除项。要查看特权，则展开“认证”图标并双击“特权”。要创建特权，右键单击“特权”并单击“新建”。

特权组 EIP 提供缺省特权组。特权组是相关特权的集合。例如，名为“管理 EIP”的特权组包含通常与管理 EIP 系统关联的五种特权:

- EIPAdminServer
- EIPAdminEntity
- EIPAdminTextEntity
- EIPAdminTemplate
- EIPAdminInfoMining

要查看或修改预分配给特权组的特权，则展开“认证”图标并双击特权组名称。要创建特权组，右键单击“特权组”并单击“新建”。

特权集 EIP 提供几种缺省特权集。特权集是定义用户角色的特权的集合。例如，名为 ClientUserCreateAndDelete 的特权集包含与客户机最终用户角色关联的 17 种特权，例如 Delete（删除项）、ItemAdd（添加项）等等。如果为客户机用户创建用户标识，并分配特权集 ClientUserCreateAndDelete，则用户可以登录到内容服务器并执行该特权集中包含的 17 个用户角色的任一角色。要查看或修改特权集，则展开“认证”图标，单击“特权集”并双击特权集名称。要创建特权集，右键单击“特权集”并单击“新建”。

管理用户和组时，您可以将特权与用户和／或用户组相关联。将特权集分配给用户组时，该组中的所有用户都可执行关联的特权集中包含的所有角色。

切换产品视图和数据库

如果您将 Content Manager 和 Enterprise Information Portal 作为企业解决方案的组成部分，则您可以从一个用户界面访问这两个系统管理客户机。过去，如果您安装了这两种产品，则必须打开两个独立的客户机。从一个客户机视图切换到另一个客户机视图提供了一种修改信息的便捷方式（同时适用于两台客户机）和对其它产品的快速访问。

要在不注销的情况下从管理 EIP 切换到管理 Content Manager，请转至主系统管理窗口并使用左侧窗格上的下拉菜单选择 Content Manager。

要在联合数据库之间切换，请转至客户机窗口中的左侧窗格并双击某个联合数据库图标。

您也可以在不退出客户机并登录到新数据库的情况下管理不同的数据库。管理客户机将显示 cmbds.ini 文件中列出的所有管理数据库的图标。要切换到另一个数据库，请单击图标。如果新的数据库具有不同于您在登录到客户机时输入的用户标识，则会提示您输入另一个用户标识。

对管理客户机的改善和增强

EIP V8.2 的特点在于 EIP 管理客户机的显著增强，包括:

改进的向导和对话框

新的对话框使得管理用户更加容易。新的向导使得定义及修改实体和搜索模板更加容易。用户仍可选择使用 EIP V7.1 支持的对话框。

共享的管理客户机

在同一系统上安装 EIP V8.2 和 Content Manager V8.2 时，这两个产品将共享一个管理客户机。如果您是这两个产品的管理员，则您登录客户机一次就可以在客户机中在这两个应用程序之间进行切换。您也可以在不注销并再次登录的情况下在管理数据库之间进行切换。

域管理员

您可以创建仅具有已定义域的管理特权的域管理员。

单一注册和 LDAP 支持

EIP 现在使用 Windows Active Directory 和 LDAP，使用户具有对多个内容服务器的单一注册访问权。

将管理客户机连接到本地管理数据库

如果在安装管理客户机的相同服务器上安装管理数据库，则连接本地客户机和服务器所需要的信息已存储在 cmbds.ini（一个存储数据库连接信息的文件）中。您不必执行任何安装后配置而可以使用本节中的步骤立即连接。**要求**: 如果使用 EIP 数据库安装实用程序创建其它本地数据库，则必须先手工修改 cmbds.ini 中所需要的信息才可连接到新的数据库。

1. 单击**开始-->程序-->Enterprise Information Portal for Multiplatforms 8.2-->管理**
2. 从“服务器”字段中的下拉列表选择本地数据库。
3. 输入管理员用户标识和密码并按“确定”。
4. 打开系统管理客户机。**技巧**: 如果使用了“EIP 第一步”，则样本数据库显示在客户机的左窗格中。

将管理客户机连接到远程管理数据库

下面是将 EIP 管理客户机连接到远程 AIX、Windows 或 Solaris 数据库的两种方法:

- 通过 RMI 服务器连接（更多信息请参阅 *Planning and Installing Information Integrator for Content*）。
- 定义连接，方法是使用 DB2 配置助手编制数据库目录以及使用 EIP 服务器配置实用程序定义服务器连接参数。该实用程序将信息（例如数据库模式名称、别名、操作系统等）复制到名为 cmbds.ini 的文件中。当启动系统管理客户机时，您可以登录的服务器的列表取自在 cmbds.ini 中定义的服务器。

要求: 您必须对每个远程数据库分别编目。每个远程数据库必须列在 cmbds.ini 文件中，才能从管理客户机连接到它。

技巧: 如果您是资深用户，则可以跳过服务器配置实用程序步骤，并在文本编辑器中修改 cmbds.ini。cmbds.ini 的缺省路径是 C:\Program Files\IBM\CMgmt。

重要事项: 如果安装该产品的人已为您想要连接的远程数据库配置了数据库目录值，则您不必为该数据库执行 DB2 CCA 步骤。但如果安装者没有输入数据库目录值或您想连接到其它远程数据库，则您必须使用 DB2CCA 并以其它数据库的连接参数修改 cmbds.ini 文件。

步骤 1 — 使用 DB2 配置助手对远程数据库编目

DB2 配置助手（CCA）对 DB2 中的远程 EIP 数据库编目。要使用 DB2CCA 对远程数据库编目，您必须知道远程服务器主机名、数据库名称和数据库实例端口号，并且必须为远程数据库定义别名。

步骤 1a - 1f 说明如何找到数据库名称、模式名称和连接端口号。您必须知道这些名称和连接端口号才能配置名称和端口号，以配置管理客户机与远程数据库之间的连接。

1. 找到远程数据库连接信息:
   a. 使用具有 DB2 管理权限的用户标识登录到远程 AIX、Windows 或 Solaris 服务器。
   b. 输入 db2 list db directory
   c. 选择要连接的管理数据库的名称。请注意安装了数据库的 DB2 实例，因为不同的实例可以有不同的连接端口号。
   d. 输入 db2 connect to <database> user <userID> using <password>
   e. 输入 db2 list tables 并注意数据库模式名称（服务器配置实用程序需要该名称）。
   f. 找到与远程管理数据库关联的连接端口号:

      在 Windows 上:

      1) 打开远程 Windows 服务器上的 DB2 控制中心。
      2) 右键单击本地机器的某个可用实例。
      3) 选择 “设置通信...”。
      4) 选择 TCP/IP 选项右边的 “属性” 按钮。端口号将列在窗口中。

      在 AIX 或 Solaris 上

      1) 输入 cd /usr/etc
      2) 输入 cat services
      3) 滚动服务的列表，直至找到远程数据库的数据库实例的连接端口号为止。例如，如果数据库安装在 db2inst1 上，则连接端口可能是 50000。
      4)
2. 使用 DB2 配置助手对远程数据库编目。更多信息请查阅 DB2CCA 帮助文件。
   a. 登录到安装了管理客户机的 Windows 服务器。您必须使用具有完全 DB2ADM 特权的用户标识登录。
   b. 从 “开始” --> “程序” 菜单浏览到 DB2 配置助手。
   c. 遵循 DB2 配置助手提示来对远程数据库的连接进行编目和测试。

d. 如果 DB2 CCA 连接测试成功，则遵循『步骤 2 — 使用服务器配置实用程序』中的步骤，或直接修改 cmbds.ini 文件，以定义存储在 cmbds.ini 中的远程数据库连接参数

步骤 2 — 使用服务器配置实用程序

服务器配置实用程序提示关于远程数据库的连接信息（端口号、主机名等），并将数据存储在 cmbds.ini 中。

1. 单击**开始-->程序-->IBM Enterprise Information Portal for Multiplatforms --> 服务器配置实用程序**。
2. 在字段中输入信息（请参阅表 4）。

表 4. 服务器配置实用程序

字段	信息	注解
服务器	选择数据库类型，或者是 Content Manager 或者是 EIP。	此处“服务器”是指数据库类型，而非安装了数据库的服务器的名称。**技巧**: 只要您的系统在同一台机器上包含 Content Manager 和 EIP 管理客户机，就可以使用管理客户机来管理这两类数据库。
服务器名称	输入您当前连接到的数据库的别名。要求: 您必须使用在 DB2CCA 中定义的相同别名。	别名提供在您的工作站上标识远程数据库的唯一名称。别名长度限制在八个字符内。例如，如果远程数据库名称是 ICMNLSDB，则别名可能是 REMOTE1。
模式名称	输入在创建远程数据库时分配的模式名称。	ICMADMIN 是 EIP 和 Content Manager 数据库的缺省模式名称。
主机名	输入安装了远程数据库的计算机的名称。	输入全限定主机名或输入安装了远程数据库的计算机的 IP 地址。
操作系统	从下拉框中选择一个操作系统。	选择 AIX、Sun Solaris 或 Windows。OS/390 选项在 EIP 8.2 中不起作用。
端口号	输入分配给远程数据库的端口号。	50000 是安装在 Windows、AIX 和 Solaris 上的 EIP 和 Content Manager 数据库的缺省连接端口号。
远程数据库名称	输入远程数据库的名称。使用大写字母。	ICMNLSDB 是 EIP 和 Content Manager 数据库的缺省名称。
节点名称	输入远程 EIP 或 Content Manager 数据库的节点名称。	节点名称是分配给远程数据库的唯一名称，与为远程数据库创建的别名相似。要查找安装在 Windows、AIX 或 Solaris 服务器上的数据库的节点名称: a. 打开 DB2 命令行会话。 b. 在 db2=> 提示符下，输入 LIST NODE DIRECTORY c. DB2 显示在远程服务器上安装或定义的所有数据库的节点名称和其它数据。
启用单一注册	如果在数据库安装期间启用了单一注册，则单击此选项。	缺省设置是未选中（禁用）。

表 4. 服务器配置实用程序 (续)

字段	信息	注解
安全性选项	如果在数据库创建期间选择了该选项，则单击“客户机认证”。	缺省设置是“服务器”。

3. 单击“确定”。

步骤 3 — 测试远程数据库连接

1. 登录到安装了管理客户机的 Windows 服务器。
2. 单击**开始-->程序-->Enterprise Information Portal for Multiplatforms 8.2-->管理**。
3. 从“服务器”字段中的下拉列表选择远程数据库别名。该名称与在服务器配置实用程序和 DB2 配置助手中定义的别名匹配。
4. 输入与远程数据库关联的用户标识和密码。
5. 单击“确定”。打开管理客户机。

定义文档类型

EIP 提供了对一些文档类型的查看器支持。如果对某服务器定义了一种文档类型，则可以在其本机应用程序中启动这种类型的文档。例如，如果将 Lotus Word Pro® 文档存储在 Content Manager OnDemand 服务器中，则可以设置 EIP 以使具有 `.lwp` 扩展名的文档在 Lotus Word Pro 而不是客户机文档查看器中启动。

要定义文档类型，修改 CMBROOT 目录（通常是 *x*:\Program Files\IBM\CMgmt）中的 `cmbcc2mime.ini` 文件。该文件包含关于如何开发定制 MIME 定义的指示信息。该文件将内容类转换为 MIME 类型的流，这样客户机可以从内容服务器读取内容。

注意事项：在启动基于 MIME 类型的应用程序时，仅显示基本对象。对文档所做的任何标记都不显示。如果文档有多个部分，则只显示第一部分。两个文件中的 MIME 类型必须匹配。

更改服务器 MIME 类型文件（cmbcc2mime.ini）

在添加服务器 MIME 类型时，请验证正在添加的文档类型是否是为该文件创建的 MIME 类型。关于更多信息，请参阅 Web 站点: `ftp://ftp.isi.edu/in-notes/iana/assignments/media-types`。

要将值添加到 `cmbcc2mime.ini` 文件，请完成以下步骤:

1. 在文本编辑器中打开 `cmbcc2mime.ini`。
2. 对用户定义的值使用以下格式:
   - 内容类从 4096 开始
   - 在内容类值之后跟等号（=）
   - 在等号之后跟 MIME 类型。如果它不是该内容类的标准 MIME 类型，则请执行以下步骤:

a. MIME 类型由一个类型和一个子类型组成。有效的类型有 application、text、image、model、message、audio 和 video。
b. 类型之后跟斜杠（/）
c. 如果要创建子类型，则用于那个文档的标记之前必须使用（x-），例如:

```
x-mydocumentclass (4096=application/x-mydocumentclass)
```

- 必要时对每个新的 MIME 类型重复 2b 和 2c。

技巧: OnDemand 内容服务器将文件扩展名而非内容类数值映射为 MIME 类型流。

第 3 章 使用 EIP 管理客户机功能部件

这一部分说明由 EIP 管理员执行的一些常见任务。

创建联合搜索

联合搜索是从客户机应用程序发出的、同时搜索一个或多个内容服务器的查询。EIP 为您提供了为联合搜索创建搜索模板的工具。由于每个内容服务器采用不同方法存储和组织信息，所以搜索模板必须考虑到每个服务器的这些差异。搜索模板将联合实体及其联合属性映射为本机属性以搜索内容服务器。

创建联合搜索涉及以下内容:

- 使用 EIP 连接器定义与内容服务器之间的连接
- 创建联合实体
  - 定义联合实体
  - 创建联合属性
  - 将联合属性映射为本机属性
  - 指定参数
- 创建搜索模板
  - 定义搜索模板
  - 定义搜索条件
  - 定义模板设置
  - 指定客户机用户的访问权限

EIP V8.2 中提供了两个向导，使得创建联合实体和搜索模板更加容易。联合实体向导包含服务器目录清单，可以对它进行过滤，使得查找本机属性比较容易。它还可为联合属性生成有效的缺省参数，从而减少了错误配置的可能性。搜索模板向导帮助您创建搜索条件。它还帮助您设计搜索条件和显示结果的外观和操作。它甚至为您提供了搜索模板出现在客户机应用程序中的可能外观的预览。另外，用于创建 EIP V7.1 的联合实体和搜索模板的对话框对于哪些喜欢它们的人来说也可用。

所有向导、对话框和字段都在 EIP 联机帮助中有所记述。

定义服务器

您必须在连接到服务器和执行服务器目录清单之前定义服务器。右键单击服务器图标并单击“新建”时，客户机将显示 EIP 支持的所有连接器。定义服务器之前，您必须了解有关连接器的一些基本信息:

- 安装程序选择哪些连接器？安装的连接器在 `cmbcs.ini` 配置文件中列出。在 Windows 服务器上，缺省路径为 *x*:\Program Files\IBM\CMgmt。请询问 AIX 或 Solaris 管理员 `cmbcs.ini` 文件的位置。
- 安装程序选择本地还是远程连接器选项？`cmbcs.ini` 文件包含本地或远程连接器类型。

- 如果系统针对 RMI 配置，RMI 服务器是否已启动？要在本地 RMI 服务器上启动 RMI，请使用**开始→程序→IBM Enterprise Information Portal for Multiplatforms 8.2→启动 RMI 服务器**。如果系统使用远程 RMI，则请在 `cmbsvclient.ini` 中查找安装 RMI 连接器的远程服务器。请咨询 RMI 服务器管理员以获取更多信息。
- 如果安装 EIP 的人员包含了 CM for AS/400 连接器，哪些信息包含在名为 `frnolint.tbl` 的网络表中？AS/400 `frnolint.tbl` 处于 `%CMBROOT%` 中。
- 如果在定义包含关系数据库的远程内容服务器（如 Content Manager V8 和 DB2、DataJoiner 和 Information Catalog），则必须从您正使用客户机的工作站对数据库编目或添加数据库。

以下列表提供了您在定义服务器时采取的一般步骤:

1. 右键单击“服务器”并选择**新建**。
2. 从列表中选择一个服务器。出现“新建服务器”窗口。
3. 在“常规”选项卡上的“服务器名称”字段中输入服务器名称和描述。对于某些服务器，您仅需输入数据库名称。对于其它服务器，您需输入安装数据库的服务器的全限定名称。
4. 如果必要，指定初始化参数。某些服务器需要初始化参数，如连接字符串和配置字符串。其它服务器仅需要数据库名称。
5. 单击“测试服务器”连接。EIP 使用您输入的用户标识和密码登录到某些服务器，从而启动管理客户机。如果服务器需要不同的用户标识和密码，EIP 会提示您输入特定于您定义的内容服务器的有效用户标识和密码。

技巧: 您也可以定义一种不是预定义服务器类型的内容服务器类型，但是必须为新的服务器类型提供 Java 或 C++ 接口类和服务器定义类。您还需要 Java 连接器运行服务器目录清单。关于添加内容服务器的指导，请参阅 *Workstation Application Programming Guide* 和联机 API 参考。

如果对内容服务器的配置不成功，请参阅 *Messages and Codes* 获取关于如何诊断各种情况或者关于接收到的错误消息的更多信息。

您还可以咨询要连接到的服务器的管理员以获得更多帮助。

定义服务器指南

本节提供了帮助执行初始服务器定义的指南。

连接到 DB2（关系）数据库

本节适用于 DB2、DataJoiner、JDBC、ODBC、信息目录和 Content Manager V7 和 V8 服务器。

- **重要信息**: 您必须在定义服务器之前对每个 DB2 数据库进行编目。可以使用 DB2 CCA 对数据库进行编目，也可以使用 DB2 命令提示符。请联系 DB2 管理员以获取更多信息。
- 在“常规”选项卡上的“服务器名称”字段中，您必须输入要连接到的数据库的名称。输入服务器名称时请使用大写字母。
- 定义 DB2、DataJoiner、JDBC、ODBC 和信息目录时，请单击“初始化参数”选项卡并输入与您连接到的数据库表相关联的模式名称，例如，`SCHEMA=ICMADMIN`。
- 定义 Content Manager V7.1 或 V8.2 服务器时，只要求您输入数据库名称。请勿更改“初始化参数”选项卡中的缺省设置。

- 定义 Content Manager V7.1 服务器时，您必须使名为 fronlint.tbl 的网络表在本地驱动器的 x:\CMBROOT 上。网络表包含 EIP 需要用于定位和登录到 Content Manager V7.1 库服务器的主机名、端口号和服务器类型信息。如果您定义了多台 Content Manager V7.1 服务器，则定义服务器之前每台服务器都必须在 frnolint.tbl 文件中有一个独立条目。
- 要连接到 DB2 DataJoiner，请确保对于 DB2 通用数据库中定义的数据库实例，Enterprise Information Portal 的认证方法定义为服务器。
- 要连接到 DataJoiner 2.1，必须从 DataJoiner Web 站点下载绑定程序并绑定 DataJoiner 数据库，然后才能定义 DataJoiner 服务器。

连接到文本搜索服务器

要定义文本搜索服务器，您必须首先定义与文本搜索服务器相关联的 Content Manager V7.1 服务器。

从下拉框将文本搜索服务器名称输入到“选择关联的 Content Manager V7.1 服务器”。此框在“关联服务器”选项卡上。

Content Manager V7.1 服务器和文本搜索服务器必须已开机且正在运行，EIP 才能连接到它们。

连接到多台 Content Manager for AS/400 服务器

如果您使用不止一台 AS/400 服务器，则必须定义网络表中的附加服务器。网络表（frnolint.tbl）位于 *x*:\<cmbroot> 中。对于新的服务器，请输入服务器名称、连接类型（例如，TCP/IP）、主机名、端口和服务器类型。对于第一台服务器，安装者将在安装期间输入服务器、主机名和端口值以创建 frnolint.tbl。

以下是存储在 fronlint.tbl 中的信息的一个典型示例:

```
/*  VI/400 Network Table */
SERVER: VI400 REMOTE TCPIP
        HOSTNAME = vi400
        PORT    = 29000
        SERVER_TYPE = FRNLS400
```

配置 Extended Search 连接器

为定义 Extended Search 服务器而输入的信息取决于以下两个因素:

- 安装 Extended 服务器的 Web 服务器类型 — Domino Web 服务器、WebSphere、IIS。
- 为安装了 Extended Search 的 Web 服务器定义的端口号。

定义 Extended Search 连接器时，请遵循以下步骤:

1. 在“常规”选项卡上的“服务器名称”字段中输入安装了 Extended Search 的 Web 服务器的全限定主机名。
2. 如果安装程序在安装 Extended Search 时对 Web 服务器端口号选择了缺省设置，则请在“初始化参数”选项卡上的“端口号”字段中输入 80。
3. 在“应用程序标识”字段中输入 Demo。按照显示的内容输入名称。
4. 在“密码”字段中输入 Demo。
5. 在“附加参数”字段中:

a. 如果您知道 Extended Search 安装在 Domino Web 服务器上，且安装程序在安装期间对 Web 服务器和 Extended Search 端口号使用缺省端口号设置，请不要更改两个分号。
b. 请参阅以下部分获取有关如何对使用定制设置安装的 Extended Search 服务器修改“附加参数”字段的信息。

如果已使用 WebSphere 应用程序服务器配置了 Extended Search 连接器，或者如果 Extended Search 端口不是 6001，或者如果 Extended Search 服务器名称不同于 Web 服务器名称，则您需要配置连接器以查找 ES servlet 的正确相对路径、ES 端口号和 ES 服务器名称。

如果已使用 WebSphere 配置了 Extended Search，或者如果 Extended Search 端口不是 6001，或者如果 Extended Search 服务器名称不同于 Web 服务器名称，则可以通过完成以下步骤来创建配置文件，例如 desclient.cfg:

将目录设置为应用程序或样本所在的目录。创建配置文件，例如，desclient.cfg。此文件不是 Enterprise Information Portal 提供的。

在 desclient.cfg 中，添加 DESHOSTNAME=（ES 主机名）

如果已使用 Domino 应用程序服务器配置了 ES，则设置 DESREQURI=/servlet/ESAdmin。如果已使用 WebSphere 应用程序服务器配置了 ES，则设置 DESREQURI=/lotuskms/ESAdmin，其中 *servlet* 是支持 Extended Search 连接器的 HTTP 服务器上的目录路径。

如果应用程序服务器是 WebSphere，则 DESREQURI 应该为 /lotuskms/ESAdmin，而不是 /servlet/ESAdmin。

如果您打算从瘦 / 胖客户机搜索 ES 源，请在管理客户机的 DES 服务器定义对话框中定义名为

"DESCFGPATH=<desclient.cfg 的绝对路径>"

的附加参数。

如果要运行 ES 样本，请在命令行参数中传递 desclient.cfg 的绝对路径。

示例 1:

```
TConnectDES es.stl.ibm.com  user password
PORT=80;DESAPPID=Demo;DESAPPPW=password;DESCFGPATH
=<desclient.cfg 的绝对路径>;
```

示例 2:

```
java TConnectDES es.stl.ibm.com  user password
PORT=80;DESAPPID=Demo;DESAPPPW=password;DESCFGPATH=
<desclient.cfg 的绝对路径>;
```

定义信息目录服务器

您必须在定义服务器之前对信息目录服务器进行编目。在“服务器名称”字段中输入服务器名称，例如，SAMPLE1。在“初始化参数”选项卡中，输入 SCHEMA=<与 SAMPLE1 关联的模式名称>。

定义 OnDemand 服务器

OnDemand 服务器和库服务器守护程序必须运行，您才能定义 OnDemand 服务器。您可以在定义 EIP 中的服务器之前 ping OnDemand 服务器，从而验证服务器和守护程序是否正在运行。

在“常规”选项卡上输入安装了 OnDemand 的服务器的全限定主机名。

在“初始化参数”选项卡中输入安装 OnDemand 服务器时指定的端口号。如果安装 OnDemand 的人员在 OnDemand 安装期间选择了缺省端口值 0，则请在端口号字段中输入 0。如果安装者选择了不同的端口号，则请输入端口号，前面带有 # 号。例如，#5000 可能是在 Windows 服务器上为 OnDemand 选择的备用端口号。

如果您在定义安装在运行版本 4 软件的 AS/400 服务器上的 OnDemand 服务器，则必须在“附加参数”字段中输入以下信息：STATECONNECT=#1。

如果您在定义安装在运行版本 2.1 软件的 AS/400 服务器上的 OnDemand 服务器，请输入在 OS/390 V2.1 服务器上安装 OnDemand 时指定的定制端口号。

OnDemand 需要套接字在连接期间保持活动。

定义文本搜索服务器

要定义文本搜索服务器，您必须先定义与文本搜索服务器相关联的 Content Manager V7 服务器。

在“常规”选项卡上的“服务器名称”字段中输入文本搜索服务器的名称。从“关联服务器”选项卡上的下拉框中选择关联的 Content Manager V7.1 服务器。

Content Manager V7.1 服务器和文本搜索服务器必须已开机且正在运行，EIP 才能连接到它们。

定义 Domino.Doc 服务器

在服务器名称字段中输入到 Domino.Doc 服务器的服务器名称和库名称的路径。例如，oakley/DominoDoc1/Lib.nsf。

如果使用本地连接器，则必须在具有 EIP 客户机的工作站上安装 Domino Doc Desktop Enabler。如果使用 RMI，则必须在 RMI 服务器上安装 Domino Doc Desktop Enabler。Domino Doc Desktop Enabler 必须与 Domino Doc 服务器版本相同。

请不要修改“初始化参数”选项卡上的两个分号。

定义 ImagePlus for OS/390 服务器

定义 ImagePlus for OS/390 服务器时，您必须获取以下参数以连接到服务器。以下列表包含样本值:

- FAF 端口号: *3061*
- FAF 应用程序标识: *01*
- FAF 协议: *4000*
- FAF IP 地址: *9.67.43.83*
- 对象分发管理器 CICS: *4000*
- 对象分发管理器 IP 地址: *9.67.43.83*

- 对象分发管理器端口号: *3082*
- 对象分发管理器终端标识: *将此字段保留为空白*
- 附加参数: *FAFSITE=CS61;;*

对 Content Manager ImagePlus for OS/390 使用跟踪

如果无法连接到 Content Manager ImagePlus for OS/390 服务器，则跟踪可以帮助您解决问题。如果安装了 Content Manager ImagePlus for OS/390 的连接器，则可以通过修改位于 cmbroot 中的 eypapi.ini 文件为 ImagePlus for OS/390 打开跟踪。

eypapi.ini 文件包含以下行:

```
; Path where the IPFAF files are stored
;    (MUST NOT have a trailing '\')
; -- default is the <ROOT Directory>\
;
IPFAFPath=d:\cmbroot
; Flag for Logging (EYPLmmdd.LOG files)
;        --  default is Logging OFF (0)
; --  0  All Logging OFF
; --  1  Log files created only error conditions logged
; --  2  Log files created all conditions logged
;
Logging = 0

   ;-------------------------------------------------------
;
; Flag for Logging the FAF Parameters Types created by APIs
;        --  default is Logging OFF (0)
;        --  0  Parameter types Not logged
;        --  1  Log Faf Parameter Types
;
FafTypeLogs = 0
```

IPFAFPath

指定写入日志的目录。日志文件名为:

EYP*mmdd*.LOG

其中 *mmdd* 是创建日志的月和日。

Logging

指定创建日志文件的时间。

0 不进行记录。缺省设置是 0。

1 创建的日志文件仅包含错误条件。

2 创建的日志文件包含所有条件。

FafTypeLogs

指定 API 创建的 FAF 参数类型的日志记录。

0 不记录参数类型; 缺省设置是 0。

1 记录 FAF 参数类型。

处理 OnDemand 连接器: TCP/IP 调节和套接字

一个已知的 Windows 问题会在连接到 OnDemand 服务器时影响性能。在 OnDemand 服务器上重复搜索和检索期间，很多 Windows 套接字会打开和关闭。两个缺省 Windows 设置会影响 EIP 和 OnDemand 服务器之间的繁重流量:

- 当应用程序关闭 Windows 套接字时，Windows 会将套接字端口置于 TIME_WAIT 状态中 240 秒；在此期间，端口不可复用。
- Windows 将应用程序可以使用的端口数限制为 5000。

为避免可能导致的问题，请使用 Windows 注册表编辑器更改超时等待时间和端口号的值。

- 将超时等待时间的值从 240 秒更改为一个较低数值（有效范围为 30-300 秒）。键的名称为 HKEY_Local_Machine\System\CurrentControlSet\services\Tcpip\Parameters\TcpTimedWaitDelay。
- 将最大端口数从其缺省值 5000 更改为一个较高数值（有效范围为 5000-65534）。键的名称为 HKEY_Local_Machine\System\CurrentControlSet\services\Tcpip\Parameters\MaxUserPort

关于 TcpTimedWaitDelay 和 MaxUserPort 的更多信息，请参考 Windows 文档。

处理 Extended Search 连接器

本节描述在 EIP V8.2 中对 Extended Search 连接器所做的更改。

通过在 DESLOCALE 键中传递语言环境值来支持用户定义的语言环境。如果直接调用 ES 连接器，您可以在命令行中传递此对值。您可以在 ES 属性的**附加**参数中设置此值。

技巧: Extended Search 服务器软件包含在 EIP V8.2 产品包装盒中。

创建联合实体

在定义了到内容服务器的连接之后，创建联合搜索的下一步是创建联合实体，该实体将成为搜索模板的构件。本节解释了联合实体以及如何使用创建联合实体向导。

了解联合实体

大多数时候客户机应用程序用户不希望在逐个服务器的基础上搜索信息，而是希望实施单一联合搜索。搜索模板允许客户机应用程序用户将其搜索绑定在单一查询中。作为 EIP 管理员，您可以创建这些搜索模板，以便在客户机应用程序中使用。创建搜索模板之前，必须先创建联合实体，该实体将其联合属性映射到内容服务器上的本机属性。

例如，DB2 将信息存储在表中，这些表的列显示存储在某个表中的信息的属性。名为 Customer_Demographics 的表可能包含诸如 Name、Pol_Number、Address、Phone 和 Occupation 等列。

另一方面，Content Manager 使用项、项类型和属性代替表和列。存储在 DB2 中的相同信息可以用实体名称 CustInfo 存储。其属性可以为 CustName、Acct、HomeAddres、HomePhone 和 Job。在这两种情况中，同样的信息以不同方式存储和标识。

EIP 解决了不得不说明内容服务器存储相同信息的所有不同方式的问题。联合实体为您保存此信息的跟踪。联合实体并不实际存储数据；它存储了关于每个内容服务器如何存储数据的元数据。创建联合实体时，您需要将它的所有属性映射到希望查询的内容服务器上的相应本机属性。

对上面给出的示例，您可以创建名为 Policy_Info 的联合实体，它具有联合属性 Policy_Name、Policy_Number、Home_Address 和 Job_Title。然后就可以将联合属性映射到每个相应的本机属性了。

EIP 可以生成包含此信息的服务器目录清单。服务器目录清单包含此信息，且创建联合实体向导允许您获取可以在内容服务器上过滤的服务器目录清单。仅在使用“创建联合实体”向导时提供过滤功能。如果使用手工（非向导）方法创建联合实体，则没有过滤选项。一旦生成服务器目录清单，您就可以开始将联合属性映射到本机属性了。

将联合属性映射到本机属性还远远不够。每个本机属性还可以具有不同特性。属性可以为: （1）可空、（2）可查询、（3）可更新以及（4）可文本搜索。根据选择的数据类型，您还可拥有关于数据长度、精度、范围、最小和最大值的选项。

定义这些特性时，您无法使它们较已经由映射为联合属性的本机属性定义的特性更加严格。向导提供了满足此条件的缺省特性。如果定义了缺省联合属性的特性之后希望还原到向导建议的缺省特性，您仍可选择缺省设置。

总之，联合属性被映射为多个内容服务器上的相应本机属性。每个联合属性的特性都包含本机属性的所有特性。一旦创建联合实体，您就拥有了到达存储在不同内容服务器上信息的路径。此后您就可以使用联合实体创建用于特殊查询的搜索模板。

使用创建联合实体向导

“创建联合实体”向导是 EIP V8.2 中的新内容。虽然您可以使用与 EIP V7.1 和更早版本中相同的对话框创建联合实体，但是该向导使您更易于创建联合实体。

要使用向导创建联合实体，请遵循下面列出的步骤:

1. **定义联合实体**命名并描述联合实体。您还可以决定是否希望该联合实体是可文本搜索的。
2. **定义联合属性**命名并修改联合属性。
3. **映射联合属性**将联合属性映射到本机属性。此处提供了一些工具，可以用来获取服务器目录清单，选择希望映射的本机属性以及在以后修改您的映射。
4. **定义特性**定义每个联合属性的特性。可以定制特性，也可以接受缺省设置。
5. **确认联合实体**复查已为联合实体选择的设置。可以返回上一面板修改设置。完成时请单击**完成**。

这些步骤与使用向导相对应。请参阅 EIP 联机帮助获取关于如何使用该向导的更多信息。

创建联合文本索引

文本搜索引擎可以与 Content Manager V7.1 和更早的内容服务器集成，所以您可以自动建立索引、搜索和检索存储在 Content Manager 中的文本信息。用户可以通过搜索单词或词组查找文档。文本搜索服务器既支持单字节字符集又支持双字节字符集。

如果将 Content Manager V7.1 和更早版本的服务器与文本搜索引擎一起使用，则可以创建联合文本索引。然后可将联合文本索引映射到 Content Manager 文本搜索服务器上的 Content Manager 文本搜索索引。

创建联合文本搜索索引时，您可为组合搜索启用该索引，即同时搜索本机文本索引和本机属性。当为组合搜索启用联合文本索引时，您也可以将那个索引映射到联合实体。然后您就可以把由联合实体映射的本机属性与其联合属性一起，映射到文本搜索服务器上的本机文本搜索索引。

创建搜索模板

创建联合实体之后，您就可以创建搜索模板了。请记住，搜索模板使用联合实体作为对内容存储位置的映射。创建搜索模板时，您仍必须定义希望搜索的内容、希望对搜索结果执行的操作以及具有使用模板的许可权的人员。虽然对每个模板只能使用一个联合实体，但是多个模板可以使用一个联合实体。也可以将联合实体的属性的任意组合作为搜索条件进行搜索。要创建搜索模板，请使用搜索模板向导完成这些步骤:

1. 定义搜索模板
2. 定义搜索条件
3. 确定搜索设置
4. 指定访问特权

这些步骤对应于搜索模板向导中的步骤。请参阅 EIP 联机帮助获取完成搜索模板创建过程的详细信息。

定义搜索模板

启动向导之后，它会提示您定义搜索模板。请作好以下准备:

- 提供搜索模板的名称和描述
- 选择搜索模板的联合实体。**限制**: 每个搜索模板仅可使用一个联合实体。
- 选择联合文本索引（如果可用）

技巧: 联合文本索引的复选框仅在您使用 Content Manager V7.1 和更早版本的文本搜索引擎时适用。如果对文本搜索使用 DB2 TIE，则它为参数搜索并且可以象在搜索模板中那样进行配置。

定义搜索条件

定义搜索模板之后，该向导会提示您执行以下操作:

1. 选择一种搜索类型，属性或文档。文档仅当在先前步骤中选择了联合文本索引时可用。
2. 命名搜索条件
3. 选择联合属性
4. 选择可用运算符
5. 提供缺省搜索字符串（仅适用于文档搜索）

该向导提供了一个下拉菜单，它列出了与选定联合实体关联的所有实体属性。这些属性将成为搜索模板的搜索条件。向导还提供可用运算符的列表。

技巧: 您可以对每个模板创建多个搜索条件，也可以从模板删除现有条件。

确定搜索设置

此面板允许您定义缺省搜索设置、条件设置以及显示值设置。这些设置中的每一个都具有缺省值，您可以修改它。要修改这些设置，请单击每个设置的相应按钮。

“缺省设置”窗口使您能够:

- 控制当客户机应用程序用户希望使用搜索模板时服务器不可用将发生的情况
- 定义参数搜索的通配符
- 指定保存搜索结果的文件夹的名称
- 选择搜索必须使用所有（AND）还是任意（OR）条件

“条件设置”窗口允许您控制搜索条件的次序、结果显示列的次序、列头以及列宽。

“显示值设置”窗口为您提供定义搜索结果的显示值的途径。例如，如果值 Weekday 在一台服务器上为 Monday，而在另一台服务器上为 Mon.，则您可以指定使用 Monday 作为搜索结果显示值。

指定特权

除了定义查找位置（使用联合实体）、查找内容（搜索条件）、如何显示结果（设置）之外，您还必须定义具有对搜索模板的访问权的人员。

搜索模板向导的“指定特权”窗口提供了将对模板的访问权指定给现有用户或用户组的工具。

给一个用户指定搜索模板的访问特权不会把映射到该模板的内容服务器的访问权限授予此用户。用户必须满足每个独立内容服务器的安全性需求。在将对搜索模板的访问权指定给用户之前，您必须使用访问控制表以及用户管理确保用户具有适当的特权。

使用向导搜索用户或用户组时，EIP 仅返回对请求的内容服务器有适当访问权的用户。

第 4 章 管理用户访问

用户不能在没有用户标识、密码或特权集的情况下访问 EIP 系统。在创建用户并为它们指定特权之前，必须确定哪些用户具有对系统的访问权以及他们的工作需要执行哪些操作。您不希望用户在不了解删除某对象的后果时就有权删除该对象。另一方面，您也不会希望用户因为未被赋予正确特权而无法进行他们的工作。因此，您需要在为用户指定特权之前确定每种工作所需的任务类型。

用户在 EIP 系统中创建对象时，必须定义其他用户将对该对象具有的访问权。创建对象的用户必须定义可以访问该对象的人员以及可以对该对象执行的操作。此定义在 EIP 系统中称为访问控制表，或 ACL。

创建用户标识和密码

如果您希望在系统管理客户机中定义的用户标识也用于 DB2 认证，该用户标识必须遵循 DB2 命名规则。DB2 命名规则对您希望用于超级管理员或连接用户标识的用户标识适用。您不能使用以下词:

- USERS
- ADMINS
- GUESTS
- PUBLIC
- LOCAL
- 在 SQL Reference 中列出的任何 SQL 保留词。

用户标识不能以下列字符开头:

- SQL
- SYS
- IBM

您可以使用以下字符:

- A 至 Z **限制**: 一些操作系统允许区分大小写的用户标识和密码。在您的操作系统文档中查看它是否允许区分大小写。
- 0 至 9
- #
- $

限制: 用户标识不能超过 30 个字符。

了解 DB2 管理权限

当登录到系统管理客户机时，您有两种级别的认证: 一种是数据库级别，另一种是产品级别。当您启用管理域功能时，管理员分为两类: 超级管理员和次级管理员。通常，只有超级管理员有权访问系统管理客户机。

超级管理员必须有 DB2 特权: db2admin 特权，也就是需要对 DB2 的完全管理特权。必须使用 db2admin 特权在操作系统中定义此用户标识。此操作系统标识的密码用于连接到 DB2 以及登录到库服务器。不使用为库服务器定义的密码。Content Manager 特权: 使用完全 Content Manager 管理特权（ “AllPrivs” ）在库服务器中定义此用户标识，以执行所有管理活动。

次级管理员不要求 DB2 特权。次级管理员仅管理库服务器的某些部分，因此次级管理员以下面两种方法中的一种登录到系统管理客户机:

- 如果用户标识是一个操作系统用户标识，那么操作系统中的密码用于连接到 DB2 以及登录到库服务器。
- 如果用户标识不是操作系统用户标识，那么 cmbfedenv.ini（对于 Enterprise Information Portal）或 cmbicmenv.ini（对于 Content Manager）中加密的用户标识和密码对就用于连接到 DB2，而“登录”窗口中提供的用户标识和密码用于登录到库服务器。

关于登录到库服务器的更多信息，请参阅下一节。

次级管理员也需要 EIP 特权。他们需要域管理特权用于所有子域管理活动。

使用 INI 文件连接到 DB2

INI 文件中的每个条目都包含库服务器名称和用于连接到 DB2 的一对加密的用户标识和密码。此加密的用户标识（称为连接用户标识）和密码是在安装产品时定义的。连接用户标识必须区别于系统管理员的用户标识。Enterprise Information Portal 使用 cmbfedenv.ini 连接到 DB2，而 Content Manager 使用 cmbicmenv.ini 来连接。缺省连接用户标识是 ICMCONCT。在安装期间，库服务器和资源管理器的密码包含在三处: cmbicmenv.ini 文件包含用于访问库服务器的用户标识和密码。操作系统定义对库服务器和资源管理器所驻留的数据库的访问权。ICMRM.properties 文件包含资源管理器用户标识和密码

如果使用了 INI 文件，也就是用户标识不是操作系统用户标识，那么 INI 文件中的用户标识和连接用户标识都必须存在于库服务器中。

连接用户标识必须定义在库服务器和操作系统中。它要求有 UserDB2Connect 特权。要更改 INI 文件中的连接用户标识和密码，请从管理客户机窗口选择**工具 --> 更改数据库标识 / 密码**。

更改资源管理器的库服务器和系统管理员密码

如果需要更改资源管理器的密码，那么需要更改库服务器用于登录资源管理器的密码以及资源管理器的系统管理员密码。**重要事项**: 当更改库服务器和系统管理员用于登录到资源管理器的密码时，请按顺序完成以下步骤:

1. 登录到系统管理客户机。
2. 展开“资源管理器”树。
3. 单击想要修改的资源管理器并展开其树。
4. 单击“服务器定义”并选择“属性”。打开“服务器面板”窗口。
5. 更改“密码”字段中的密码。
6. 单击“确定”。
7. 右键单击展开的资源管理器（从步骤 2）并选择“属性”。打开“资源管理器属性”窗口。

8. 更改“密码”字段中的密码并单击“确定”。

更改数据库访问密码

如果需要更改数据库访问密码，则需要更改数据库连接的操作系统密码和 ICMRM.properties 文件，这样资源管理器可以标识新密码。

要更改数据库连接的操作系统密码，则执行以下步骤:

1. 根据您的操作系统，浏览至用户和密码实用程序。
2. 单击 ICMRM。
3. 选择“设置密码”。
4. 输入新密码。

要更改 ICMRM.properties 文件，请完成以下步骤:

1. 打开 ICMRM.properties 文件。缺省位置是: *X*:\WebSphere\AppServer\installedApps\icmrm.ear\icmrm.war\WEB-INF\classes\com\ibm\mm\icmrm\ICMRM.properties，其中 *X* 是安装 Content Manager 的驱动器的位置。
2. 更改 DBPassword 以匹配操作系统密码。
3. 保存 ICMRM.properties 文件。

更改了数据库密码以后，数据库需要重新启动，或者让它发出两个或三个错误，直到它自己复位为止。

有关在系统管理客户机中更改资源管理器的密码和其它字段的详细指示信息，请参阅系统管理联机帮助。

从 LDAP 导入用户

LDAP 支持按企业级别而不是逐个系统地管理用户的标识和密码。 EIP 使用三种 LDAP 技术: IBM Directory（在以前的版本中称为 IBM SecureWay Directory）、Windows 2000 Active Directory 和 Lotus Domino Directory Notes 通讯录（NAB）。用户密码驻留在 LDAP 服务器上。当用户登录到 或 Enterprise Information Portal 时，由 EIP 数据库中的用户概要文件认证用户标识和密码并检查用户标识的具体特权。安装 EIP 期间，LDAP 可能已经启用。如果 LDAP 未在安装期间启用，您可随时激活它。

要启用 LDAP，请选择 **开始 ⟶ 程序 ⟶ EIP for Mulitplatforms ⟶ LDAP 用户标识导入调度程序**，然后启动系统管理客户机。打开“LDAP 配置”窗口（工具 --> LDAP 配置）。在“服务器”页面上选择“启用 LDAP 用户导入和认证”复选框，并提供 LDAP 服务器信息。

启用了 LDAP 之后，可以通过在“新用户”窗口中单击 LDAP 按钮来导入用户。这可允许用户从 LDAP 服务器有选择地导入到 EIP。另一种方法是，可使用 LDAP 用户标识导入调度实用程序成组地导入用户。在登录期间，库服务器自动连接到 LDAP 服务器以认证用户。如果 LDAP 服务器出于任何原因不能验证用户密码，则认证失败。

您可以通过进入主系统管理客户机窗口并单击**工具 -> LDAP 配置**来修改 LDAP 服务器配置。还可以通过从 EIP 上的“开始”菜单进入 LDAP 用户注册表导入实用程序，来更改当前 LDAP 服务器。关于规划 LDAP 的信息，请参阅 *Planning and Installing*

Your Content Management System。关于如何在系统管理窗口中配置 LDAP 服务器信息的信息，请参阅系统管理客户机联机帮助。

关于规划 LDAP 的信息，请参阅 *Planning and Installing Your Content Management System*。关于如何实现 LDAP 的信息，请参阅系统管理客户机联机帮助。

介绍特权

管理客户机提供特权组、特权集和单个特权。如果管理 Content Manager/EIP 组合的系统，则特权对于客户机的这两部分是公用的。构建到客户机的特权可以帮助您理顺

特权组 特权组是用户任务的集合，目的是帮助管理员在“特权集”对话框中创建新的特权集或用户角色。

特权集 特权集是一个用户角色集合。

特权 特权代表用户操作。例如，

示例 1 — 特权: 您希望将特权 ClientScan 和 ClientImport 分配给一组用户，这些用户通常仅将客户机用于把文档扫描和导入到 Content Manager。如果您有多个通常执行该任务的用户，则将创建一个用户标识（例如，user1）。然后将特权 ClientScan 和 ClientImport 与用户标识 User1 相关联。然后将 User1 分配给名为 Group1 的组。当任何输入 user1 的最终用户登录到他们的客户机并访问 Content Manager 时，该用户将仅能够扫描和导入文档。

示例 2 — 特权组: 您有一组资深的最终用户，他们需要有访问所有典型客户机任务的特权。您将创建一个用户标识（例如，user2）。接着将 user2 分配给某个组（例如 group2）。随后将名为 ClientTaskAll 的特权组关联到 user2。当任何输入 user2 的最终用户登录到他们的客户机并访问 Content Manager 时，该用户将能够执行名为 ClientTaskAll 的特权组中包含的所有任务。

示例 3 — 特权集: 您有一组需要只读权限的用户。您将创建一个用户标识（例如，user3）。接着将 user3 分配给某个组（例如 group3）。然后将名为 ClientUserReadOnly 的特权集关联到 user3。当任何输入 user3 的最终用户登录到他们的客户机并访问 Content Manager 时，该用户将仅能够执行名为 ClientUserReadOnly 的特权集中包含的任务。

创建特权集

规划 EIP 系统配置时，您还必须确定哪些人将具有对系统的访问权以及这些用户将对系统上的对象具有什么程度的访问权。EIP 系统通过特权定义访问权。

特权以特定方式授予访问特定对象的权限。特权包括如创建、删除和选择存储在系统中的对象等的权限。指定给用户的一组特权就是特权集。

您管理访问权限的第一个任务是为用户创建特权集。*特权集*标识了用户可以执行的任务或操作。特权集将特权结合起来并根据某些用户类型进行调整。例如，您可能希望一组管理员管理文档转送服务器，另一组管理员管理域。管理员登录时，EIP 会检查管理员的特权集。

系统管理客户机有很多预定义的特权，您可以将其一同分组到一个特权集中。然后指定为个人用户创建的特权集。您不能将特权集指定给用户组。

创建特权组

特权组类似用户的用户组。您可以创建特权组以将相似特权放在一起，从而便于查找您想要包含在特权集中的特权。例如，如果您将两个特权指定给了系统中几乎每个用户，则您不用在每次创建特权集时搜索很多特权，而可以将这两个基本特权分组至一个称为 BasicPrivs 的特权组。

将特权集指定给用户

系统管理客户机有很多预定义的特权，您可以将其一同分组到一个特权集中。然后指定为单独用户创建的特权集。您不能将特权集指定给用户组。

您可以创建特权名称，但不能创建特权集本身。您需要与系统程序员合作，创建尚未定义到系统管理客户机的任何特权。

您可以使用 EIP 所带的特权集，也可以创建自己的特权集。

为用户指定授权特权集

为了防止用户创建具有比他们自己更多特权的用户标识， EIP 采用了授权特权集。当您给用户标识指定授权特权集时，您赋予他们在其获得授权的特权范围内创建用户标识的权限。例如，您可以赋予某个用户标识一组系统管理特权以管理域。但可能想要确保该用户标识不拥有创建用户的特权。所以当您创建此用户标识时，您应当在授权特权集字段中选择 “Noprivs”。实际上，用户标识可以管理域，但不能为该域创建用户。

将用户指定到资源管理器

为允许用户访问特定资源管理器，您需要将资源管理器指定到一个该用户可以访问的域。关于将资源管理器指定到域的更多信息，请参阅第 31 页的『将资源管理器指定到域』。

将用户指定到集合

为允许用户访问集合，您需要将资源管理器上的集合指定到一个该用户可以访问的域。关于将集合指定到域的更多信息，请参阅第 32 页的『将集合指定到域』。

创建用户组

通常，具有相同工作描述的用户也具有相同或相似的任务，并因此具有对系统中对象的相同访问权。您可以将具有公共访问需要的用户一同分组到一个用户组。不能嵌套用户组。

用户组只不过是具有相似任务的个人用户的方便分组。不要为用户组指定特权集。用户组中的每个用户都具有他或她自己的特权集。用户组使得为系统中的对象创建访问控制表变得更加容易。

如果启用了域，则请在将用户标识指定到组之前，检查用户组是在特定域中还是 PUBLIC 域中（请参阅第 31 页的『管理域』获取关于域的更多信息）。请确保用户组处于您希望用户标识所处于的域中。如果要明确为某个域创建用户标识，则可以在 “用户组” 窗口中单击**新建用户**。然后可以将创建的用户添加到用户组，并确保该用户处于同一域中。

创建访问控制表

您为用户提供了他们完成任务所需的特权。在个别基础上，对象会有某些访问控制问题。

访问控制表（ACL）是由一个或多个个人用户标识或用户组及其关联特权组成的列表。您可用 ACL 控制用户对 EIP 系统中对象的访问。可与访问控制表关联的对象有:用户存储的数据对象、项类型和项类型子集、工作列表和流程。

特权集定义个人用户使用系统的最大能力，ACL 限制个人用户对对象的访问权。具有非用户特权集定义的特权的 ACL 不能授予用户该特权。只有具有该特权的用户可以在对象上使用它。ACL 将限制用户访问，它不会授权更多访问。访问控制表在管理系统时提供了另一级别的安全性。

将特权集指定到访问控制表

您添加到访问控制表（ACL）的每个用户标识都需要与之关联的特权集。用户标识和特权集定义了哪些用户具有对对象的访问权以及他们对对象具有何种访问权。

用户不能访问任何对象，除非他们处于 ACL 中。要将用户或用户组添加到 ACL，您就需要选择 ACL 的用户标识和特权集并单击**添加**。对于每个已定义的 ACL，您会发现用户标识和组列在“访问控制表”窗口中。您可以通过添加或除去用户标识和组来修改此表。关于创建和修改 ACL 的更多信息，请参阅系统管理客户机联机帮助。

创建域

域是管理数据库中由一个或多个管理员管理的部分。域由用户标识、用户组、访问控制表、特权集、访问控制表、资源管理器和 SMS 集合组成。域对于用户来说不可见，所以您如何命名域仅对您和管理它们的系统管理员有意义。用户不了解您把他们限制到了管理数据库的一部分，也就是说他们只了解该域中的项。

域限制对管理数据库的子部分的管理和用户访问。具有管理数据库全部特权的管理员可以将受限的管理特权赋予另一个管理员。拥有全部特权的管理员，即超级管理员，可以访问管理数据库的所有部分，而拥有有限的特权的管理员，即子管理员，只可访问管理数据库的某个部分。

域限制了子管理员对访问控制表（ACL）的访问权。只有超级管理员才能创建子管理员可用于添加或删除用户标识和用户组的 ACL。子管理员不能创建、更新或删除 ACL。

子管理员可以共享不同组合的超级管理员责任，但只能针对他们自己的域。通过创建域并指定管理那些域的管理员，超级管理员可以委派子任务，同时全神贯注于整个系统并对其进行有效管理，而子管理员管理特定于他们的域的用户和任务。

启用域之前，请考虑以下条件:

- 不能禁用域
- 资源管理器、集合、用户标识和用户组一次仅可存在于一个域中。
- 特权集和访问控制表可以一次存在于多个域中。
- 除 PUBLIC（共享）域之外，域不重叠
- 任何在超级管理域中创建的对象都不能被移动，不论它是系统生成的还是用户创建的。

要启用域，请转至文件菜单，选择**工具 → 管理域**，然后选择**启用管理域**。您需要重新启动系统管理客户机以使这些域生效。关于如何配置域的管理数据库的特定指导，请参阅系统管理客户机联机帮助。

管理域

根据特权集，您管理整个管理数据库或管理特定的域。拥有对管理数据库全面访问权的管理员是超级管理员。子管理员拥有对特定域中的对象的全面访问权。

每种类型的管理员具有创建、检索、更新和删除他们的域中的对象（包括用户和集合）的能力。子管理员仅可查看和检索他们的域中的对象，并列出或检索 PUBLIC 或共享域中对象。

访问域

子管理员不能更改对象的域。但是他们可以访问自己域的内容并列出或检索 PUBLIC 或共享域中的任何对象。

超级管理员具有对管理数据库中所有域的访问权。他们可以创建对象并将其指定到域。对于某些对象（如特权集和 ACL），只有他们才能进行创建以供子管理员使用。

子管理员只能对他们域中的任何对象执行创建、检索、更新和删除（CRUD）。

将用户指定到域

创建用户标识时，您可以选择将其指定到域，也可以将其保留在缺省域中。您可以以后通过用户特性来更改用户标识的域。

用户标识每次只能访问一个域。您不能将用户添加到 PUBLIC 或共享域。

只有超级管理员具有创建域以及将用户指定到那些域的权限。一个域可以具有不止一个子管理员，但只有超级管理员可以来定义那些管理员是谁，方法是赋予他们特权集中的系统管理员特权。“新建用户”或“用户特性”窗口中的**授权特权集**字段将指示子管理员具有域中的哪些管理特权。

将用户组指定到域

将用户组指定到域将更改为该用户组中每个用户标识指定的域。用户标识每次只能访问一个域。因此，您指定的组中包含的所有用户标识也将移动到新域。

用户组名称不可每次仅可处于一个域中。您可以将用户组指定到 PUBLIC 或共享域。

将特权集指定到域

您添加到域的所有用户标识都必须具有关联的特权集。如果您不包括关联特权集，则用户无法执行他们的任务。存储特权集并使特权集对于任何用户都是可用的最好地方是 PUBLIC 或共享域。

将资源管理器指定到域

您可以通过将用户指定到特定域来限制他们对某些资源管理器的访问。定义新资源管理器供 管理数据库访问时，您拥有一个选择域的选项。

所有资源管理器的缺省值是 PUBLIC。如果您不希望每个人都能够访问资源管理器，您需要给其指定一个域。如果看不到可将资源管理器指定到的域，则您仍可定义资源管理器，然后创建所需的域。定义适当的域之前，请打开资源管理器特性并选择域。

将集合指定到域

您可以通过将用户指定到特定域来限制他们对资源管理器上某个集合的访问。如果资源管理器处于 PUBLIC 域中，您就可以将集合指定到任何其它已定义的域。但如果该资源管理器已定义给特定的域，那么您就不能将该集合指定给另一个域，即使要将集合指定给 PUBLIC 域也不行。

用户需要具有对资源管理器的访问权，以访问它上面的集合，因此在没有对它上面的集合部署相同限制的情况下，您不能限制对资源管理器的访问。

将用户从一个域移动导另一个域

您可能想要从一个域中除去特定用户并将他们添加到另一个域。请考虑使用“用户定义”窗口中的**描述**字段，用此方法来记住用户被分到哪些用户组中。这可能使得此任务变得较为容易。

重要信息：此任务将非常花费时间，并且如果您没有正确地操作，会导致访问系统发生问题。必须是超级管理员才能更改用户的域。

请仔细遵循这些步骤:

1. 查找出该用户所属的所有组。
2. 对于该用户所属的所有组，请将这些组移动到 PUBLIC 域，或者从所有组中除去该用户。
3. 将任何与此用户相关联的资源管理器移动到 PUBLIC 域，然后再移动每个移动到目标域的资源管理器所有集合。
4. 创建，*而不是移动*所有与目标域中的用户相关联的特权集（如果它们尚未在目标域中）。
5. 创建，*而不是移动*所有与此用户相关联的访问控制表（如果它们尚未在目标域中）。
6. 通过打开用户的“特性”并更改用户的域来将用户移动到目标域。
7. **可选**：您可以将在步骤 1、2 和 3 中移动的组和资源管理器从 PUBLIC 域移动到目标域；但是，只有当源域中不再有与所移动的组和资源管理器相关联的用户时，您才可以这么做。否则，组和资源管理器需要留在 PUBLIC 域中，以使不同域中的用户可以共享它们。

提醒：任何时候用户都不可以在 PUBLIC 域中。用户无法被共享。

将用户组从一个域移动到另一个域

重要信息：如果您没有正确地操作，此任务可能导致系统访问出问题。必须是超级管理员才能更改用户组的域。

遵循这些步骤，将用户组移动到不同的域:

- 如果用户组是空的，请从它的当前域中删除该用户组，然后重新创建组并将它指定给目标域。

- 如果用户组不是空的，请遵循以下步骤:
  1. 查找到属于此组的所有用户。
  2. 将组从它的当前域中删除，这会删除所有用户。
  3. 重新创建组并将它指定给目标域。
  4. 将所有这些用户添加到这个新创建的组中。

将资源管理器从一个域移动到另一个域

必须是超级管理员才能更改资源管理器的域。

要将资源管理器移动到另一个域，请遵循这些步骤:

- 如果资源管理器不包含集合，请将资源管理器移动到目标域，方法是：打开它的特性并将域更改为目标域。
- 如果资源管理器包含集合，请遵循这些步骤:
  1. 将资源管理器移动到 PUBLIC 域。
  2. 通过打开“特性”并选择目标域来将集合移动到目标域。
  3. 通过打开“特性”并选择目标域来将资源管理器移动到目标域。

将集合从一个域移动到另一个域

必须是超级管理员才能更改集合的域。

要将集合从一个域移动到另一个域，请遵循这些步骤:

1. 查找出集合所属的资源管理器。
2. 将相关联的资源管理器移动到 PUBLIC 域。
3. 通过打开“特性”并选择目标域将该集合移动到目标域。
4. 通过打开“特性”并选择目标域来将资源管理器移动到目标域。

将特权集从一个域移动到另一个域

因为特权集可以驻留在多个域中，所以您可以将它们添加到目标域而不需移动它们。

将访问控制表从一个域移动到另一个域

因为访问控制表可以驻留在多个域中，所以您可以将它们添加到目标域而不需移动它们。

第 5 章 管理信息发掘

本节以 Information Mining 是什么以及可以在商务环境中如何使用 Information Mining 开始。然后是关于 Information Mining 第一步、Information Mining 中的概念和 Information Structuring Tool 的各节，最后是关于性能调整的备注。

什么是 Information Mining?

Enterprise Information Portal Information Mining 是一种关键技术，它通过从许多方面自动化信息抽取和分析来帮助公司以较低成本为用户提供相关信息的方便访问权。

Information Mining 中的第一个难题是如何方便地访问未结构化文本中的信息以由计算机来使用。使用目前的技术，仍无法完全解释用非严格的自然语言描述的仅是事实的知识。但是，采用了模式识别技术和启发式搜索的工具已经能够从任意自由文本中抽取有价值的信息。抽取的信息范围从标识所谓的重要词语（如文档中提到的名称、机构或地方）到文档的整个摘要。

而 Information Mining 不仅仅能够从单个文档中抽取少量信息，它还具有更多功能。处理大量文档集合时，信息发掘开始起作用。“发掘”隐喻用于知识发现过程（即从单一文档中标识和抽取编码的信息、将此信息存储为元数据），以及对跨越文件集合以检测感兴趣的现象、模式或趋势而分布这些功能的分析过程。

Enterprise Information Portal Information Mining 服务

Enterprise Information Portal Information Mining 服务提供了一个基础结构，用于创建与维护与个别文档或文档集合相关的信息。此信息称为元数据。表现文档内容并存储为元数据的信息的示例包括:

- 标题
- 摘要或总结
- 名称、术语或表达式
- 文档所属类别

Information Mining 与将文档与元数据相关联的传统元数据存储器的不同之处在于 Information Mining 提供了自动创建元数据（即使它不是明确可用的）的功能。发掘和检索算法可以访问大量文档集合中的相关信息，它使用元数据指导检索过程或者对元数据运行统计模型以查找在查看集合的个别文档时可能不明显的文档间感兴趣的关系。

由于发掘和检索操作在定义良好的元数据集上进行处理而不考虑原始文档内容，所以这些过程的速度可以通过将元数据保持在所谓 *Information Mining 数据存储器*的专用存储器中来显著加快。这显著加速了对元数据的访问，因为应用程序可以从单一资源库存取元数据且不需要返回任意远程内容服务器。由于元数据对检索和导航至关重要且常用于限制搜索结果，所以很多步骤都可以在不返回内容服务器的情况下执行。

使用元数据存储器的另一个优点是它将与文档关联的数据和实际文档分开保存。明显地，对于仅可用于读访问的文档（如 Web 上的外国文档），在同一资源库中保存内容和元数据根本构不成一个选项。

Enterprise Information Portal Information Mining 服务提供用于自动创建元数据的以下机制，名为:

- **归类** — 根据用户定义的分类法，给文档指定一个或多个类别。归类组件包含了一个名为 *Information Structuring Tool* 的应用程序，这个程序提供了用于创建和维护分类法的图形用户界面。
- **摘要** — 抽取文档中最重要的语句，从而可以帮助用户决定是否要阅读整个文档。用户可以指定目标摘要的长度并以此方式直接影响抽取的元数据的复杂性和文档中信息量之间的平衡。
- **语言标识** — 确定编写文档的语言。这是应用其它 Information Mining 服务之前有用的预处理步骤。
- **信息抽取** — 自动识别文本文档中重要的词汇项，如名称、术语和表达式。
- **群集** — 将文档集划分为相似的组或群集。群集是从文档集合自动派生的。

Information Mining 服务的组成部分

Information Mining 应用程序通常包括以下任务:

1. 以可以浏览和导航的方式组织数据
2. 访问完全不同的数据源
3. 使用高级搜索操作来过滤出预测或趋势所需的数据

图 1 说明了这些 Information Mining 任务。

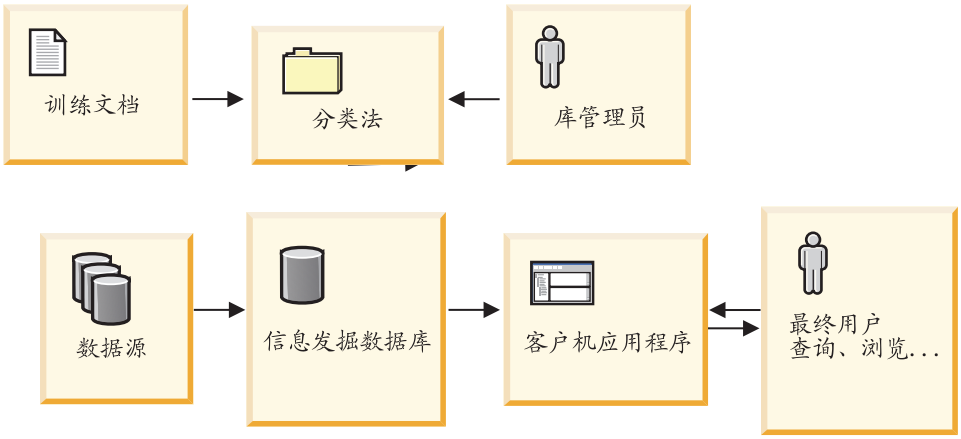

图 1. Information Mining 任务

要利用 Information Mining 功能，必须组织文档使它们可以被浏览和导航。此任务典型地由库管理员或知识工程师执行。库管理员使用 Information Structuring Tool（IST）对将被进一步利用的文档中的数据定义分类法、分层主题特性描述。IST 是一个带图形用户界面的应用程序，允许您创建和维护分类法。类别已经训练，只要分类法是稳定的，就可以访问不同的数据源。

文档可从任何内容服务器导入，或使用网上搜寻器服务（仅作为 JavaBeans 时可用）从 Web 导入，并且可以分配给类别。

随后应用于文档内容的文本分析和元数据创建功能可用作非可视的 JavaBeans 级别上的编程接口和 Java 服务 API。

Information Mining JavaBeans 是用于快速应用程序开发的软件组件，它们与 JavaBeans 一致。Java 服务 API 包含全部 Information Mining 功能，它们作为创建应用程序的各个构件。提供基于 JavaBeans 的样本和样本 JSP 以支持应用程序的创建。

标识文档内容是处理文档的所有 Information Mining 操作的先决条件。此任务的子步骤包括:

1. 标识文档代码页
2. 标识需要处理的文本部分，也就是说，忽略标记信息或二进制数据，如图像。

由于内容服务器中的文档可以任意组织，Information Mining 服务提供写入特定模块的方法，这些模块从文档格式中标识和抽取有关文本部分。在 Information Mining 服务中还提供一个缺省模块，它涵盖范围广泛的常用文档格式。关于受支持格式的列表，请参阅第 105 页的第 9 章，『文档格式』；关于如何使用样本缺省模块的详细信息，请参阅 *Application Programming Guide*。

为每个选定文档创建元数据会涉及: 处理文档内容和应用统计方法或基于知识资源的启发式方法，例如，字典或频率简要表。

Information Mining API 支持以下操作:
- 摘要
- 分类
- 语言标识
- 信息抽取
- 群集

所有文档的已创建元数据存储在 Information Mining 数据存储器中。

一旦数据存储器被填满了，则另一个用于执行文档选择的选项将变为可用的，即基于此数据存储器中信息的文档选择。高级搜索操作结合了带类别的正文查询，这样就将搜索文档限制于属于某类别。

在商务环境中使用 Information Mining

支持任何 Information Mining 技术实现的组织的基础结构通常至少由以下角色组成:
- 普遍意义上的 IT 系统管理员，不必限制于 Information Mining
- 应用程序员
- 库管理员或知识工程师
- 处理 Information Mining 应用程序的人员（最终用户）

根据应用程序的性质，您还可以找到以上列出的角色之外的更多特定角色:
- Web 设计者
- 建筑师或顾问

第 38 页的图 2 说明了这些 Information Mining 角色和操作。

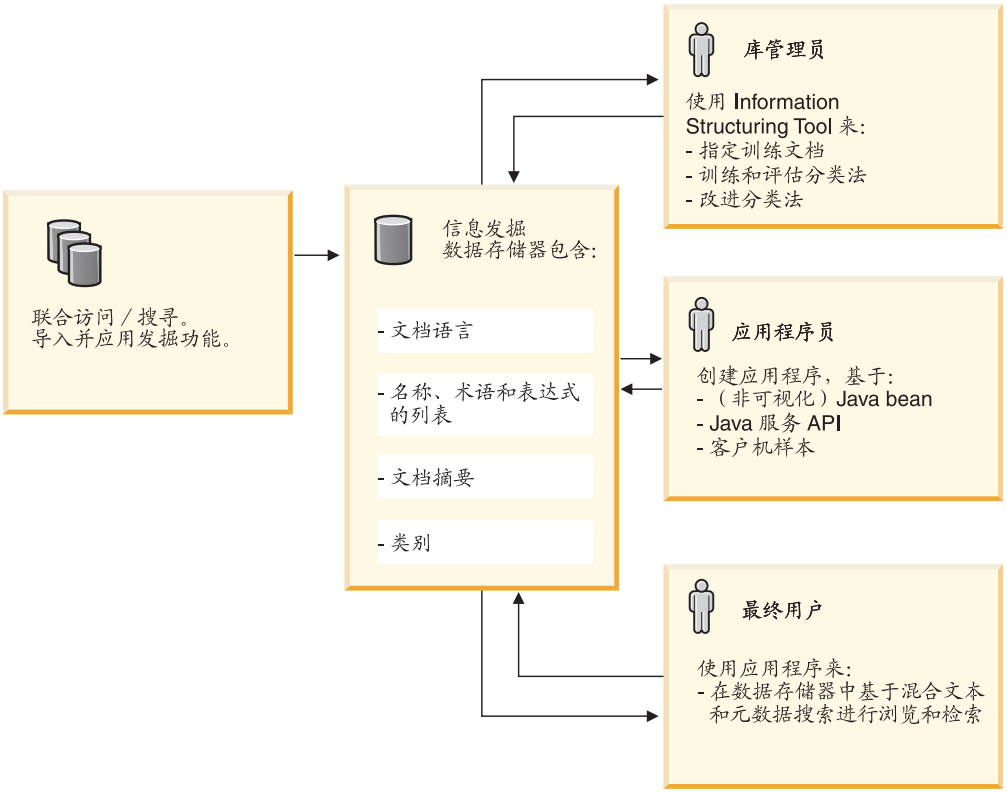

图 2. Information Mining 角色和操作

*系统管理员*设置硬件与软件环境并维护必需的资源，例如，文件系统空间和访问权限。系统管理员安装必需的 Enterprise Information Portal 组件并配置内容源和 Enterprise Information Portal 管理应用程序，这样就可以访问各种资源库。系统管理员还管理数据库级别的 Information Mining 元数据存储器。

*应用程序员*使用 JavaBean 或服务 API 创建应用程序。关于 JavaBean 样本，请参阅 *Application Programming Guide for Windows*；关于服务 API 的描述，请参阅 *Application Programming Guide*。应用程序员可以在 *Web 设计员*的帮助下进行此操作。

*库管理员*或*知识工程师*负责设置并维护用于发掘和检索的文档集合和资源。库管理员使用 Enterprise Information Portal 管理应用程序创建元数据映射和搜索模板，并使用 Information Structuring Tool（请参阅第 53 页的『构建分类法』）定义目录和分类法。用来自数据源或 Web 的文档填充 Information Mining 元数据存储器、使用应用程序员编写的应用程序也是库管理员的责任。

*最终用户*使用应用程序员创建的应用程序，执行基于库管理员或知识工程师所创建和维护的资源的信息发掘和检索任务。根据最终用户和库管理员之间的工作分发，最终用户还可以参与从内容服务器选择文档以及填充 Information Mining 元数据存储器。

使用 Information Mining 的一个示例

Electro 公司是一家为大规模市场生产电气设备的公司。其业务量至少涵盖五种不同的具有大范围独立配置的产品。

销售部门拥有客户对某些设备应用区域的首选项的信息。这些用法简要表解释了顾客如何使用产品以及各种配置选项。每个简要表都与特定市场、首次展示以及关系管理策略相关。

服务部门拥有关于哪些部件组成设备、它们如何装配、这些部件的供应商是谁以及各部件的可维护性与可靠性的信息。

合同管理部门保存转售商和转包商的信息。他们还可访问对特定类型合同有效的条件和条款相关的法律文档。

最近，他们注意到对某些顾客的销售显著下降。应用各种电气设备的方法之间的竞争已经有所改变，而顾客的期望也已经有所更改，从而配合这些新技术的推进。

为了跟上这些更改并受益于它们，Electro 公司成立任务组开发将公司带回商务领域的策略。

第一步是设计 IT 基础结构来访问此相关信息，这样规划者就可以制定快速且博识的决策。

以下是此类信息的示例:

- 关于竞争者产品、功能、价格和顾客可接受性的数据
- 从客户观点出发，对于 Electro 公司产品线的优点和弱点的认识
- 这些产品通常使用的领域中的趋势和远景

此信息驻留在完全不同的数据源中，它们具有不同的硬件和软件平台、组织级别（例如，分层的、索引的或平面文件）和文档类型（例如，数据库记录或 HTML）。

第 40 页的图 3 说明了使用 Information Mining。

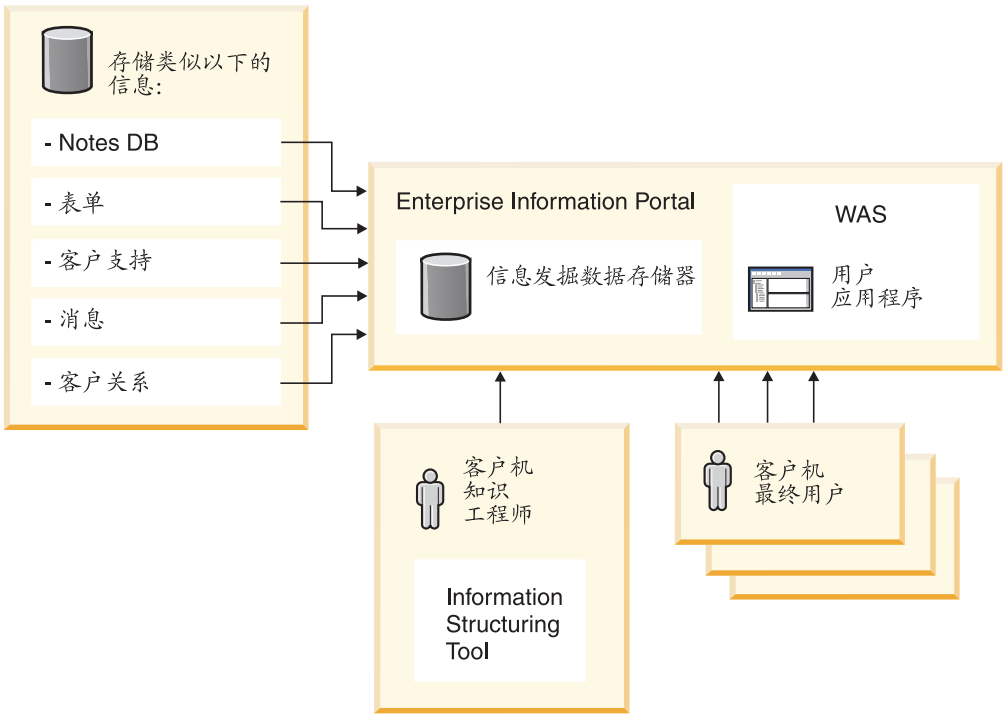

图 3. Information Mining 的一个示例

任务组决定创建一个门户，以从驻留在规划者客户机上的瘦客户机（浏览器）访问所有这些完全不同的源。他们选择 Enterprise Information Portal，因为它提供所有构件，用来创建这样一个门户网站，甚至允许创建旧数据源的定制连接器，这些数据源是在商务应用区域中少有标准时创建的。

涉及到以下步骤:

1. 设置硬件和软件基础结构
2. 定义数据源的访问方法，设置必要的连接并创建到相关数据的映射
3. 以规划者可以浏览并在此数据中导航的方式组织数据
4. 创建最终用户应用程序

步骤 3 是 Information Mining 开始活动的位置。一旦基础结构是可用的并正在运行，数据源已标识，相应的连接已建立，而且有关的映射已创建，则所需的数据就可从单一点上访问，而且此数据的子集也可使用联合搜索来定义。下一个问题是如何为预测或趋势过滤出所需的数据，以及如何以适合策略规划过程的方式组织此数据。

任务组定义了负责维护、组织和更新策略规划信息的*知识工程师*。为了从驻留于各种数据源的大量文档的集合抽取相关信息，知识工程师要会见参与过去策略规划的员工，从而了解过程如何交互以及经过证明的好的实践，并且搜索顾客关系和支持数据库。

使用 Enterprise Information Portal 的搜索功能，可以容易地通过客户姓名、地址或设备属性来访问来自这些数据库的文档。但是，确定用法简要表所需的信息隐藏在文本中，访问此信息的唯一方法是使用 Enterprise Information Portal Information Mining 服务以智能方式来分析文档内容。

这些服务提供的一种有用的信息类型是文档内容的主题特性描述，称为类别，例如，*此文档是关于 PDA 的*。Information Mining 归类服务通过分析文档内容将其指定到类别。类别是在称为分类法的主题层次结构中构造的。明显可用的和自动创建的元数据驻留在资源库中，这些资源库由称为目录的、帮助加快访问和检索的 Information Mining 服务维护。

使用 Information Structuring Tool，知识工程师可以定义说明顾客使用设备的方式的目录。

第 41 页的图 4 显示一个目录。

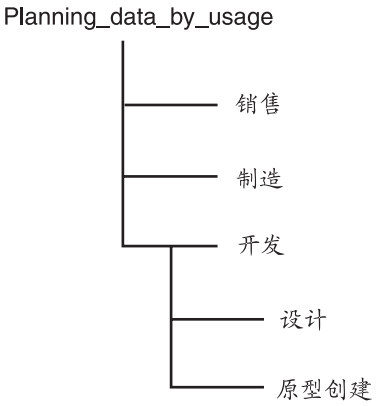

图 4. 一个目录示例

通过对所有客户和他们在销售和支持数据库中使用设备的方法执行联合搜索，知识工程师能够识别在分类法中代表每个类别的训练文档集合。这会导致当更密切地查看有关数据时显示新类别，从而引起分类法的重组。

一旦分类法是稳定的而且每个类别有足够的训练文档数，则知识工程师就使用 Information Structuring Tool 训练分类法。训练创建归类模型，可用来通过使用归类服务而将文档指定给类别。

同时，IT 部门的程序员使用 Enterprise Information Portal 瘦客户机和 JavaBean 或 Java 服务 API 来为战略规划人员创建最终用户应用程序。

此应用程序由许多为在策略规划中使用而定制的搜索模板组成。通过使用这些模板，规划人员可以将从不同的后端组合检索到的文档植入目录。批量载入目录时，Information Mining 服务会自动将类别指定到文档。如果已经标识了新的用法简要表，则知识工程师从而可以使用规划者标识为培训材料的新文档重新组织分类法。目录进行了重新培训且新结果传递到了规划者。

以上示例说明，通过结合 Information Mining 功能来反映客户的期望和需要，象 Electro 公司这样的公司可以与市场的变化保持同步，从而保持公司的竞争力。

支持的语言和格式

Enterprise Information Portal Information Mining 服务支持以下语言（请参阅第 42 页的表 5）：

表 5. 支持的语言

语言	语言标识	信息抽取	摘要	分类	群集
英语	x	x	x	x	x
德语	x		x	x	x
法语	x		x	x	x
丹麦语	x				
芬兰语	x				
意大利语	x		x	x	x
挪威语	x				
葡萄牙语	x		x	x	x
西班牙语	x		x	x	x
瑞典语	x				
韩国语	x		x	x	x
日语	x	x	x	x	x
中文（繁体和简体）	x		x	x	x

要获取受支持文档格式的列表，请参阅 第 105 页的第 9 章，『文档格式』。

概念

消耗的信息量是不断增长的。大多数组织有数量更大且不断增长的联机文档，这些文档包含潜在价值巨大的信息，例如，客户反馈数据、在竞争日益激烈的市场中至关重要的战略信息、或者是提供洞察新的和正变化的商业机会的信息。Information Mining 服务被设计用作处理大量联机文档的应用程序。

系统体系结构

第 43 页的图 5 说明了 Information Mining 系统体系结构。

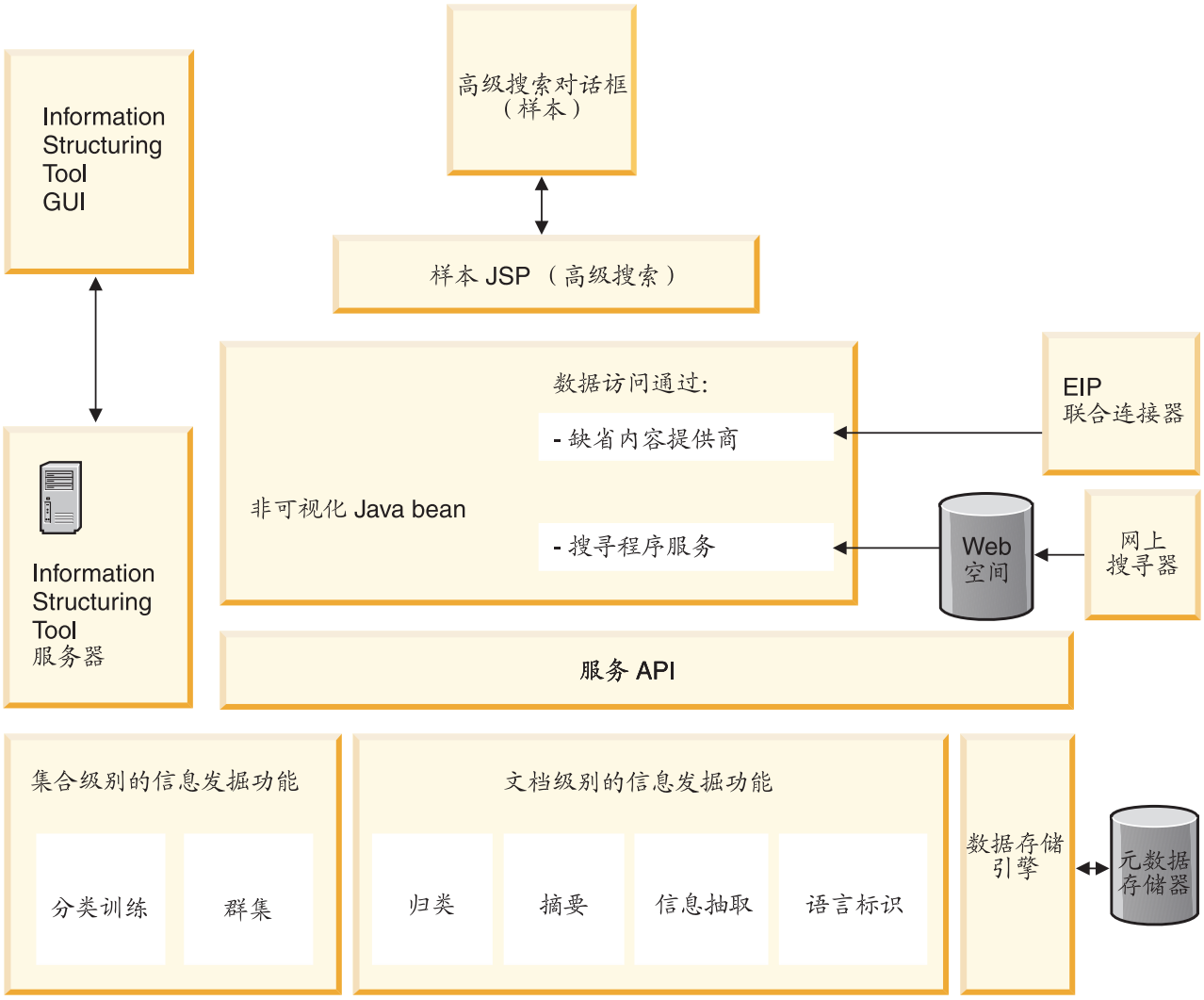

图 5. Information Mining 系统体系结构

图右边的方框是指 Information Mining 服务所使用的组件，但我们并不认为它们是服务的一部分，即:

- 联合连接器（OO API 的部件）
- 网上搜寻器

Information Mining 功能有不同的层，即:

1. Java 服务 API

   此层显示 Information Mining 功能和元数据持久，作为一个相容的 Java API。

2. 非可视 JavaBean

   此层使用现成可用的组件，这些组件基于应用来自标准 bean 的事件类型和约定的 JavaBean 规范。

3. 样本 Java Server Pages

   此级别由使用非可视 JavaBeans 的样本代码组成，样本代码举例说明用于高级搜索（换句话说，就是带类别限制的正文搜索）的应用程序。

4. Information Structuring Tool

一个带图形用户界面的应用程序，用于创建和维护分类法。

Information Mining 概念

为了全面了解并能够有效地使用 Information Mining 功能，本节论述它的主要概念。

Information Mining 服务提供了一个基础结构，用于创建与维护与个别文档或文档集合相关的信息。此信息关于一个被称为其**元数据**的文档。

库是 Information Mining 数据库内容的概念视图。库包含一个目录集。

目录是文本文档的元数据存储器，它包含:

- **目录模式**，用于为每个文档定义哪些属性被存储。
- **分类法**，是**类别**的分层树结构。
- 基于文档训练结果的**归类模型**，可用来自动将类别指定给文档。此模型是用 **Information Structuring Tool** 生成的，在那里可创建和训练分类法。该模型被用作为对归类服务的输入。

模式指定可以为目录中的文档而生成或存储的属性的名称和类型。模式是预定义的，它包含以下属性:

- 字符串类型的 IKF_CONTENT
- 字符串类型的 IKF_TITLE
- 字符串类型的 IKF_AUTHOR
- 字符串类型的 IKF_CATEGORIES
- 字符串类型的 IKF_SUMMARY
- 字符串类型的 IKF_LANGUAGE
- 字符串类型的 IKF_FEATURES
- 字符串类型的 IKF_COMMENTS
- 时间戳记类型的 IKF_DATE
- 整数类型的 IKF_IDNUMBER

目录根据目录模式创建**记录**，以存储从导入的文档中抽取或创建的信息。记录有一个唯一的标识和一套名称值对。唯一标识（也称为持久对象标识或 PID）将所创建的记录链接回原来的文档源。

第 45 页的图 6 显示一个样本记录。

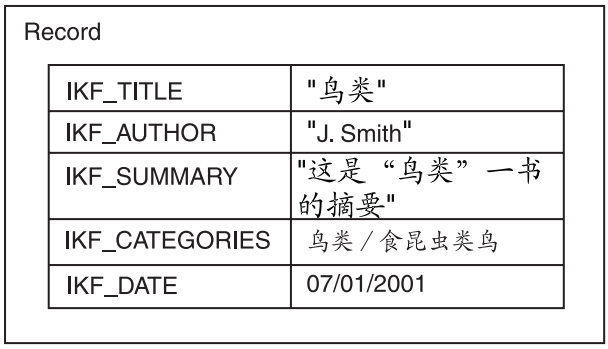

图 6. 样本记录

如果使用 JavaBean 或 服务 API 设置 Information Mining 记录的值，则确保您使用的值在定义的大小限制范围之内。否则，您将得到 `DKIKFSizeOutOfBoundsException`。已定义的限制范围是:

表 6.

键	最大大小（字节数）
IKF_CONTENT	209715
IKF_TITLE	2048
IKF_AUTHOR	2048
IKF_CATEGORIES	8192
IKF_SUMMARY	8192
IKF_LANGUAGE	8
IKF_FEATURES	524288
IKF_COMMENTS	8192

一旦创建了记录，就通过将其指定到一个类别而将该记录存储在目录中。即使也可以根据存储在记录中的另一个值来选择类别，但通常，可用归类结果来选择适当的类别。记录必须指定到类别，因为这还包括建立文档内容索引操作（以使文本搜索可以进行）。每个目录都有一个文本索引，这说明所有搜索结果一直自动位于目录搜索范围中。

数据存储引擎是一个组件，它维护对持久数据存储器的访问。

Information Mining 工具

Information Mining 服务提供用来处理联机文档的功能。包括以下这些功能:

- Information Structuring Tool 创建和维护目录。
- 语言标识服务自动地检测写文档所用的语言。
- 归类服务自动地将文档指定给您以前使用 Information Structuring Tool 定义的类别。
- 摘要服务分析文档中的词语和句子，用以生成文档的摘要。
- 信息抽取服务自动地在文本中识别重要的项，而不需您定义依赖于域的词汇。
- 群集服务将文档集划分为组或群集。每个群集中的文档共享公共的特征。群集不是预定义的；它们是自动派生的。
- 高级搜索搜索文档中的文本（文档存储在局限于特殊类别的目录中）。

Information Structuring Tool

Information Structuring Tool 是一个基于 Web 的应用程序，它提供了一种创建并维护一组目录（称为库）的方式。目录用来存储从文档中抽取的元数据，并且目录与用来组织文档的分类法（它使用预定义的组织）相关联。分类法是根据文档主题内容对其进行分类的类别的层次结构。

例如，通过使用 Information Structuring Tool，库管理员可以定义一个说明鸟类喂食习惯的目录。

图 7 显示样本目录。

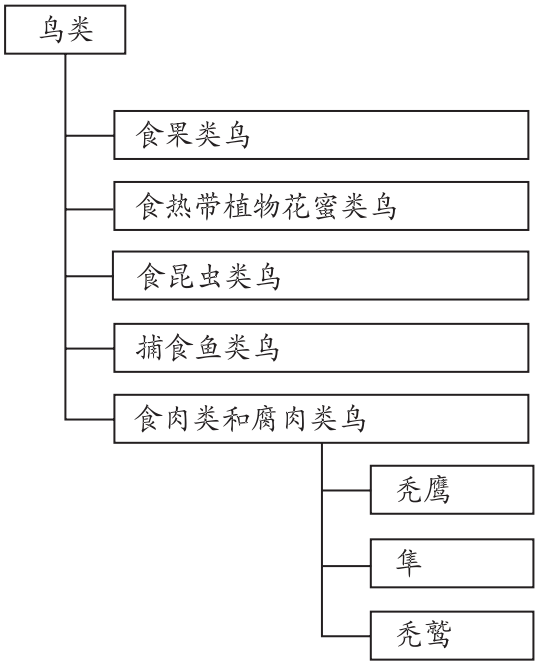

图 7. 样本目录

结构良好的类别系统对在大量数据中查找有关信息有极大的帮助。选择的类别与对文档集合的使用意图相匹配，并且类别必须已经事先用样本文档训练过。随后，归类服务可以使用用 Information Structuring Tool 创建的归类模型，来自动地将类别指定到文档。

Information Structuring Tool 功能包括:

- 创建、重命名和删除目录。
- 编辑目录的描述。
- 创建、重命名和删除类别。
- 向类别添加训练文档和从类别中除去训练文档。
- 查看训练文档的内容。
- 启动和停止目录训练过程。
- 获取关于目录中训练数据的质量的反馈。

关于如何安装和使用 Information Structuring Tool 的详细信息，请参考第 53 页的『构建分类法』。

语言标识

语言标识服务在给定的一组语言中选择一种与写文本文档所用的语言最接近的语言。

语言标识服务返回已排序的语言列表和每个文档的可信度值。语言键是根据 ISO 标准 639 以两个字母指定的。可信度值是指示文档与语言匹配程度的度量，它由 0（差）到 1（好）之间的浮点数来表示。语言标识算法被设计用来确定单一语言文档中的语言。因此，对于多种语言的文档，不能保证按可信度值的排序反映了文档的正确语言。

可以检测以下语言:

- 英语 EN
- 德语 DE
- 法语 FR
- 丹麦语 DA
- 芬兰语 FI
- 意大利语 IT
- 挪威／博克马尔语 NB
- 挪威／尼诺斯克语 NO
- 葡萄牙语 PT
- 西班牙语 ES
- 瑞典语 SV
- 韩国语 KO
- 日语 JA
- 简体和繁体中文 ZH

可以设置以下特性:

- **maxResults**（仅当使用 Java 服务 API 时可用）:

  可为每个文档确定并返回的最多语言数。它是一个大于或等于 0 的整数。缺省值是 1，表示返回最好等级的结果。如果值设置为 0，则所有被识别的语言都按等级顺序返回，除非语言的可信度值低于 0.01。

可以将语言标识用作为另一个 Information Mining 服务的预处理步骤。例如，可以在进行文档抽取之前，使用语言标识来查找所有英语或日语文档。

归类

归类是将类别指定给文档以及用预定义的组织模型（用 Information Structuring Tool 创建的）来组织文档的一种方法。

因此，在可以使用归类服务之前，您必须定义并训练分类法，以 Information Structuring Tool 来创建一个这样的模型使用。

归类结果包含类别和指示文档与类别的匹配程度的可信度值。对每个文档都会返回一组这样的结果。列表是根据返回的可信度值排序的。

可以设置以下特性:

- **maxResults**:

  对每个文档返回的最大类别数。缺省值是 –1，表示返回所有的类别。结果列表是分等级的。

- **minConfidence**:

指定给文档的可信度值指示文档与类别的匹配程度。参数 minConfidence 指定最小值，范围从 0（不太合适）到 1（较为合适）之间。缺省值设置为 0，表示返回指定给文档的所有类别。结果列表是按从较为合适到不太合适分等级的。

- **catalogName**:

  这指定了要用于归类的目录。为了使用 Information Structuring Tool 进行归类，可以创建和训练目录。

关于受支持语言的列表，请参阅第 41 页的『支持的语言和格式』。

摘要

文档摘要由从文档中抽取的句子集合组成，是文档内容的特性描述。例如，摘要工具可以帮助您确定文档是否是相关的并且应该完整阅读，或者当作为查询结果的一部分而返回时，来帮助您决定是否跟随文档的链接。

摘要结果包含作为单个字符串和数据结构（矩阵）摘要，应用程序可以使用它来选择个别句子并确定它们是否是互相临近的。

摘要服务可以不同的方式来使用。这些方式确定了如何使用值 `maxLength` 和 `ratio` 用来确定摘要的长度。

- **maxLength**:

  摘要中的最大句子数。所创建的摘要不会比 maxLength 更长。缺省值是 3。

- **ratio**:

  摘要的句子数与文档总长度相比的比例。所创建摘要的长度由文档的总长度决定。缺省值是 0.1。

- **mode**:

  确定设置摘要长度所需的 maxLength 和 ratio 之间的关系。有以下不同的方式:

  – MODE_LESS_THAN_MAXLENGTH:

    摘要最多可有 maxLength 个句子。这是缺省方式。

  – MODE_EQUALS_RATIO:

    摘要的句子数完全符合 ratio。这是由 ratio 乘以文档中句子的总数确定的。

  – MODE_EQUALS_RATIO_BUT_AT_MOST_MAXLENGTH:

    摘要至少有句子比例数（ratio 乘以文档中句子总数），但句子数不多于 maxLength 个。

关于受支持语言的列表，请参阅第 41 页的『支持的语言和格式』。

信息抽取

分析文档时，重要任务是抽取提供关于文档内容的信息的项。这些关键元素可用于:

- 指示重要信息，帮助评估文档是否应受关注。
- 查找并存储用于精炼查询的关键概念
- 作为收集有关文档的标准

关键元素的示例有: 词汇项（如词语）、名称或多个词语构成的术语。

对于英语，信息抽取服务将查找到的关键元素进行标准化，而且如果这些关键元素是指同一个实体或表述同一个概念，将文本中出现它们的地方分组在一起。例如，文档

中出现 James J. Smith、Mr. Smith、James 和 Smith，则通过将它们映射到同一张标准化表单中来将它们标记为指向同一个人。词形的变化也被映射到它们的标准化表单中，例如，children 变为 child。

但是，对于日语，所有关键元素都按它们出现在文档中的样子来抽取。除日期、时间和货币表达之外不进行标准化，这里的标准化遵从 ISO8601 和 ISO4217。

信息抽取服务使您能分析文档的以下内容:

- 单个词和多个语的词汇项。例如，announcement、value、product cycle
- 地名、人名和组织名称。例如，Washington、Bush、Data Management Academy
- 缩写。例如，MB（兆字节）
- 关于日期、金额和数字的术语。例如，11 Jan. 1958、01/11/58、$30、thirty pence、4.5、5000

可以指定抽取三种类型的信息:

- 名称
- 术语
- 表达

通过使用 Java 服务 API，您还可以标识子类型（如下所列）和可信度值（表示子类型与所抽取词汇项的匹配程度）。可信度值的范围在 0（差）和 1（好）之间。名称、术语和表达式类型的子类型包括:

- **名称**
  - 地方。例如，Montreal 或 London
  - 人。例如，Tim Brown
  - 组织。例如，Smith 和 Son
  - 未知。例如，Smashing Pumpkins、Silicon Valley、CCTV（不带全称的缩写）
  - 其它。例如，AIS Plan、ISO Conference、Internet、Privacy Act Officer、JCAHO Performance Report
- **术语**
  - 未指定的术语。例如，entertainment conglomerate、art world、class variable、source code、data definition、process improvement initiative
- **表达式**
  - 基数。例如，four、fifty、70
  - 序数。例如，fourth、fiftieth
  - 百分比。例如，12%、sixty percent
  - 日期。例如，07/28/98
  - 时间。例如，18 hrs、4 o’clock
  - 金额。例如，DM90、thirty pounds
  - 缩写。例如，NY

信息抽取只对英语和日语文档起作用。可以使用语言标识服务作为预处理步骤，标识文档集合中的不是英语或日语的文档。您可以将信息抽取服务和其它发掘功能结合在一起，例如，将它用作摘要服务的预处理步骤，可只对那些关于 Bush 作为总统而非德克萨斯州州长的文档进行摘要。

群集

群集工具安排文档的集合，以使类似的文档分组在一起，而在不同组（群集）中的文档根据其内容而互相区别。这样，群集可以作为提供大文档集合的概观和标识有关文档的一种方法。同样，它还可以用来支持使用 Information Structuring Tool 构建分类法，方法是：在应用程序区域内群集训练文档。群集对于以下操作也是很有用的：查找集合内的相似文档（可能向您指出新趋势或新技术），以及查找重复的或特别相似的文档（可能会使您有兴趣进行比较分析）。

群集是将文档组织到群集中迭代过程，这样在每个群集内各文档的内容都尽可能相似，而不同的群集之间内容则尽可能不同。群集是将文档集合作为一个整体来处理的，这与上面的 Information Mining 服务形成对比，而与归类或摘要相似（在文档级别上起作用）。群集是通过以下操作进行的：互相比较每个文档中具有代表性的特征，并且根据文档特征的相似程度来对文档进行分组。

群集阶段中，没有新的文档可以添加到文档集中。

可以设置以下特性:

- **maxClusterCount**

  可返回的最大群集数。

- **minClusterCount**

  可返回的最小群集数。

- **clusterFeatureCount**

  每个群集返回的标签（关键字）数。

但是，这些值不约束群集器，仅仅是作为原则上的边界。群集服务的输出是一个结果列表。

群集只对英语文档起作用。可以使用语言标识服务作为预处理步骤，标识文档集合中的不是英语的文档。

高级搜索

与在整个 Enterprise Information Portal 内容服务器上执行的标准 Enterprise Information Portal 搜索相比，所谓的高级搜索只搜索那些其标识存储在 Information Structuring Tool 所创建的目录中的文档。为了进一步缩小搜索的范围，高级搜索不仅仅查询搜索文本，还将此搜索限制于特定类别中的文档。

可以设置以下参数:

- **catalogName**:

  这指定了要用于搜索的目录。您可使用 Information Structuring Tool 来创建和训练目录。

- **maxResults**:

  每次查询返回的最大搜索结果数。缺省值是 0，表示返回所有结果。

可以提交以下类型的查询:

1. 纯文本查询。此搜索返回所有与文本查询相匹配的文档。结果按相关程度顺序列出。
2. 纯类别搜索。此搜索返回指定到该类别的所有文档。结果按随意的顺序列出。

3. 文本和类别组合搜索。此搜索返回与文本查询相匹配的所有文档并指定到该类别。结果按相关程度顺序列出。

针对系统提交的高级搜索查询总是绑定到特定目录。这也称为目录搜索范围。不可能有交叉目录搜索，因为目录代表导入文档的视图，必须是独立的。

用于查询字符串的 BNF（查询语法）如下:

```
query_string         ::= term
term                 ::= ( term )
                     ::= single_term
                     ::= compound_term
single_term          ::= category_term
                     ::= text_search_term
                     ::= string_term
                     ::= number_term
compound_term        ::= term binary_bool_operator term
                     ::= unary_bool_operator single_term
category_term        ::= ( DKIKFCategory category_operator category_path_value)
text_search_term     ::= ( "attribute name" CONTAINS text_search_value )
string_term          ::= ( "attribute name" string_operator string_value )
number_term          ::= ( "attribute name" basic_operator number_value )
binary_bool_operator ::= AND | OR
unary_bool_operator  ::= NOT
category_operator    ::= >= | =
string_operator      ::= LIKE | basic_operator
basic_operator       ::= > | < | <= | >= | != | =
category_path_value  ::= "category path"
text_search_value    ::= "'string'"
string_value         ::= "string"
number_value         ::= "integer" | "decimal number"
```

- 终结符（字符串和数字）代表公共术语。
- category_operator '=' 将搜索范围限制于仅一个类别。
- category_operator '>=' 将搜索扩展到此类别及其在类别树上的所有子类别。
- 在 CONTAINS 子句中的字符串搜索可以包括: 表示单个字符的通配符（ "_" ）和表示任意数量字符的通配符（ "%" ）。例如, _LOB 可以与 BLOB 和 CLOB 匹配, 而 %name 可与 filename 匹配。只有已标志为可搜索的模式属性（例如 IKF_CONTENT）方可使用字符串运算符 CONTAINS 来查询。
- 在 LIKE 子句中的搜索字符串还包括 SQL 中使用的通配符。
- 关于当前受支持属性名称的完整列表，请参考第 44 页的『Information Mining 概念』。

查询示例:

- 纯文本查询:

  ```
  ("IKF_CONTENT" CONTAINS "'southern Africa'") AND NOT
  ("IKF_CONTENT" CONTAINS "'Cape'")
  ```

- 纯类别查询:

  ```
  ("DKIKFCATEGORY" >= "birds/Fruit eaters")
  ```

- 文本和类别组合查询:

  ```
  ("IKF_CONTENT" CONTAINS "'South Africa'") AND
  ("DKIKFCATEGORY" >= "birds/Birds of prey and scavangers/Falcon")
  ```

- 属性查询:

  ```
  ("IKF_SUMMARY" LIKE "humming birds in the tropics")
  ```

  或

```
("IKF_FEATURES" LIKE "Goethe") AND ("IKF_TITLE" = "Faust")
```

编程接口

Information Mining 功能可用于构建以下应用程序:

- Java 服务 API
- Information Mining JavaBeans

Java 服务 API 集成了全部 Information Mining 功能（作为 Information Structuring Tool 的一部分的目录维护除外）作为 Enterprise Information Portal 服务。它基于 Java RMI 提供客户机 / 服务器通信。

使用 Java 服务 API 的应用程序可以:

- 确定编写文档的语言
- 创建文本文档的摘要
- 对文档分配类别
- 从文本文档抽取信息，例如名称、术语或表达式
- 将相似的文档分为一组
- 存储和查阅目录中的文档的元数据
- 对限于某些类别的文档和对属性（例如摘要）执行文本搜索

Java 服务 API 能够使用直接方法调用以本地方式运行，或者使用 Java 远程方法调用（RMI）以远程方式运行。运行远程方式使您能将一台服务器配置为运行 Web 应用程序的应用程序服务器，将另一台服务器配置为执行文本分析、建立索引和进行搜索的 Information Mining 服务器。通过服务器任务机制，可将完整的任务发送到 Information Mining 服务器（远程机器）并在该机器上执行所有处理。

使用网上搜寻器时，必须使用 JavaBean 来实现相应的访问机制。网上搜寻器访问在 Java 服务 API 级别上不可用。

关于 Information Mining Java 服务 API 的详细描述，请参考 *Application Programming Guide for Windows*。

第 52 页的图 8 说明了 Information Mining 远程配置。

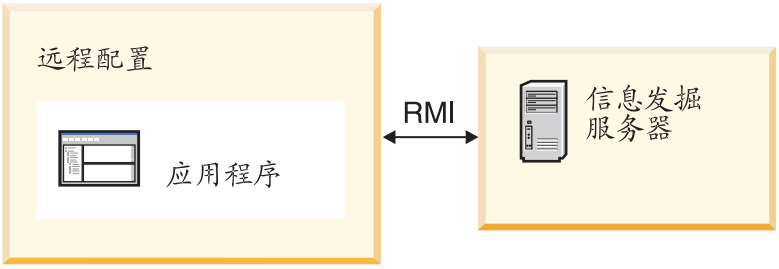

图 8. Information Mining 远程配置

Information Mining bean 是用于快速应用程序开发的高级 Java API，并且是根据 JavaBean 规范构建的。bean 不支持服务器任务处理，所以出于性能原因，所有使用 bean 的应用程序开发应当在同一台机器上进行。

每个 bean 都可由事件驱动来使用，而且其中一些 bean 提供可以直接调用的方法。通过 Enterprise Information Portal bean 使用的结果事件支持可以完成与现有 Enterprise Information Portal bean 的集成。这意味着联合搜索和网上搜寻器结果事件是与 Information Mining 服务兼容的，并且 Information Mining 结果可以由 EIP 处理。关于 Information Mining bean 的详细描述，请参考 *Application Programming Guide for Windows*。

第一步

Enterprise Information Portal Information Mining "第一步" 是一个故事板教程，它向知识工程师、管理员或应用程序员展示如何在基于现实情况的商业环境中应用 IBM Information Mining 技术。该教程的结构如下:

- 简介
- 可以开始第一步之前需要做的工作
- 组织样本数据
- 访问此数据
- 使用样本客户机
- 除去样本数据
- 更多阅读参考

要访问 Information Mining 第一步，请执行 `<CMBROOT>\ikf\firststeps\first_steps.html`。

要在 Windows 上运行 Information Mining "第一步"，单击 "开始" 按钮并选择 **Enterprise Information Portal for Multiplatforms 8.1→<信息发掘>→第一步**。

Enterprise Information Portal Information Mining "第一步" 还用作安装验证。

构建分类法

Information Structuring Tool 是一个基于 Web 的应用程序，它提供了一种创建并维护称为库的一组目录的方式。目录用来存储元数据，并且与在目录中组织信息的分类法相关联。分类法是根据文档主题内容对其进行分类的类别的层次结构。例如，分类法的最高级别可以包含类似*商业*、*文化和运动*类别，其中在下一个级别中*运动*又细分为*团体运动*和*田径运动*。*团体运动*的再下一个级别细分为*足球*、*棒球*、*网球*。

通过将培训文档分配给类别并训练目录，Information Structuring Tool 会创建一个归类模型，该模型最后可由归类服务用来将类别分配给文档。

安装 Information Structuring Tool

在 servlet 容器中，必须将 Information Structuring Tool 部署为 Web 应用程序，例如 IBM WebSphere Application Server（WAS）servlet 引擎。

但要注意，不允许有两个 Information Structuring Tool Web 应用程序（例如，一个名为 IST1，另一个名为 IST2）针对同一个 Information Mining 实例来工作。

在部署 Information Structuring Tool 之前，确保已安装并配置了正确的 WAS 版本。详细信息请参考 *Planning and Installing Enterprise Information Portal*。

部署 Information Structuring Tool 所需要的用户访问权为:

- 对于 Windows: 管理员权限
- 对于 AIX: root 用户特权
- 对于 Sun: root 用户特权

入门

该应用程序提供基于 Web 的界面，可以在该界面上定义、维护和训练分类法。有两个框架。左边的框架称为*类别*视图，用于创建并维护分类法。右边的框架称为*记事本*，提供使用该应用程序的信息。记事本显示了用来训练和评估该目录的文档的一系列选项卡。要使用该应用程序，必须现在左边的类别视图框架中创建目录或类别。

访问权

在 Information Structuring Tool 中，由 Enterprise Information Portal 维护用户名和密码信息。只能输入 Enterprise Information Portal 已知的用户名和密码。

当启动 Information Structuring Tool 时，会出现一条安全警告消息，因为用来上载培训文档的 Java Applet 需要对文件系统具有读权限。如果拒绝，则不能上载训练文档，从而也不能训练分类法。

Information Structuring Tool 可以在多用户环境中运行。要允许多个用户查看同一个分类法，Information Structuring Tool 提供了锁定机制来控制对目录及其类别的访问。

用户可以在开始使用目录前通过选择和锁定目录来显式地锁定某条目录，或者可以开始使用目录，例如，添加培训文档，在这种情况下会自动锁定目录以防止访问冲突。其它用户可以查看此目录，但不能更改该目录中的任何内容，直到它再次由锁定该目录的用户解锁为止。

请注意，如果关闭其中部署了 Information Structuring Tool 的应用程序服务器，会删除所有锁定。

定义分类法

目录是分类法的定位点，分类法是由类别组成的树形结构。

定义新目录并选择适当类别的步骤如下:

1. 确定要定义并创建新分类法的类别。

   在目录视图中，请选择**库**并单击鼠标右键。显示出菜单。选择**新建目录**，创建了一个目录图标。通过输入编目名称并单击 **Enter** 键重命名此图标。创建了一个具有相同名称的文件夹。这是根类别。记事本的内容有所更改。

   将新目录添加到库的另一种方法是，导入在 Information Structuring Tool 外部（例如，在文件系统中）创建的现有分类法。更多信息请参考第 56 页的『上载训练文档』。

   要有效地使用分类法，必须锁定目录。当创建目录时会自动将其锁定。要锁定现有类别，请选择该类别并单击鼠标右键。显示出菜单。选择**锁定类别**，类别状态图标有所更改。图标用于显示不同类型的目录状态:

   

分类法树已折叠并且未被任何用户锁定。

分类法树已展开并且未被任何用户锁定。

分类法树已折叠并且已由当前用户锁定。

分类法树已展开并且已由当前用户锁定。

分类法树已折叠并且已由另一个用户锁定。

分类法树已展开并且已由另一个用户锁定。

图 9 是两个目录的示例。

图 9. 目录示例

2. 一旦创建并锁定目录，就可以添加、重命名或删除类别。首次创建新目录时，将同时创建与该目录名称相同的根目录。您可以重命名此类别。要执行此操作，请选择该类别并单击鼠标右键。选择**重命名**，输入新名称，然后按 **Enter** 键。

要添加新类别，突出显示要对其添加新子类别的类别，并单击鼠标右键。选择**新建类别**，输入新名称，然后按 **Enter** 键。树结构中同一级别上的类别名称必须唯一。

不能删除根类别。只有当目录已删除时才能删除它。删除类别还会除去类别中的所有子类别（树结构中的较低类别）、所有培训文档和所有记录。

3. 与选定类别相关的进一步描述信息显示在**特性**选项卡中。首次访问时描述字段为空。要添加或编辑目录描述，请单击**编辑描述**按钮。显示出一个窗口，可以在其中输入描述。

选择训练文档

归类模型的质量严重依赖于分配给每个类别的训练文档的质量。

培训文档必须是 Enterprise Information Portal 支持的文档格式之一。对于受支持格式的列表，请参阅第 105 页的第 9 章，『文档格式』。

选择合适的训练文档集合是必需的。文档应当满足以下条件:

- 能表现类别
- 包含相当数量的描述性文本，没有很多标记或词汇列表
- 都是以同一写作样式编写的，例如，如果其它文档是报表样式，则要避免散文样式
- 所有文档长度相当，最好不要太长；培训文档还应当与想要使用归类服务进行归类的文档具有大致相等的长度

建议对每个类别使用大约 40 个培训文档的集合；但是，如果选定的类别更一般，则需要更多文档。必须仔细选择类别，并且类别必须有意义；对人工建立索引含糊和不清楚的类别和训练文档肯定会在自动处理期间引发问题。

上载训练文档

要添加培训文档，请选择相应目录，显示出**培训文档列表**窗口。要将文档添加到此列表，请单击**添加文档...**按钮。

出现**添加训练文档**窗口。上载文件后不需要关闭此窗口；它可以随后用来将文档上载至另一个类别或另一个目录。

要添加训练文档，请单击**浏览...**按钮，并在**打开**窗口中选择相关文件或目录。

可以上载同一目录中的一个或多个文件或者整个目录。如果所选目录为空，会得到通知。

这使您能够导入并使用在 Information Structuring Tool 外部（例如，在文件系统中）创建的现有分类法。

如果文件系统中的所选目录（例如，Development）包含子目录 Design，并且分类法树中的类别 Development 也包含子类别 Design，那么子目录中的文件会添加到子类别中。如果子类别不存在，则创建它，并且将文件添加到此新建的子类别。

选择文档语言和所有选定文件的格式。除了纯文本文件，始终使用格式“自动检测”。

要将文件添加到培训文档的列表，请单击**提交**按钮。

图 10 显示了添加训练文档的过程。

图 10. 添加训练文档

显示**上载过程**窗口。可以从此窗口取消文档上载过程。如果取消了上载过程，则不会再添加文件。会除去已作为训练文档添加的文件。但如果在上载期间添加了任何子类别，则它们不会从目录中删除。

如果无法成功上载文件，会自动显示**上载状态**窗口，该窗口说明了无法上载文件的原因。例如:

- 具有相同名称的文件已存在。
- 文件是空的。
- 文件无法上载到服务器。

在“上载状态”窗口，按下**训练文档**选项卡以返回到“训练文档”列表。其中显示所有已成功上载到该类别的训练文档。使用**上一页**和**下一页**按钮查看列表中的所有文档。

如果要将相同的文档作为一个训练文档添加到多个类别，则将文件上载到第一个类别然后将其复制到其它类别。不要再次上载此文件。要进行复制，在“训练文档列表”窗口中选择一个或多个文档并按下**复制**。在出现的窗口中，按**浏览...**选择要将文档复制到其内的一个或多个类别，然后按**提交**。

在文件上载期间不允许以下操作:

- 解锁目录
- 从 Information Structuring Tool 中注销
- 启动目录训练和评估

- 重命名目录
- 删除目录的上载状态信息

但是允许以下操作:

- 启动另一个文件上载过程，或者上载到同一类别或者是不同类别
- 使用另一个目录

图 11 显示了训练文档列表。

图 11. 训练文档列表

如果在文件上载期间关闭浏览器，这些文件将作为训练文档来添加（如果它们已被传送到服务器）。如果没有，则不会添加文件。

上载完所有训练文档后，选择目录以开始评估。

评估归类模型

一旦定义了分类法并且已将训练文档指定给每个类别，那么应当评估分类法。评估分类法可以帮助您估计训练文档与预定义的分类法的符合程度。这是一个交互过程，它包含以下步骤:

1. 启动评估
2. 评定评估结果
3. 更改分类法或训练文档
4. 重新运行评估

在每个评估迭代期间，评估过程会:

- 将训练文档分割为训练集（大约为文档的 80%）和测试集（大约为文档的 20%）。它使用训练集训练目录，并对测试集使用归类服务。
- 检查是否已将文档以足够高的可信度值分配给正确类别。范围在 0 到 1 之间，1 意味着文档是理想的。您可以设置可信度值。缺省值是 0.5。

可以在3 至 5 次迭代之间进行选择。3 次迭代是缺省值并且给出了分类法的强势和弱势的有用分析。当选择 5 时，所有文档将在训练和测试集中。

按下**启动评估**开始评估过程。

对于每个类别，在每个评估迭代中将计算以下内容:

- *正确*文档。在评估期间分配给 c 的类别 c 中的训练文档数量。
- *出站*文档。在评估期间分配给不同类别的类别 c 中的训练文档数量。
- *入站*文档。在评估期间分配给类别 c 但源自不同类别的训练文档数量。
- *未分配*文档。在评估期间未分配给任何类别的类别 c 中的训练文档数量。这可能包括已分配给某个类别但降至给定可信度值之下的文档。

显示评估过程在目录级别进展情况的概要，包括:

- 指示评估正运行或已停止的评估状态
- 列出上次评估日期的上次评估
- 完成的评估迭代数量
- 迭代的全部数量
- 平均全局精度，显示正确文档（即初始分配给该类别的文档以及入站文档）百分比
- 平均全局重调用，显示类别中正确文档百分比
- 正确文档，也就是正确分配的文档数量
- 错误放置的文档，也就是入站或出站文档的数量
- 未分配的文档

精度和重调用与分配的可信度值紧密相关。如果可信度值降低，精度会下降而重调用会增加，反之亦然。高精度值意味着已正确分配了许多训练文档；另一方面，高重调用值意味着已将大多数训练文档指定给某个类别，换言之，没有或只有极少的未分配文档。

评估结果

通过按下**评估结果**选项卡可以获取详细的评估结果。如果选择目录并按下该选项卡，会显示整个目录的结果；如果选择一个目录，则显示该目录的结果。

第 60 页的图 12 显示了目录级别的评估结果:

图 12. 目录级别的评估结果

从全局目录结果开始。红色（趋于临界）和蓝色（未到临界）值表示类别或训练文档存在错误。类别的质量由其训练文档定义，所以建议研究一下文档移入（入站）和移出（出站）类别的趋势。

对类别或训练文档所作的更改在很大程度上取决于对精度和重调用的强调是否均等，或者仅对精度的强调多一些。精度值越高，该类别与分类法中的其它类别相比就越特殊。另一方面，重调用值越高，未分配的训练文档就越少。

评估结果显示在两个级别:

1. **目录**级别:

   对于目录中的每个类别:

   - 精度和重调用百分比
   - 入站文档百分比
   - 出站文档百分比
   - 正确文档百分比
   - 未分配文档百分比

2. **类别**级别:

   对于文档类型入站和出站:

   - 训练文档和起源或目的地类别

   对于文档类型正确和未分配:

   - 训练文档

解释评估结果

下一节建议如何解释评估结果，但要始终记住，分类法是作为一个整体来操作的，这意味着对分类法的一节所作的更改可能会对在分类法的其它地方产生的结果具有相反的影响。

图 13 显示了类别级别的评估结果:

图 13. 类别级别的评估结果

从目录级别的评估结果开始，选择具有潜在较低精度和再调用值以及具有标记为红色（趋于临界）或蓝色（未到临界）的训练文档值的类别。对于依次排列的每个类别（可从左边的框架中选择），请检查以下内容:

- 许多**入站**文档:
  - 类别从其它许多类别中获取文档:
    - 类别不够特别。将这些文档复制到此类别。
    - 通过选择更合适的训练文档或将类别分割为子类别，使该类别更特别。
  - 类别从一个或两个类别获取文档:
    - 将这些文档复制到此类别。
    - 检查源类别。保留此类别有意义吗？或应当合并这些类别吗？要合并，请将培训文档复制到想要保留的类别并删除其它类别。
- 许多**出站**文档:
  - 类别失去的文档被许多其它类别获得:
    - 类别不够特别。选择更合适的训练文档或考虑删除该类别。
  - 类别的文档被一个或两个类别获得:
    - 将这些文档复制到其它类别。

- 检查此类别。保留此类别有意义吗？或应当合并这些类别吗？

- 许多**未分配的**文档:
  - 在以下方面将未分配的文档与正确的文档作比较:
    - 大小

      例如，如果未分配的文档比较短，则考虑将两个文档连接起来形成一个较长的新文档，以匹配正确文档的大小。
    - 样式

      如果样式不同，则删除该文档。
    - 主题

      如果主题略微不同，但仍然与正确文档中的主题相关，则考虑查找包含此主题的其它训练文档并将它们上载到同一类别或新建的类别中。

    如果重调用和精度值比较理想，并且相同文档在重复的评估迭代期间保持未分配，则将它们删除。

只有对分类法作出更改（例如删除类别或合并两个类别，或移动了许多文档），才需要再次启动评估过程。在需要重新启动评估过程之前，可以在类别级别作较小的更改，比如添加新的训练文档或复制文档。请注意，没有撤销功能可用于撤销所做的更改。

当没有引起注意的红色和蓝色值或者得到超过 90% 的精度和重调用级别时，可停止评估。

训练目录

在评估了分类法并对结果感到满意时，需要使用所有训练文档对其进行训练，以生成特殊类型的元数据，称为归类模型，以后可以使用 Information Mining 归类服务将其用来归类新文档。

要开始训练阶段，选择需要训练的目录，并按下**训练**选项卡，然后按下**开始训练**按钮。

显示在右边的框架顶部的是选定目录的名称。您还可以看到目录的培训状态。培训状态的类型为:

- 文档无法归类。训练目录以更新归类器。（这也是新创建目录的缺省值。在此情况下，**上次训练**日期为空。）无论目录为新目录并且还未训练，还是最新的训练结果无效（例如，因为自上一次训练目录以后已重命名或删除了类别），试图使用归类服务来归类新文档会导致错误。
- 目录中的训练文档集已更改。而文档仍根据上一次的训练结果进行了归类。训练目录以更新归类。
- 正在对目录运行训练。
- 目录是最新的，不需要训练。

要停止训练过程，请单击**停止训练**按钮。请注意，如果培训停止，则目录培训过程必须重新运行。

无法将新的训练文档上载到正在训练的目录。

性能调整

当在目录中创建记录时，可搜索文本属性（例如 IKF_CONTENT）的文本索引会大量增加，并且搜索性能大大降低。要优化存储并增强性能，需要定期地重组文本索引，尤其在对索引做了大量更新之后。

文本索引的重组最好在计算机的低峰时间开始，例如在夜间。要运行索引组织，请切换至:

- 对于 Windows: `...\ikf\IkfReorg.cmd`
- 对于 AIX: `.../ikf/bin/IkfReorg.sh`
- 对于 Solaris: `.../ikf/bin/IkfReorg`

参数包括: IkfReorg <UserID><Password><DBName>

使用 IBM Web Crawler

本节描述并解释了如何配置 IBM Web Crawler 功能部件。如果选择了“功能部件”复选框，EIP 安装程序会安装此功能部件。

EIP V8.2 包含网上搜寻器、Lotus Notes 网上搜寻器、用于从已搜寻文件抽取数据的摘要器、基于 HTML 的文档、配置示例和支持的实用程序。IBM Web Crawler（也称为 GCS）需要 Java V1.3 或更高版本。

IBM Web Crawler 是基于 Java 的内容网上搜寻器和发掘器。当指向内容时，它会获取并发掘该内容。

IBM Web Crawler 可以在内部网、外部网或因特网 Web 中搜寻内容、在 Lotus Notes 数据库中本机或者通过 Domino 搜寻内容和在本地文件系统搜寻内容。将 IBM Web Crawler 构建为可以方便地添加新协议。内容可能为任意类型，例如，HTML、Notes 附件和多媒体。

IBM Web Crawler 可以从多种类型的内容中发掘元数据和文本。例如，可以通过以下方面发掘 HTML 内容:

- URL
- 标题
- 主体
- 上次修改时间
- 元标记，例如作者、关键字、描述等

从给定类型内容的一组预定义发掘器中进行选择。内容和/或发掘的元数据保存在本地磁盘中。IBM Web Crawler 可以使用网络解决方案 Outside In 技术从 200 多种类型的内容中抽取文本，是在搜索应用程序中使用的理想合作功能。还已经将 IBM Web Crawler 构建为可以方便地添加新发掘器。

IBM Web Crawler 可用于 Windows NT 4.0 和 Windows 2000 操作系统。可以用大约半个小时的时间安装、配置和使用 IBM Web Crawler。在 500 MHz 的 PC 机上，每秒钟大约可以获取和发掘 10 个文件。在测试时它可以达到 100 万个对象（200000 Notes）。它支持多个用户和每个用户的多个搜寻/发掘配置，并能够支持用户首选的本地语言。

IBM Web Crawler 功能

该安装程序安装两个文件:

x:/<install directory>/run
: 用于 Web 批处理文件和样本配置的 IBM Web Crawler。

x:/<install directory>/notes-run
: 用于 Notes 批处理文件和样本配置的网上搜寻器。

x:/<install directory>/lib
: IBM Web Crawler .jar 和 .zip 以及过滤文件。

配置和运行用于 Web 的 IBM Web Crawler

本节描述如何配置和运行用于 Web 的 IBM Web Crawler。用于 Web 的 IBM Web Crawler 访问 HTTP、FTP、新闻或文件服务器并创建 HTML 文档和其它对象的摘要。摘要为文件，每个文档或对象一个，包含元数据和全文本。

基本配置

本节包含解释如何编辑 XML 格式的 IBM Web Crawler 配置文件的指示信息。下面提供了两个样本配置以帮助您入门:

- 以 DB2 UDB 使用 IBM Web Crawler 的 config-db2.xml 文件。
- 不用 DB2 UDB 使用 IBM Web Crawler 的 config-sample.xml 文件。

1. 打开命令提示。
2. 将目录更改为安装了 IBM Web Crawler 的 run 子目录。例如，如果在 Windows 服务器上安装了 IBM Web Crawler，则输入 cd *x:*<cmbroot>\gcs\run。如果在 AIX 上安装了 IBM Web Crawler，则输入 cd /usr/lpp/cmb/gcs。

   技巧: 保留原始文件的副本是很重要的。文件中的错误会中断 IBM Web Crawler。在编辑时要特别注意。
3. 要以 DB2 UDB 数据库运行 IBM Web Crawler（伸缩性较强，速度较慢），则编辑 config-db2.xml 文件。例如，在命令提示符下输入 edit config-db2.xml。
4. 要不用 DB2 UDB 数据库运行 IBM Web Crawler（伸缩性较差，速度较快），则编辑 config-sample.xml 文件。例如，在命令提示符下输入 config-sample.xml。

   要运行 *n* 个 URL（无数据库）的搜寻，则需要机器上有大约 *n*/1000 MB 的内存以容纳搜寻的 URL 元数据。例如，要搜寻 500,000 个 URL，则需要 512 MB 的内存。要利用此内存，则编辑 crawlweb.bat 文件并增加 JVMXmx 的值。

配置网上搜寻器用于信息发掘

要使用网上搜寻器进行信息发掘，也就是能够将信息发掘功能应用到搜寻的文档，需要以下特定配置设置，这些设置与在上述样本配置文件（config-sample.xml 和 config-db2.xml）中给定的设置不同:

```
<globals ...

     max-urls="5000"
     temp-filepool-class="FullPathFilePool"
     summaries-dir="webspaces/ikf/disks/1/"
     summaries-filepool-class="DistributedEIPFilePool"
     ... >
  ...
</globals>
```

```
<group-list>
    <group ...>
       ...
         <summarizer-config>
            ...
               <resource-handler content-type="*"
                                 summarizable="EipHtmlSummarizable"
                                 summary-maker="EipHtmlRawSummaryMaker" />
               ...
         </summarizer-config>
       ...
     </group>
</group-list>
```

这些设置是信息发掘组件的样本配置文件（im-crawler-config-sample.xml）的一部分，该文件位于以下目录:

在 Windows 上:

```
<CMBROOT>\samples\java\beans\infomining\webcrawler\
```

在 UNIX（AIX 和 Solaris）上:

```
<CMBROOT>/samples/java/beans/infomining/webcrawler/
```

配置 IBM Web Crawler DB2 选项

要配置 DB2 选项，必须创建数据库。这会需要 DB2 管理员权限。可能需要切换至 DB2 管理员帐户。可以在 DB2 的允许范围内命名数据库，但如果数据库名不是 gcs，则必须更新网上搜寻器配置文件中的 dbname。

如果您具有数据库管理员权限，那么可以在 DB2 命令提示符下运行以下命令以创建数据库:

```
db -createdb <user><password>[database_name]
```

如果不指定数据库名，则使用 gcs。一旦创建了数据库，通过发出以下命令来添加 IBM Web Crawler 表:

```
db -createtables<user><password>[database_name]
```

IBM Web Crawler 数据库和表的创建必须是完整的，以与 IBM Web Crawler 一起使用 DB2。

需要以下配置文件设置（在 urlpool-config 部分中）以使用新数据库: dbname:

- 数据库名称（如上面所创建的）: 例如 gcs。
- 用户名: 用户的名称，例如: db2admin
- 密码: 用户的密码，例如: db2admin。

将数据库、用户名和密码特性设置为合适的值。不要更改高速缓存大小或驱动程序。继续编辑文件以为系统设置搜寻范围。

设置搜寻范围

要建立搜寻范围需要这些配置文件设置，不管是否使用 DB2。

检查 crawler-config 部分中的以下设置，并根据需要相应地设置这些条目。

seed list

一个或多个起始绝对 URL。该 URL 必须是可用的。使用浏览器进行验证，例如: `http://www.<mysite>.com/`

content-type-pattern-list

仅当有文件扩展名与这些模式相匹配时，才搜寻页面上查找到的 URL，例如: `htm*`

include-pattern-list

仅当它们与这些模式相匹配时，才搜寻页面上查找到的 URL，例如: `<mysite>.com`

还可以设置这些条目:

recursion-depth

从任意开始点进行搜寻的链接间的最大距离。对于无限制深度使用 -1。

exclude-pattern-list

仅当它们与这些模式不匹配时，才搜寻页面上查找到的 URL，例如: `*cgi-bin*`

system properties

要通过防火墙从未 socks 化过的机器进行搜寻，还需要在此文件中设置 `socksProxy` 值。

启动 IBM Web Crawler

如果编辑了 `.xml` 配置文件，则保存该文件。

要启动 IBM Web Crawler，使用 `crawlweb` 批处理文件和配置文件。打开命令提示符并输入:

- 对于 Windows: `crawlweb.bat<CONFIGFILE>`
- 对于 AIX: `crawlweb.sh<CONFIGFILE>`

要与 DB2 UDB 一起运行，则输入: `crawlweb config-db2.xml` 并按下 Enter 键。若不与 DB2 UDB 一起运行，则输入: `crawlweb config-sample.xml` 并按下 Enter 键。

技巧: 计划定期地报告搜寻 / 摘要过程。当目标已搜寻时，将摘要写入在 summaries-dir 中配置的位置。缺省的摘要器将原始对象加上元数据起始标记写为树中的 .html 文件。在搜寻期间或搜寻之后，可以检查日志文件以获取其它信息。

高级配置

在此处可以学习配置选项。对于配置示例，请参阅第 91 页的第 7 章，『IBM Web Crawler 样本文件』中的 `config-sample2.xml` 文件。样本说明了以下内容的配置:

- 网上搜寻器和摘要器线程
- 图形监视器
- 记录选项
- SOCKS
- Lotus Domino 搜寻
- 多种内容类型
- 更多排除项
- 使用 `InsoSummarizable` 获取对象（如 `.pdf` 文件）的摘要

请参阅 config.dtd 文件获取配置文件中可用参数的正式定义。**建议**: *请勿*编辑此文件。制作该文件的副本并且重命名副本。

IBM Web Crawler 配置文件

该配置文件是一个 XML 文件，它告知 IBM Web Crawler 要收集哪些基于 Web 的资源以及如何摘要它们。本节描述了可以在 config.xml 中设置的每个元素和属性。关于如何使用用于 Notes 的 IBM Web Crawler 的信息，请参阅第 77 页的『IBM Web Crawler for Notes』。

IBM Web Crawler 检查配置文件的内容是否与 gcs-config.dtd 一起编译。如果有重大错误，例如，没有要搜寻的 URL，那么 IBM Web Crawler 会退出并打印错误消息。对于小问题（未知属性或值），该程序会将警告记录到日志文件并继续执行。**建议**: 编辑之前请备份配置文件。文件中的错误会中断 IBM Web Crawler。

IBM Web Crawler 带有样本配置文件。

<gcs-config>

gcs-config 文件包含两个部分: **globals** 和 **group-list**。请参阅第 91 页的第 7 章，『IBM Web Crawler 样本文件』获取 gcs-config 文件的示例。

globals

globals 元素捕捉 IBM Web Crawler 的设置，如文件系统、性能和网络信息。

group-list

group-list 元素配置组的搜寻与摘要，其中组是诸如商业或网络域之类资源的集合。

<globals>

globals 元素表示 IBM Web Crawler 的全局设置。该设置编码为全局属性和子元素。

以下列表定义了全局属性。关于全局子元素的定义，请参阅第 69 页的『<logger-config>』。

max-urls

要搜寻的 URL 的最大数目。它应当为正整数，缺省值为 100000。

summaries-dir

在其中写入资源摘要的目录。缺省情况下使用 summaries/ 目录。

summaries-filepool-class

用于资源摘要的文件池的类型。这确定了摘要文件如何命名以及使用什么样的子目录结构（如果存在）。缺省情况下使用 FullPathFilePool，它为主机创建了一个目录，然后使用与 URL 相同的子目录结构和文件名。

num-crawlers

要使用的网上搜寻器线程数。它应当为正整数，缺省值为 20。

num-summarizers

要使用的摘要器线程数。它应当为正整数，缺省值为 5。使用这些步骤配置 num-crawlers 和 num-summarizers:

1. 将网上搜寻器设置为以 MHz 为单位的机器速度的 1/20。例如，在 600 MHz 的系统上，使用 *30*。
2. 将摘要器的数量设置为步骤 1 中数值的 1/4，例如，*8*。

3. 进行试运行时留意 “Windows 任务管理器性能” 面板。如果 CPU 经常不止一秒处于 100%，则请返回步骤 1 并设置一个较小数值，例如，上次数值的 3/4，直到 CPU 基本上不再处于 100% 为止。

   如果试验期间，您发现文本监视器总是报告摘要器以下消息：ToDo 数值恰好小于已配置的摘要器数目，则您可以在不违反步骤 3 的情况下减少摘要器的数目（越少越好）并增加网上搜寻器的数目（越多越好）。要获得最佳性能，请使用您可以获得的最快网络并尽可能跨多个独立磁盘散布摘要、数据库、临时空间和日志。

text-monitor

设置为 on 时，text-monitor 会每 5 秒一次将 IBM Web Crawler 的状态打印到标准输出。给出小数值时，text-monitor 设置文本输出的刷新之间的时间（以秒为单位）。缺省设置为 off。

graph-monitor

设置为 on 时，graph-monitor 会通知 IBM Web Crawler 使用图形 GUI 显示其状态。给出小数值时，graph-monitor 设置监视器 GUI 的刷新之间的时间（以秒为单位）。缺省设置为 off。

log-file

指定要使用的主日志文件。缺省值为 `log/log.txt`。

技巧：您可以在 logger-config 元素中指定附加记录器信息。

log-priority

设置缺省日志优先级。输入信息的值：warn 或 error。缺省值为 warn。

技巧：您可以在 logger-config 元素中指定附加记录器信息。

temp-dir

在其中写入临时文件的目录。**技巧**：此目录中的所有文件都可以由 IBM Web Crawler 删除。您应该不需要将其从缺省设置 *x*`:/temp/gcs` 更改为其它值。

temp-filepool-class

要对临时文件使用的文件池的类型。**建议**：请勿将其从缺省设置 `TempFilePool` 更改为其它值。

content-dir

IBM Web Crawler 写入内容文件的目录。通常，`content-dir` 与 `temp-dir` 相同。

content-filepool-class

要对内容文件使用的文件池的类型。通常，它与 `temp-filepool-class` 相同。

how-often-to-gc

请求碎片收集之间要搜寻的 URL 数。**建议**：请定义一个大于等于 50 的整数。缺省设置为 100。

max-resource-pool-size

等待进行摘要的资源的最大队列大小。**建议**：请定义一个大于等于 10 的整数。缺省设置允许每个摘要器具有 10 个等待资源。

connect-timeout

定义在网络上连接超时之前等待的毫秒数。缺省值为 4000。有效范围为 1000-60000。

read-timeout

定义在网络上读取超时之前等待的毫秒数。缺省值为 6000。有效范围为 1000-60000。

cookies

定义是否检查 HTTP 头中的 cookie 以及是否将它们存储在数据库中。缺省设置为 off。可以通过将该值设置为 on 来启用 cookie。

locale 定义用于摘要和记录日志的语言。缺省值为 en_US。

全局子元素包括 logger-configs、urlpool-config 和 system-properties。

<logger-config>

`logger-config` 文件提供了一些高级控制，包括记录的内容、格式化方式以及日志文件写入的位置。缺省 `log-file` 和 `log-priority` 指定为全局属性。关于记录日志的更多信息，请参阅第 74 页的『登录 IBM Web Crawler』。

category

正在配置的记录器的类别，例如，`gcs.crawler`。如果不指定，则配置缺省记录器。请记住，对特殊类别的设置会影响所有子类别。

priority

记录消息时所必须具有最小优先级。如果不指定，则此记录器会从其父类别（最后从全局缺省 log-priority）获取优先级。

log-file

定义日志文件写入的位置。如果它以 “+” 开始，则除了任何其它（父）日志文件外还将使用此日志文件。如果不指定，则使用父日志文件（最后使用全局缺省 log-file）。

提示: 请注意不要对多个记录器使用相同的 log-file，因为它们将彼此覆盖。

log-layout

定义用于打印到日志文件的每个消息的布局。

<urlpool-config>

`urlpool-config` 文件配置存储 URL 的 IBM Web Crawler 的组件。URL 池具有多个选项。您可以将该池存储在内存中，可以使用 DB2 或不为每个 URL 存储许多信息的特殊小内存版本。如果不指定 `urlpool-config` 元素，则 URL 池存储在内存中。`urlpool-config` 可以具有子 `urlpool-param` 元素，用于指定诸如数据库信息的事项。

urlcontainer-class

要使用的 URL 容器的类型。指定:

- 利用 DB2 UDB 搜寻的 DB2URLContainer
- 不利用 DB2 UDB 搜寻的 MemoryURLContainer（缺省）。
- 不利用 DB2 UDB 搜寻并使用附加内存（存储一些引用的 URL 和其它信息）的 BigMemoryURLContainer。

urlcollection-class

要使用的 URL 集合的类型。指定:

- 利用 DB2 UDB 搜寻的 DB2URLCollection
- 不利用 DB2 UDB 搜寻的 MemoryURLCollection（缺省）。

- 不利用 DB2 UDB 搜寻并使用附加内存（存储一些引用的 URL 和其它信息）的 BigMemoryURLCollection。

<urlpool-param>
用以将参数传递到 urlcollection-class。例如，请参阅第 91 页的第 7 章，『IBM Web Crawler 样本文件』中使用 DB2 UDB 的样本配置中的数据库连接信息。

name 定义参数名称。

value 定义参数值。

技巧: 使用这些参数时请小心，因为它们没有错误检查。

<system-properties>
系统特性代表一列系统特性设置。

<property>
例如，请参阅高级配置样本中使用 SOCKS 网关的配置。

name 参数名称。

value 参数值。

作为选择，您可以配置 IBM Web Crawler 使用以下方法通过 PROXY 网关访问外部服务器:

```
<system-properties>
  <property name="proxySet" value="true"/>
  <property name="proxyHost" value="proxy.hostname"/>
  <proxy port name="proxyPort" value="80"/>
</system-properties>
```

技巧: 这些参数上没有错误检查，因此使用时请小心。

<group-list>
group-list 是一个或多个组元素的列表。

<group>
组元素表示单独一组将以相似方式搜寻与摘要的资源。每个组必须具有一个唯一名称属性以及至少一个 crawler-config 子元素（通知它要搜寻的内容）。如果不想使用缺省摘要器，则组可以具有一个子 summarizer-config 元素。**提示:** 覆盖组（两个或多个组中具有同一 URL）会导致意外结果。多重分组的 URL 仅与在其中找到它们的第一个组相关联。

为此组**命名**一个唯一名称（必需的）。

<crawler-config>
使用这些规则设置搜寻范围。网上搜寻器检索 seed-list 中的每个 URL、从它们的内容分析 URL 并将与以下内容匹配的 URL 添加到 to-be-crawled 列表:

- content-type-pattern-list 中的至少一个规则
- include-pattern-list 中的至少一个规则
- exclude-pattern-list 中的零个规则。

crawler-config 还需要一个单独的属性: **recursion-depth**。recursion-depth 定义了网上搜寻器可以从每个种子遍历的链接数范围。缺省值为 -1，表示无限深度。

<seed-list>
这是 URL 种子的列表，可能带有认证信息。

<seed>
Seed 代表启动网上搜寻器的 URL 种子，它具有 URL 属性且可能带有认证信息。每个种子必须是一个绝对 URL，例如，`http://<your.server>.com/`。请避免重定向、不可用或指向非正文页面的种子。指向已编辑为包含种子的页面是有用的。这样的页面很容易使用浏览器更新、复查以及测试。

URL 用于启动搜寻的种子 URL。

<authentication>
如 rfc2617 中所定义，为受*基本认证*保护的种子 URL 发送的可选认证。

username

用于认证的用户名。

password

用于认证的密码。

例如:

```
<seed url="http://your.server.com/"><authentication username="me"
password="mine"/></seed>
```

<content-type-pattern list>
这是模式列表，用于包含以文件扩展名标识的要搜寻的内容类型。与此列表中任意 url-name-pattern 匹配的所有 URL 文件扩展名（.html、.gif、.doc 等等）都将通过测试。不具有扩展名的 URL 在缺省情况下通过测试。如果 `content-type-pattern-list` 未指定或为空，则仅接受不具有文件扩展名的 URL。

<include-pattern list>
这是包含要搜寻（例如，按服务器或域名）的 URL 的模式列表。与此列表中的任意 `url-obj-pattern`、`url-regex-pattern`、`url-name-pattern` 或 `url-predicate-pattern` 匹配的所有 URL 都会通过此测试。如果 `include-pattern-list` 未指定或为空，则接受所有 URL。

<exclude-pattern-list>
这是要从正在搜寻的内容排除 URL 的模式列表。与此列表中的任意 `url-obj-pattern`、`url-regex-pattern`、`url-name-pattern` 或 `url-predicate-pattern` 匹配的所有 URL 都不会被搜寻。如果 `exclude-pattern-list` 未指定或为空，则不拒绝任何 URL。

<url-obj-pattern>
这是将 URL 的不同部分（协议、主机等等）与通配符相匹配的模式。它可以在 exclude-pattern-list 和 include-pattern-list 中使用。每个部分的模式可以在开头和/或末尾具有 “*” 通配符，该通配符与所有内容都匹配。但是，任意模式的中间位置都不能具有通配符。该匹配是不区分大小写的。所有省略的 URL 部分模式都会自动匹配。

以下列表包含了一个 Java 和 IBM Web Crawler 如何分解 URL http://www.ibm.com/products/index.html?query#ref 的示例:

- 协议为: http
- 主机为: www.ibm.com
- 端口为: -1（未指定）
- 文件为: /products/index.html?query
- 路径为: /products/index.html
- 目录为: /products/
- 文件名为: index.html
- 扩展名为: .html
- 查询为: query
- 引用为: ref

以下列表提供了 url-obj-pattern 的每个元素的更多详细信息:

protocol
URL 协议必须匹配的通配符模式，例如，http。

host URL 主机必须匹配的通配符模式，例如，*.ibm.com

port URL 端口必须匹配的通配符模式，例如，80。

file URL 文件必须匹配的通配符模式，例如，*.htm*。URL 的文件部分从主机之后的第一个斜杠开始，可以包含查询，但不能包含引用。http://www.ibm.com/products/index.html?query#ref 的*文件*部分为 /products/index.html?query。

path URL 路径必须匹配的通配符模式，例如，*.html。URL 的路径部分从主机之后的第一个斜杠开始，不包含查询或引用。在该示例中 http://www.ibm.com/products/index.html?query#ref 的*路径*部分为 /products/index.html

dir URL 中的目录必须匹配的通配符模式，例如，/products/。目录是路径中从第一个斜杠开始到最后一个斜杠结束的部分。在该示例中，http://www.ibm.com/products/index.html?query#ref 的*目录*部分为 /products/。它不包含查询或引用。请注意，遗漏了最后一个斜杠的错误的 URL（例如，http://www.ibm.com/products）将无法与目录正确匹配。在错误 URL 的示例中，URL 目录为 /。

filename
URL 中的文件名必须匹配的通配符模式，例如，index.html。文件名是路径中跟着最后一个斜杠的部分。在该示例中，http://www.ibm.com/products/index.html?query#ref 的*文件名*部分为 index.html。它不包含查询或引用。

extension
URL 文件扩展名必须匹配的通配符模式，例如，htm*。可能时首选使用 content-type-pattern-list。

query URL 查询必须匹配的通配符模式。

ref URL 引用必须匹配的通配符模式（不用于 HTTP）。例如，<url-obj-pattern host="*.ibm.com"/> 将与所有 IBM 站点上的 HTML 页面相匹配。

<url-regex-pattern>

url-regex-pattern 是与使用规则表达式的 URL 匹配的模式。它既可用于 exclude-pattern-list，也可用于 include-pattern-list。它使用 com.ibm.regex 包（regex4j），且具有 Perl 5 规则表达式的大多数功能性。它可以具有两个规则表达式，其中一个是 URL *必须*匹配的，另一个是 URL *不能*匹配的。还可以指定其它选项，如 i 指不区分大小写。请参阅 Regex4j 规则表达式获取详细信息。

match URL 必须匹配的 Perl 5 样式规则表达式。

no-match

URL 必须不匹配的 Perl 5 样式规则表达式。

options

可选的修饰成分，如 i 表示不区分大小写。

例如，<url-regex-pattern match="^http://www\.ibm\.com/.*\.html?$"/> 将与主 IBM Web 站点上的 HTML 页面匹配。

<url-name-pattern>

这是一种简单模式，具有与全 URL 或 URL 文件扩展名匹配的通配符。它可以在 content-type-pattern-list、include-pattern-list 和 exclude-pattern-list 中使用。它可以在模式字符串的开头和／或末尾具有与所有内容匹配的 “*”。但是，名称的中间不能具有通配符。该匹配是不区分大小写的。

但是，<url-name-pattern name="*.ibm.com/*"/> 将与 IBM 站点上的所有文件匹配，而 <url-name-pattern name="*.ibm.com/*.html"/> 无效，因为中间位置存在通配符。

URL 字符串必须匹配的 **name** 通配符模式，在开头和／或末尾带有可选的 “*” 通配符。

<url-predicate-pattern>

此模式将装入用于与 URL 匹配的 Java UnaryPredicate 类。它可以在 exclude-pattern-list 或 include-pattern-list 中使用。该类必须具有 public boolean execute(URL url) 方法，如果 URL 与谓词匹配，则返回真。

class 全限定 UnaryPredicate 类名。

<summarizer-config>

这是摘要器的配置，带有子 resource-handlers 的列表。当前每组只能具有一个 summarizer-config。

<resource-handler>

根据资源的内容类型，例如（text/html）或文件名扩展名（htm）确定对资源（如 Web 页面或新闻组文章）执行的摘要的类型。资源准备好进行摘要时，IBM Web Crawler 会依次检查 resource-handler，并使用第一个与内容类型或文件扩展名匹配的 resource-handler。如果都不匹配，缺省情况下将使用 Copy2RdfSummarizable 和 Copy2RdfSummaryMaker。您可以通过在列表底部添加一个不带有 content-type 或 file-extension 的 resource-handler 来覆盖它。

resource-handler 还可以拥有 summarizer-param 子模式，它可将特殊参数传递到其 SummaryMaker 类。

content-type
资源的 content-type 必须匹配的通配符模式，例如: *htm*

file-extension
资源的文件扩展名必须匹配的通配符模式，例如，htm*

summarizable
资源 Summarizable 类名，例如: HtmlRawSummarizable

summary-maker
资源 SummaryMaker 类型，例如: HtmlRawSummaryMaker

content-type 和 file-extension 模式允许通配符。模式可以在模式字符串的开头和/或末尾具有 *，它可与所有内容匹配。但是，中间不能具有通配符。该匹配是不区分大小写的。

如果 content-type 和 file-extension 模式都匹配，则 resource-handler 将匹配，且未指定模式将总是匹配，则

```
<resource-handler content-type="*htm*"
summarizable="*HtmlRawSummarizable" summary-maker=
"HtmlRawSummaryMaker"/>
```

将与内容类型为 text/html 的所有文件相匹配，不管它们的文件扩展名是什么。

对于 summarizable 和 summary-maker，如果类处于 com.ibm.IBM Web Crawler.summarizer.resource 包中，则不必指定其全路径。

<summarizer-param>
这些是传递给 SummaryMaker 类的特殊参数。用法特定于该类。

name 参数名称。

value 参数值。

技巧: 这些参数上不能存在错误检查，因此使用时请小心。

登录 IBM Web Crawler
这是 IBM Web Crawler 记录日志功能的介绍。

IBM Web Crawler 提供了对记录内容、记录位置以及格式化方式的有力控制。例如，您可以选择将搜寻到的每个页面的响应代码写入一个文件，将 IBM Web Crawler 状态（搜寻的 URL 数、工作的线程数等）写入另一个文件，将摘要的 URL 写入第三个文件，将所有 IBM Web Crawler 警告写入第四个文件，并将网络实用程序软件包中的所有日志消息写入另一个文件以进行调试。

请参阅第 93 页的『IBM Web Crawler 日志分析文件示例』获取日志分析文件的样本。

日志的使用
记录日志对于网络/Web/搜寻/摘要记帐、与其它应用程序组件进行通信以及 IBM Web Crawler 调试都很有用。

搜寻和发掘记帐可以显示您感兴趣的广泛范围的功能部件，例如错误配置的服务器、缺少的页面和每种内容类型的对象数。loganalysis.pl Perl 脚本提供了日志摘要记帐的样本。应用程序可能需要来自 IBM Web Crawler 的信息，如何时已除去内容。

配置记录器

您可以在 IBM Web Crawler 配置文件中指定一个或多个记录器的配置。全局元素的 log-priority 和 log-file 属性会建立缺省的记录日志策略。

要扩展记录日志策略，请创建 logger-config 语句作为全局元素的子元素。每个语句选择 IBM Web Crawler 日志消息的一个子集，将它们转送到一个特殊文件并使用特殊格式编写它们。所记录消息的子集是使用优先级和类别属性选定的。合法优先级值为: trace、debug、info 和 warn（不区分大小写）。

- 设置优先级的值可以确定 trace 的详细程度最高时记录器的详细程度。
- trace 和 debug 是维护级别 — 消息是以英语硬编码的。
- info 和 warn 是用户级别，支持本地语言。
- info 会产生很多消息。通过指定 warn 优先级可以减少消息输出。

记录日志配置示例

不带日期 / 时间 / 线程信息，将源 / 目标超链接记录到文件 log/fromto.txt

```
<logger-config category="gcs.url.fromto" priority="info"
 log-layout="%m\n"log-file="log/fromto.txt"/>
```

将进行了摘要的对象记录到文件 log/resources.txt

```
logger-config category="gcs.summaries.list.resource"
 priority="info" log-file="log/resources.txt"/
```

记录跳过的 URL 和跳过的原因

```
<logger-config category="gcs.url.skipped"
priority="info" log-file="log/urls_skipped.txt"/>
```

记录特殊处理的 HTTP 响应代码

```
<logger-config category="gcs.http.302"
priority="info" log-file="log/urls_redirected.txt"/>

<logger-config category="gcs.http.404"
priority="info" log-file="log/urls_not_found.txt"/>
```

记录所有摘要器类别的消息，包括它们的优先级

```
 <logger-config category="gcs.summarizer"

priority="TRACE" log-file="summarizer_trace.txt"
log-layout="%d: %t: %c: %p: %m\n"/>
```

疑难解答

如果发生问题，首先要检查的事项如下:

种子列表上的页面是否可及?

页面必须存在（避开重定向的种子），且必须可从您的系统（并且通过 SOCKS，如果使用了 SOCKS）访问。

种子列表上的页面是否为常规 HTML?

框架、Flash、javascript 以及其它这样的项不是作为种子的良好选择。请选择常规 HTML 页面。

如果使用 DB2 UDB，您是否已经进行了搜寻?

DB2 UDB 会对您搜寻的内容保持跟踪。如果所有页面都已处理，则 DB2 UDB 无操作。使用 db -emptytables 命令启动新搜寻。

如果使用 DB2 UDB，配置文件的数据库访问条目是否正确?

如果数据库连接失败，则搜寻也将失败。

仔细检查过配置文件编辑了吗?

错误会终止 IBM Web Crawler。是否忽略了限制性的 max-urls 或 recursion-depth 值？

仍有问题吗?

编辑配置文件并将日志优先级更改为 “debug”。现在再次启动网上搜寻器，然后在它停止之后，检查日志文件。

选择摘要器

摘要器的用途在于获取资源（如 Web 页面）或主机（如 Web 服务器）并产生包含您感兴趣的信息的文件（以易使用的格式）。

IBM Web Crawler 包含各种摘要器，用于处理不同内容类型、从资源抽取各种类型的数据并以不同文件格式输出。本节描述了可用摘要器的功能部件和需求。如果所有这些摘要器都不具有您期望的功能，则您也可以编写自己的摘要器。

IBM Web Crawler 中有两种摘要器。资源摘要器产生单个资源（如 Web 页面）的摘要；主机摘要器产生主机（如 Web 服务器）的摘要。目前，只有资源摘要器是可配置的。

选择资源摘要器

选择或编写资源摘要器时要考虑的事项如下:

- 输入格式是什么？HTML Web 页面、PDF、WordPro 文档、XML 文件
- 要抽取哪些元数据？HTTP 头、标题、已标注链接，主体文本
- 期望的输出格式是什么？XML、HTML、RDF

您可以使用 IBM Web Crawler `config` 文件中的 resource-handler 元素指定要用于特殊类型资源的摘要器。首先指定摘要器用于的内容类型和 / 或文件扩展名。然后指定执行操作的 Java Summarizable 和 SummaryMaker 类。summarizable 类表示要进行摘要的资源；summary-maker 类表示将产生的摘要的类型。

缺省摘要器（复制 + RDF 摘要器）

复制 + RDF 摘要器是应用于带有某种 content-type 的任何对象的摘要器，该 content-type 未明确配置为由另一摘要器进行处理。本摘要器可由任意类型的资源使用，它将编写两个文件。第一个文件时原始资源的精确副本，第二个文件是包含原始 URL、存储文件的文件名和 HTTP 头信息的 RDF 摘要。它还可以使用 `DefaultSummarizable` 和 `Copy2RdfSummaryMaker` 明确配置。

HTML 页面的摘要器（原始 HTML 摘要器）

对于 HTML 资源，原始 HTML 摘要器仅产生文件的副本，顶部注释中包含 URL 和 HTTP 头信息。它是使用 `HtmlSummarizable` 和 `HtmlRawSummaryMaker` 配置的。

```
<resource-handler content-type="*htm*"
                      summarizable="HtmlSummarizable"
                      summary-maker="HtmlRawSummaryMaker" />
```

HTML 页面的摘要器（EIP HTML 摘要器）

对于 HTML 资源，原始 HTML 摘要器仅产生文件的副本，顶部注释中包含 URL 和 HTTP 头信息。它是使用 EIPHtmlSummarizable 和 EIPHtmlRawSummaryMaker 配置的。

```
<resource-handler content-type="*htm*"
                  summarizable="EIPHtmlSummarizable"
                  summary-maker="EIPHtmlRawSummaryMaker" />
```

HTML 页面的摘要器（非写 HTML 摘要器）

此摘要器将搜寻 HTML 并跟随链接，但不写入文件的任何摘要。有时这会有用，例如，如果您希望搜寻站点上的所有 PDF 文件（使用 INSO 到 XML 摘要器）但不存储 HTML 文件。它是使用 HtmlSummarizable 和 NoWriteSummaryMaker 配置的。

```
<resource-handler content-type="*htm*"
                    summarizable="InsoSummarizable"
                    summary-maker="InsoSummaryMaker" />
```

其它内容类型的摘要器（INSO 到 XML 摘要器）

此摘要器将为多于 200 种类型的资源（如 Microsoft Word 文档、PDF 文件、PowerPoint 演示文稿及其它资源）创建 XML 摘要。它含有一些元信息，且主体文本是由网络解决方案 INSO 过滤器（需要 INSO 许可证）抽取的。它是使用 InsoSummarizable 和 InsoSummaryMaker 配置的。

```
<resource-handler content-type="pdf"
                    summarizable="InsoSummarizable"
                    summary-maker="InsoSummaryMaker" />
```

其它摘要器

如果需要对其它资源类型进行摘要、发掘其它数据或以其它格式输出，请联络 IBM 或创建定制摘要器。

IBM Web Crawler for Notes

本节描述了如何配置和运行 IBM Web Crawler for Notes。IBM Web Crawler for Notes 访问 Notes 数据库并创建 Notes 文档及附件的摘要。摘要为 XML 格式文件，每个文档或附件一个，包含对象和全文本。

先决条件

以下先决条件是运行 IBM Web Crawler for Notes 之前所必需的:

- Lotus Notes V5.0.5 或更新版本。
- PKZIP V2.50，如果您希望处理文件类型为自解压 zip 文件的附件文件。

执行测试搜寻

请选择**开始 → 程序 → 命令提示符**。在新窗口中，将目录切换到安装 IBM Web Crawler for Notes 的位置然后切换到 notes-run 子目录。例如:

```
cd c:\<install directory>\gcs\notes-run
```

Notes 搜寻受控于两个文件:

- 对其进行编辑以标识您可以搜寻的 Notes 数据库的源列表。它包括 Notes 服务器名称、IP 地址、.nsf 文件名等等。例如，源列表可以命名 34 个 Notes 数据库，搜寻的数据库是在如下所述的配置中建立的。资源列表可以是 .xml 文件或 Notes 数据库（.nsf 文件）。

- 指定源列表、搜寻哪些源、处理哪些附件类型、输出格式等等的配置文件。配置文件总是 .xml 文件。

要验证 IBM Web Crawler 是否已经正确安装，请搜寻其测试数据库。使用编辑器确保 testSources.xml 源列表具有到 test.nsf 数据库的正确路径和文件名，它处于安装 IBM Web Crawler 的 notes-run 子目录中。制作原始文件的备份副本。**建议**: *请小心编辑*: 文件中的错误会导致 IBM Web Crawler 失败。保存所有更改。

通过搜寻包含的 test.nsf 数据库测试安装。输入: crawlNotes crawlTestXml

crawlNotes.bat 文件以 crawlTestXml 作为其配置文件启动 IBM Web Crawler; .xml 会自动追加到配置文件名。IBM Web Crawler 应报告搜寻情况并摘要两个文件，每个文件都带有附件。

IBM Web Crawler 完成时，您可以查看摘要目录中的摘要，并在配置文件指定的日志摘要目录中搜寻的日志文件。

配置定制 Notes 搜寻

成功测试搜寻之后，您会希望搜寻其它数据库。

1. 创建数据库源列表。将要搜寻的 Notes 数据库添加到源文件。

   要在 XML 文件中标识要搜寻的 Notes 数据库，请通过编辑 testSources.xml 文件开始操作。要在 Notes 数据库中标识要搜寻的 Notes 数据库，请使用 Notes 打开并更新 testSources.nsf Notes 数据库。“编辑源列表”中解释了能在源文件中设置的参数。
2. 设置网上搜寻器配置。您将需要以 XML 格式编辑配置文件。
   - 如果您的源列在 HML 文件中，则请通过编辑 crawlTestXml.xml 并设置 sourcesInXmlFile 使其指向源文件来开始操作。
   - 如果您的源处于 Notes 数据库中，则请通过编辑 crawlTestNsf.xml 并设置 sourcesInNotesDB 使其指向源数据库来开始操作。第 67 页的『IBM Web Crawler 配置文件』中解释了能在配置文件中设置的参数。

一旦源列表与配置完成，即请调用 IBM Web Crawler: crawlNotes your_config

或者，如果您的源列表处于 Notes 数据库中，则启动 IBM Web Crawler（如在以下工作示例中所示）: crawlNotes crawlTestNsf

IBM Web Crawler Notes 网上搜寻器完成时，您可以查看摘要目录中的摘要，并在配置文件中指定的日志摘要目录中搜寻日志文件。

源列表参数

源列表包含可以搜寻的 Notes 数据库的描述。file.xml 格式中的源列表包含 notesDataSources 元素，带有一个或多个 oneDBinfo 元素。每个 oneDBinfo 元素包含:

id 此数据库的数字标识。配置文件中的范围参数引用该标识。

serverName
为数据库服务的服务器的名称。请对本地数据库使用空字符串 ""。

pathAndFileName
服务器上的数据库的全路径和文件名。请以 .nsf 结束路径和文件名。

viewName

要搜寻的数据库的 Notes 视图的名称。

ipAddress

可选。服务器的 IP 地址；如给出，则不使用 DNS。如果 DNS 不能解析目标服务器名称，则您可以在此处指定其 IP 地址。在 Windows 上，IP 地址可以使用 `nslookup server_name` 命令确定。

dateLastCrawled

可选。上次搜寻数据库的日期。除非您在配置文件中将更新最近搜寻的日期设置为 no，否则它都会被自动修改。

tries 可选。如果搜寻不成功（超时），您希望尝试重新搜寻数据库的次数。

fieldSubstitutions

指定 Notes 数据库字段名称在输出 XML 文档中如何被替换的映射。它包含一个或多个替换元素，每个元素具有两种属性:

- Original: 输出 XML 文件中将被替换的字段名称（如果存在）
- Replace: 用将在输出 XML 文档中替换原始字段名称的新的字段名称。

可以使用 Notes 客户机检查和更新 Notes 数据库中的资源列表。启动 Notes 并选择**文件 —► 数据库 —► 打开**。单击**浏览**以定位并打开 *x*:\<install directory>/gcs/notes-run 中的 testSources.nsf 数据库。也可以使用 Notes 客户机检查和更新附带的测试数据库。启动 Notes 并选择**文件 —► 数据库 —► 打开**。单击**浏览**以定位并打开 *x*:\<install directory>/gcs/notes-run 目录中的 test.nsf 数据库。

配置文件参数

配置文件中可以设置的参数在以下描述。可以忽略缺省列出的参数。

`sourcesInXml` 文件或 `sourcesInNotesDB` 元素标识源列表。源可以分别是包含标识要搜寻的数据库的信息的 XML 格式的文件或 Notes 数据库。

包含控制单个 `run` 的参数的 `runInfo` 元素。即，这些参数应用于以给定的网上搜寻器用法所搜寻到的 Notes 数据库:

rangeSpecify

要搜寻的数据库的标识。这些标识是在源列表的标识字段中给出的数值。它可以指定为以逗号分割的独立标识和 / 或连字符指示的范围的列表，如 1-4、15、25-31。

SummaryDirectory

指定输出摘要的根目录。摘要被写入此目录的子目录中。

MaxThreads

指定并行搜寻线程的数目。每个 Notes 数据库都由一个单独的线程进行搜寻。多个数据库可以并行搜寻。

doIncrementalCrawl

缺省值为 no。如果为 yes，则仅处理自 `'summarizeThisDateAndLater'` 以来新的 / 修改过的 Notes 文档。如果未指定 `'summarizeThisDateAndLater'`，则网上搜寻器将使用每个数据库自己的 `DateLastCrawled` 字段（在数据库源列表中指定）。如果 `doIncrementalCrawl=no`，则所有文档都将被处理，而不考虑它们的日期。

summarizeThisDateAndLater

此字段的格式为: MM/dd/yyyy hh:mm a tz, 例如, 01/01/2000 01:11 PM PDT。如果未给出日期和时间, 则对源列表中记录的上次搜寻以来（如果 doIncrementalCrawl 设置为 yes）或一直以来（如果 doIncrementalCrawl 设置为 no）的所有文档。

detachAttachments

缺省值为 yes。如果为 yes, 则拆离附件并对其进行摘要。要处理的附件文件的类型在配置文件中列出。如果为 no, 则将忽略附件。

attachmentFilenameFormat

缺省值为 l（长）。它也可以为 s（短）。长文件名对类型、服务器、数据库名称和 Notes 标识进行编码。短文件名对类型和 Notes 标识进行编码。

processAttachmentsAfterwards

缺省值为 no。如果为 yes, 则 Notes 数据库摘要期间将不对附件文件进行摘要。而是为每个附件将一条记录写入 notesCrawl-attachments.bat, 该记录指定用于摘要文件的命令。之后您编写并运行批处理文件, 以对附件进行摘要然后删除附件。过后处理附件通常需要充足的磁盘存储量。

saveAttachmentFiles

缺省值为 no。如果为 yes, 则原始附件文件在处理之后不会被擦除。此选项仅在 processAttachmentsAfterwards 为 no 时有效。如果 processAttachmentsAfterwards 设置为 yes, 则需要磁盘空间来存储保存的附件。

MaximumNumberOfDetachingErrors

缺省值为 10。处理附件时的最大错误数, 例如, 保存附件文件时磁盘空间不足, 在异常终止搜寻之前网上搜寻器将保持现状。

saveURLsToFile

缺省值为 no。如果指定该值, 则 Notes 文档项中找到的 URL 将用以下格式的名称写入文件:

databasename(不带有路径和 .nsf) + ".html".

updateDateLastCrawled

缺省值为 yes。如果为 no, 则不更新源文件中的 dateLastCrawled。

tempDirectory

缺省值为 c:\temp。此目录用于写入所有临时文件。

logSummaryDirectory

缺省值为 log。保存日志文件的目录。

loggerPriority

缺省值为 info。这些设置定义了记录器的优先级。该设置从高到低可以为 error、warn 或 info。例如, 如果记录器的优先级设置为 warn, 则只将记录具有 warn 和 error 优先级的日志消息。

包含 include 元素（标识会被处理的附件文件扩展名）的附件元素, 例如 .prz。

从服务器排除 IBM Web Crawler

由于安全性和性能原因, EIP 管理员可能要排除搜寻某些服务器或页面。您可能需要能够限制在服务器和页面上的浏览器活动。

您可以指导 IBM Web Crawler 以避免服务器或页面使用*访问策略*文件。此文件是按照 *A Standard for Robot Exclusion*（请参阅 http://info.webcrawler.com/mak/projects/robots/norobots.html）中发布的指南构建的。

- 搜寻服务器并且在这之后的周期性搜寻之前，IBM Web Crawler 将请求访问策略文件 http://*yourserver*/robots.txt。
- 文件由以下格式的各行组成

  field:<optionalspace>*value*<optionalspace>

  如果 *field* 是 User-Agent，并且 *value* 是 IBM-WebCrawler 或 *，则以下 Disallow 行将（通过下一 User-Agent 行）指定要避开的部分地址。这可以是完整的路径或部分路径，任何以该值开头的地址将不被检索。

  例如:

  ```
  Disallow: /help
  ```

  不允许 /help.html 和 /help/index.html。

  ```
  Disallow: /help/
  ```

  不允许 /help/index.html，但允许 /help.html。

  如果是空值，则允许检索所有地址。
- 这些行可用空行分隔。
- 您可通过输入 # 字符来包含注释。余下的行将被认为是注释。

这是一些示例:

- 此 /robots.txt 指定所有 robot 都应当避免此服务器。

  ```
  # disallow everybody

  User-agent: *
  Disallow:   /
  ```
- 此 /robots.txt 指定只有 IBM Web Crawler 可以搜寻此服务器，并且对网上搜寻器没有限制。

  ```
  # allow only IBM

  User-agent: *
  Disallow:   /
  User-agent: IBM-WebCrawler
  Disallow:   # disallow nothing
  ```
- 此 /robots.txt 指定所有 robot 都应当避开 temp、development 和 testing html 文档树中的地址，但名为 IBM-WebCrawler 的用户代理程序不必避开 development 和 testing 树（在此允许），并且 xyz 和 wxyz robot 应当完全避开。

  ```
  # a more realistic example

  User-agent: *
  Disallow:   /htmldocs/temp
  Disallow:   /htmldocs/development
  Disallow:   /htmldocs/testing

  User-agent: xyz
  User-agent: wxyz
  ```

```
Disallow:   /

User-agent: IBM-WebCrawler
Disallow:   /htmldocs/temp
```

第 6 章 介绍工作流

您可以使用 EIP 工作流控制业务中工作的流程和性能。当用户使用联合搜索的结果处理内容时，他们经常要决定执行哪些操作。您可以使用 EIP 工作流来事先确定希望用户如何执行工作。

您可以通过设置控制工作流组件协同工作方式的概要文件和规则使工作流自动化。同时，通过控制特权集和访问控制表中的用户访问权限，还可设置对系统访问的约束。

理解工作流

大多数商务操作可以表现为一组相关过程。工作从一个员工流向另一个员工，从一个部门流向另一个部门。某些简单过程可能只需要几步，而比较复杂的过程涉及到不同部门的许多员工。

工作流使您可以推动工作流的处理，并且对处理过程中的工作作决策。例如，XYZ 保险公司接收到大量索赔表邮件。在验证过程期间，保险索赔调停人需要收集相关文档，例如照片、估价和专家报告。雇员为此每天要花几个小时来打开、排序、归档和监控信息，还要收集相关文档以备最后批准。

在这些信息被接收和检查时，会从一个雇员传送到另一个雇员。索赔处理完成时，它可能已由多个部门的雇员处理过了。

如何使用工作流

如 XYZ 保险公司示例所示，处理文档的大多数企业执行以下某些或全部任务:

- 存档以备检索。
- 通过不同资源收集文档、表单、报告和信息，然后将其传送到要对它们进行处理的地方。
- 使进入的邮件与当前正在处理的文档匹配。

*工作流*表示工作的流程。它描述了可以对一个或多个文档或内容组执行的操作，以及该文档组在工作流中所采用的路径。工作流反映了工作执行的方式以及明确定义的作用域和边界。它定义了活动和任务的顺序，以及这些活动和任务之间的连接与关系。工作流确定了用于对工作流程作决策的标准。关于工作流创建过程的信息，请参阅 *Workstation Application Programming Guide*。关于使用带有工作流的客户机的信息，请参阅 *Installing, Configuring, and Managing the eClient*。

使用户标识和组同步

本节说明如何在 EIP、Content Manager 和 MQ Series 工作流之间使用户标识和组同步。

当 MQSeries Workflow 是系统的一部分时管理 Content Manager 或 Enterprise Information Portal 中的用户的时候，您还管理 MQSeries Workflow 中的用户。这样，如果您在 Content Manager 或 Enterprise Information Portal 中创建、修改或删除用户标识或组，也必须在 MQSeries Workflow 中这样做。

因为 Content Manager 和 Enterprise Information Portal 共享用户标识和组，如果在启用了工作流服务的系统管理客户机上创建用户标识和组，但 MQSeries Workflow 服务器不在运行，那么您会收到错误消息，说明还未在 MQSeries Workflow 服务器上创建该用户或组。用户标识仍然作为 Content Manager 或 Enterprise Information Portal 上的用户存在，但不存在于 MQSeries Workflow 服务器上。

要将内容管理用户标识和组与 MQSeries Workflow 上的用户标识和组同步，您必须运行 Workflow 用户同步实用程序。如果已手工安装了 MQSeries Workflow 服务器，那么确保您的 MQSeries Workflow 服务器已启动。打开服务并检查 MQSeries Workflow *X.X* - FMC（其中 *X.X* 是已安装的工作流版本）的状态。如果 MQSeries Workflow 不在运行，则启动该服务，或者如果已静默安装，则转至 WFInstall 目录并运行批处理文件 `CMBWFStart.bat`。完成以下步骤以运行同步实用程序:

1. 更改到安装了 Enterprise Information Portal 的目录。缺省目录是 `C:\CMBROOT`。
2. 输入 `EIPUser2WF.bat`。

   重要信息: 如果有远程工作流服务器，则必须在输入 `EIPUser2WF.bat` 前启动 RMI 服务器批处理或外壳程序文件（`cmbsvregist81.bat` 或 `cmbsvregist81.sh`）。而且，`EIPUser2WF.bat` 使用来自两个 INI 文件的信息来定位本地或远程服务器: `cmbsvcs.ini` 和 `cmbsvclient.ini`。确保如果服务器是本地的，则 `cmbsvcs.ini` 指明“本地”（LOCAL）；如果服务器是远程的，则指明“远程”（REMOTE）。`cmbsvclient.ini` 文件包含服务器位置。
3. 输入 EIP 数据库名称、用户标识、密码和模式的必需信息。

输入了必需的信息后，同步实用程序将所有用户从 Content Manager 或 Enterprise Information Portal 服务器复制到 MQSeries Workflow 服务器。实用程序完成后，应当不会接收到以下错误: 用户或组对 MQSeries Workflow 服务器未知。

如果从 Content Manager 或 Enterprise Information Portal 删除用户标识或组，则还必须从 MQSeries Workflow 服务器删除用户标识或组。如果在 MQSeries Workflow 服务器上不存在某个用户标识或组，则不能从 Content Manager 或 Enterprise Information Portal 删除标识或组。例如，您在禁用 EIP 工作流服务器选项的 Content Manager 或 Enterprise Information Portal 中创建了某用户标识或组。然后在尝试删除此用户标识或组时启用工作流服务器选项。因为该用户标识或组在 MQSeries Workflow 中不存在，您将接收到错误消息，声明用户标识在 MQSeries Workflow 服务器中不存在。要更正错误，必须运行 EIPUser2WF 实用程序以使用户标识和组同步，然后删除该用户标识和组。

重新安装启用了工作流的 EIP 服务器

如果在 EIP 服务器中启用了 EIP 工作流，并且想要重新安装一个新的 EIP 服务器，那么必须从 MQSeries Workflow 服务器除去所有 EIP 工作流数据。

在除去 EIP 服务器前，先完成以下步骤以从 MQSeries Workflow 服务器除去所有 EIP 工作流数据。**重要信息**: 按以下面描述的顺序执行这些步骤。

1. 使用 eClient 或工作流 API 终止所有工作流实例。终止工作流实例会删除所有工作项。
2. 通过使用 EIP 系统管理客户机或系统管理 API 删除工作列表和工作流模板。
3. 当启用工作流时，从 EIP 删除所有用户和用户组，这样所有用户标识和用户组也会从 MQSeries Workflow 服务器删除。

4. 为 EIP 重新配置 MQSeries Workflow 服务器:
   a. 插入 EIP 安装 CD-ROM。使用命令提示符更改到 WfInstall 目录。
   b. 在命令提示符下，输入

      ```
      fmcibie -iCMBWFAdmin.fdl  -uadmin -ppassword -o  -f
      ```

   重要信息: 如果在从 EIP 和 MQSeries Workflow 删除用户标识和组之前除去 EIP 数据库，然后尝试创建在删除的 EIP 数据库中存在的相同用户标识和组，则将得到错误消息，说明无法将用户和组添加到 EIP。错误消息是: DGL2616A: 未能添加用户: XXX -DGL2485A: 此工作流用户已存在。完成以下步骤以解决此问题:

1. 在 EIP 系统管理客户机中禁用 EIP 工作流服务。
2. 注销 EIP 系统管理客户机，然后重新登录。在禁用工作流服务后登录时，可以创建与在 EIP 中和仍然存在于 MQSeries Workflow 服务器中的用户标识和组相同的用户标识和组。
3. 在创建已存在于 MQSeries Workflow 服务器中的用户标识和组后，启用 EIP 工作流服务。

更新 MQSeries Workflow 和 EIP 数据库之间的用户标识和组

对于存在于 MQSeries Workflow 服务器中但不存在于 EIP 数据库中的每个用户标识或组，可以完成以下步骤来使用户标识和组同步:

1. 创建文件并输入以下文本:
   - CODEPAGE 1252
   - FM_RELEASE V3R3 2
   - DELETE PERSON 'User1'
   - DELETE PERSON 'User2'
   - DELETE PERSON 'User3'

   其中 User1、User2 和 User3 是想要删除的用户。您可根据需要列出任意数量的用户。
2. 保存文件并在命令提示符下运行以下命令:

   ```
   fmcibie -u admin -ppassword -i DeletePersons.fdl  -f -o
   ```

   其中 DeletePersons.fdl 是在上一步中创建的文件的名称。

对于存在于 MQSeries Workflow 服务器但不再存在于 EIP 数据库的每个组，可以完成以下步骤:

1. 创建文本文件并输入以下信息:
   - CODEPAGE 1252
   - FM_RELEASE V3R3 2
   - DELETE ROLE 'Group1'
   - DELETE ROLE 'Group2'
   - DELETE ROLE 'Group3'

   其中 Group1、Group2 和 Group3 是想要删除的组。您可根据需要列出任意数量的组。
2. 保存文件并在命令提示符下输入以下命令:

```
fmcibie -u admin -p password -i DeleteGroups.fdl  -f -o
```

其中 DeleteGroups.fdl 是在上一步中创建的文件的名称。

当从 EIP 删除用户标识并且启用了工作流支持时，可能得到以下错误:

```
未能删除用户 [RC=12]
```

如果删除用户操作失败，则系统生成 temp.log，它提供关于失败的详细信息。temp.log 文件被写入 *x:*\CMBROOT。**提示:** 如果尝试删除用于登录到 MQSeries Workflow 客户机的用户标识，则删除用户操作将失败。

规划工作流

在开始定义工作流前，必须分析企业执行的工作、在何处执行、如何执行以及由谁执行。规划步骤由管理员或商务分析者制定。

什么是最终产品？最终产品可能是由您的企业、企业的某个部门或不同部门的某些员工完成的所有工作的结果。例如，XYZ 保险公司索赔补偿过程的最终产品是发送给核准或拒绝索赔的决策控制者的信件。

分析必须处理的信息以得到最终产品、确定必须执行的操作和在哪里执行这些操作，以及决定信息如何流过工作流。

要处理的信息

考虑必须由企业内用户处理的信息。哪些输入类型支持最终产品？指定为必须处理的文档是什么？

*工作项*不是一个实际文档。*工作项*包含对文档的引用，也包含关于文档的其它信息，例如文档状态、创建日期等等。*工作项*可以是来自内容服务器的任何内容（文档或对象）。例如，XYZ 保险公司首先接收到索赔表，然后接收到后续文档，如照片、估价和专家报告。

如何处理信息

谁是处理过程中每个步骤的最佳人选？例如，行政助理可能验证索赔表是否完成，然后将其申请存档，直到从决策控制者处收到特定的文档。收到文档后，索赔调停人可能会负责把收到的文档和索赔表相匹配以及审核该文档。

可以将*工作表*看作是为一个或多个员工的使用而创建的工作队列。索赔表应当分组到*工作表*中，以便许多索赔调停人可以访问。工作列表是工作项的过滤视图。雇员仅看到允许他们看的工作列表中的项。

工作列表可以定义为以处理索赔过程的每个部分（如收集照片、估价和报告）的方式过滤工作项。工作列表还可以由来自不同工作流的工作组成。例如，一个索赔调停人的工作列表可以包含一次索赔的估价、第二次索赔的照片以及第三次索赔的专家报告。调停人在工作列表中对每项执行的操作可以有所不同。调停者可能复查估价并审核第一次索赔。在对照片执行操作前，她可能需要等待关于第二次索赔的更多信息。对于第三次索赔，她可以将专家报告发送给另一个雇员进行运作。

要执行的操作

考虑要在工作流期间对工作项的内容采取哪些操作。例如，索赔调停人可接受索赔表或因索赔表不完整而拒绝接受。*操作列表*定义了用户可对工作执行的操作。

例如，根据索赔是否满足初始要求，调停人可选择一个选项在工作流中继续完成索赔表，或选择另一个选项拒绝接受索赔表。

信息如何在过程中流动

考虑信息和活动的流动方式。例如，何时复查初始索赔表？要在过程中移动到下一步需要哪些支持文档？确定是否接受或拒绝索赔的标准是什么？此信息流是工作流的基础。

工作流由处理过程中引导工作的路径构成。输入数据源于哪里？工作流必须在某个点开始。对于 XYZ 保险公司，决策控制者提交的索赔表是开始工作流的文档。

接收到所有文档时，工作项可以继续沿着该路径到达最终操作，例如，索赔审核。

所有工作如何结合

分析要处理的信息、确定要执行的操作以及决定信息如何流动后，您已可以创建一个工作流图表（即工作流的图形表示）了。使用 EIP 工作流构建器功能部件来创建该图表。

工作流图表显示工作如何通过过程中的各种活动，同时注解活动涉及到的任务。它描述了工作流的流程、主元素和关键点。

工作流图表中的每个符号都代表工作完成的位置。必须复查保险索赔，必须收集支持文档，必须根据特定条件审核或拒绝索赔。更多有关工作流构建器中使用的过程符号的信息，请参阅第 89 页的『创建工作流』。

使用 Enterprise Information Portal 工作流组件

本部分描述了工作流组件。您可以通过管理客户机访问所有组件。**技巧**: EIP V8 工作流包括很多更改，其中包括对 V7.1 容器的更改，从而适应新的 Content Manager V8 体系结构。

使用工作流构建器

使用工作流构建器以图表方式定义并构建工作组、部门或企业的工作流。**限制**: EIP 迁移过程将用户从 V7.1 数据库进行迁移。EIP V8.2 不提供对工作流数据的任何自动迁移。您必须使用 EIP V8.2 工作流构建器重新绘制 V7.1 工作流图表并重新部署 EIP V7.1 工作流过程。

使用工作流构建器创建工作流的模型之前，必须定义特权集、访问控制表、用户、用户组、操作、操作列表和工作列表。当在管理客户机中定义工作流时，可以为整个工作流设置缺省的操作列表。您还可以将不同的操作列表分配给工作流中的每个节点。关于这些任务的更多信息，请参阅第 89 页的『定义操作列表』、第 88 页的『定义工作列表』和联机帮助。

尽管可以使用工作流构建器建立工作过程，但不能用工作流构建器来运行工作过程。您的用户使用客户机进行查看并对工作列表和工作项执行操作。关于如何对客户机进

行编程使其与 Enterprise Information Portal 工作流一起工作的更多信息，请参阅 *Workstation Application Programming Guide* 和联机 API 参考。表 7 描述了对每个工作流都通用的三个工作流图标。还有其它工作流图标，例如子工作流、事件、集合和用户出口。EIP 联机帮助详细描述了工具栏图标。

表 7. 基本工作流图标

图标	描述
	开始节点开始工作流过程。工作流过程图表必须具有且只有一个开始节点。
	停止节点中止工作流过程。每个新的工作流过程图表包含一个停止节点。创建过程时，将为您生成一个停止节点。您可以将停止节点移动到绘制表面上的任意位置。工作流过程图表必须具有且只有一个停止节点。
	工作节点与工作流流程中的特定点的工作列表和操作列表相关联。工作节点代表从中执行工作的工作流流程中的某个点。每个节点（包括开始和停止节点）都需要一组操作和指定的人员来执行这些操作集合。

使用工作流服务

Enterprise Information Portal 提供维护工作流信息的工作流服务。使用工作流构建器创建的工作流和操作列表定义在 Enterprise Information Portal 管理数据库和 IBM MQSeries® Workflow 数据库中维护。

系统管理员创建工作列表时，与工作列表关联的信息永久存储在管理数据库中。系统管理员可以使用 EIP 更新、删除和添加工作列表。系统管理员检出工作流时，工作流锁定在 Enterprise Information Portal 数据库中，并在数据库中标记为已对用户检出，以防止在该用户完成操作前有任何其他用户更新工作流。

定义工作列表

工作列表可以视为可用工作过滤器。工作列表是指定给特定用户或用户组的项的已过滤列表。当用户登录到 Enterprise Information Portal 上时，他们可以查看分配给他们的已过滤的工作项列表。使用 Enterprise Information Portal 管理客户机定义工作列表。

工作列表定义包括了管理其工作项的显示、状态和安全性。在创建工作列表的同时为每个工作列表指定规则。要管理对工作列表的访问权限，则创建工作列表的访问控制表。关于如何定义工作列表的完整描述，请参阅联机帮助。工作列表定义包括以下部分:

访问控制表

访问控制表由一个或多个独立的用户标识、用户组以及与每个用户标识或用户组相关联的特权集构成。特权集用来定义用户对工作进行访问或执行某些任务的权限。使用访问控制表来限制用户对工作列表中项的访问。

对工作列表进行过滤和排序

用户可以查看已过滤和已排序的工作列表的条件。

工作列表中的最多项目数

希望工作列表包含的项的最大数目。

定义操作列表

操作列表是用户可以对工作流中的工作执行的所有操作的综合列表。

管理客户机联机帮助提供了解释如何定义操作和操作列表的逐步指导。

创建工作流

定义了操作、操作列表和工作列表之后，您就可以使用工作流构建器创建工作流的模型。管理客户机联机帮助提供了解释如何定义操作和操作列表的逐步指导。工作流构建器提供创建工作流的可视提示。

启用工作流构建器

在此步骤中，您将启动管理数据库上的工作流。**限制**：您对工作流选择的数据库必须处于与安装 MQ Series 的服务器相同的服务器上，且必须启动 MQSeries 服务。

要启用 EIP 工作流并创建工作流定义，请执行以下操作:

1. 登录到管理客户机。
2. 如果您有多个管理数据库，则请单击希望启用工作流的数据库的图标。
3. 单击“工具”->“服务”。单击“启用工作流”。
4. 从客户机注销并再次登录。如果您有多个数据库，则请选择启用 EIP 工作流的数据库的图标。将显示工作流文件夹图标。
5. 在 Enterprise Information Portal 管理主窗口的左侧窗格中，双击工作流文件夹**工作流**。
6. 右键单击“工作流定义”图标并选择**新建**以创建一个新的工作流定义。

需求：定义工作流之前，您必须创建至少一个访问控制表、一个操作和一个操作列表。

启动 MQSeries Workflow 服务器

通过在命令提示符下输入 `cmbwfstart` 来启动 MQSeries Workflow 服务器。此时会打开两个用于 MQSeries Workflow 服务器的窗口。将这些命令窗口保持为打开状态以继续运行服务器。

如果在对 Enterprise Information Portal 初始化安装之后安装工作流，则必须为工作流功能部件配置 Enterprise Information Portal 系统。若将工作流功能部件安装在不同的工作站上而不是安装在安装了管理客户机的工作站上，则还需要更改配置。

1. 从“管理”窗口，单击文件成员**工具**。
2. 从菜单单击**服务**。
3. 选择**工作流**复选框。
4. 配置完成后，注销 Enterprise Information Portal 管理客户机并再次登录来初始化工作流功能部件。登录到 Enterprise Information Portal 管理客户机之后，**工作流定义**图标出现在左侧窗格中。

技巧：除非管理员具有管理工作流功能部件的权限，否则他们看不到**工作流定义**图标。如果您要限制对工作流功能部件的访问，请参阅与每个内容服务器相应的系统管理书籍。请参阅联机帮助获取关于授权管理员管理工作流功能部件的更多信息。

可以使用 EIP 连接器工具箱和样本从定制应用程序创建客户机，也可以使用 EIP 样本客户机进行创建。

第 7 章 IBM Web Crawler 样本文件

本节提供了两个代码样本。config-sample2.xml 样本文件提供了 <gcs-config> 配置参数的示例。日志分析样本提供了包含来自已完成搜寻的信息的报告示例。

config-sample2.xml 样本

本节中的代码样本是 gcs-config 文件的示例。

```
<!DOCTYPE gcs-config SYSTEM "config.dtd">
<gcs-config>
   <!-- Global settings: -->
   <globals max-urls="1000000"
            num-crawlers="30"
            num-summarizers="8"
            summaries-dir="summaries"
            log-file="log/LOG.txt"
            temp-dir="temp"
            log-priority="warn"
            text-monitor="60"
            graph-monitor="2"
            connect-timeout="120"
            read-timeout="100">

      <!-- sample logger settings -->
      <logger-config category="gcs.summaries.list.resource"
priority="info" log-file="log/resources.txt"/>
      <logger-config category="gcs.summaries.list.host"     priority="info"
log-file="log/hosts.txt"/>
      <logger-config category="gcs.url.skipped"             priority="info"
log-file="log/skipped_urls.txt"/>
      <logger-config category="gcs.url.fromto"              priority="info"
 log-layout="%m\n"log-file="log/fromto.txt"/>
      <logger-config category="gcs.http"                    priority="info"
log-file="log/http.txt"/>
      <logger-config category="gcs.http.connect"            priority="info"
log-file="log/connecterrs.txt"/>

      <!--use this to specify a database
      <urlpool-config urlcontainer-class="DB2URLContainer"
urlcollection- class="DB2URLCollection">
         <urlpool-param name="dbname"    value="gcs"/>
         <urlpool-param name="user"      value="xxxxxx"/>
         <urlpool-param name="password"  value="xxxxxx"/>
         <urlpool-param name="cachesize" value="1000"/>
         <urlpool-param name="driver"
value="COM.ibm.db2.jdbc.app.DB2Driver"/>
      </urlpool-config> -->

      <!--use this to specify a SOCKS proxy
      <system-properties>
         <property name="socksProxySet" value="true"/>
         <property name="socksProxyHost" value="socks2.server.ibm.com"/>
         <property name="socksProxyPort" value="1080"/>
      </system-properties> -->

   </globals>

   <group-list>
      <group name="ibm">
```

```
          <crawler-config recursion-depth="-1">
             <seed-list>
             <!-- URLs to start crawling at: -->
                <seed url="http://gcs.stl.ibm.com/gcs/testurl.html"/>
                <seed url="http://gcs.stl.ibm.com/gcs/stl.html"/>
                <seed url="http://gcs.stl.ibm.com/gcs/ibm.html"/>
             </seed-list>

             <content-type-pattern-list>
             <!-- URL file extensions that don't match these patterns
                  won't be crawled: -->
                <url-name-pattern name="htm*"/>
                <url-name-pattern name="pdf"/>
                <url-name-pattern name="gif"/>
                <url-name-pattern name="zip"/>
                <url-name-pattern name="txt"/>
             </content-type-pattern-list>

             <include-pattern-list>
             <!-- URLs that don't match these patterns won't be crawled: -->
                <url-obj-pattern host="*.ibm.com"/>
       <!-- sbo - url-obj-pattern   query="*OpenDocument*"/ -->
       <!-- sbo - url-obj-pattern   query="*OpenView*"/ -->
                <!-- url-obj-pattern   query="*OpenDocument =>
OpenDocument&amp;ExpandAll*"/ -->
       <!-- url-obj-pattern   query="*OpenView =>
OpenView&amp;ExpandAll&amp;Count=999999*"/ -->
             </include-pattern-list>

             <exclude-pattern-list>
             <!-- URLs that match these patterns won't be crawled: -->
                <!-- skip these common patterns in our intranet -->
                <url-obj-pattern file="*news*"/>
                <url-obj-pattern file="*search*"/>
                <url-obj-pattern file="*/afs/*"/>
                <url-obj-pattern file="*/.../*"/>
                <url-obj-pattern file="*bluepages*"/>
                <!-- skip personal home pages -->
                <url-obj-pattern file="*/~*"/>
                <!-- skip SOCKS: no URL should specify this directly -->
                <url-regex-pattern match=".*:1080/.*"/>
                <!-- skip gateways: recommended for mere mortals -->
                <url-regex-pattern match=".*[\?\=\+\;\%&quot;&amp;].*"/>
                <!-- else crawl gateways as configured here... -->
                  <!-- skip Domino -->
                  <url-obj-pattern   file="*.nsf*"/>
                  <!-- else crawl Domino: allow only OpenDocument -->
                  <url-obj-pattern   query="*OpenServer*"/>
                  <url-obj-pattern   query="*OpenDatabase*"/>
                  <url-obj-pattern   query="*OpenElement*"/>
                  <url-obj-pattern   query="*OpenView*"/>
                  <url-obj-pattern   query="*OpenAbout*"/>
                  <url-obj-pattern   query="*OpenHelp*"/>
                  <url-obj-pattern   query="*OpenIcon*"/>
                  <url-obj-pattern   query="*OpenForm*"/>
                  <url-obj-pattern   query="*OpenNavigator*"/>
                  <url-obj-pattern   query="*OpenAgent*"/>
                  <url-obj-pattern   query="*CreateDocument*"/>
                  <url-obj-pattern   query="*DeleteDocument*"/>
                  <url-obj-pattern   query="*EditDocument*"/>
                  <url-obj-pattern   query="*SaveDocument*"/>
                  <url-obj-pattern   query="*SearchSite*"/>
                  <url-obj-pattern   query="*SearchView*"/>
                  <url-obj-pattern   query="*&amp;login*"/>
                  <url-obj-pattern   query="*Command*"/>
                  <!-- crawl Domino: avoid OpenDocument permutations -->
                  <url-obj-pattern   query="*ExpandSection*"/>
```

```
                <url-obj-pattern   query="*Navigate*"/>
                <url-obj-pattern   query="*Start*"/>
             <!-- -->

          </exclude-pattern-list>
        </crawler-config>

        <summarizer-config>
        <!-- Copy2Rdf is the default summarizer.  For these types, use: -->

          <resource-handler content-type="*htm*"
                            summarizable="EipHtmlSummarizable"
                            summary-maker="EipHtmlRawSummaryMaker" />
          <resource-handler content-type="*pdf"
                     summarizable="InsoSummarizable"
                     summary-maker="InsoSummaryMaker" />
        </summarizer-config>
      </group>
   </group-list>
</gcs-config>
```

IBM Web Crawler 日志分析文件示例

```
D:\gcs\run\log>perl loganalysis.pl log.txt

Log.txt 中 7710 行的共用时间为 1.84 分钟。

GCS was configured for 20 crawlers
        999  total crawls attempted
        137  - total crawl failures:
                  21    GCSHttpConnection.ABANDONING
                  12    GCSHttpConnection.CONNECT_ERROR
                  16    GCSHttpConnection.UNKNOWN_HOST
                   4    HTTP 403
                  29    HTTP 404
                   2    HTTP 500
                   8    HTTP 599
                   1    Read timed out
                  39    Robots not allowed
                   4    over max redirects
                   1    unknown protocol
   --------
        862  = successfully crawled
          0  - unchanged since earlier crawl
   --------
        862  = new or changed
        468  crawled per minute


GCS was configured for 5 summarizers
        855  total summaries attempted
          0  - total summary failures:
   --------
        855  = successfully summarized
        144  gcs.summaries.list.host
        855  gcs.summaries.list.resource
        465  summarized per minute


GCS successfully crawled 134 servers to obtain 862 URL:
  afqa0854.mop.ibm.com: 15
  als1f1.yamato.ibm.com: 1
  apache.btv.ibm.com: 1
  apc.endicott.ibm.com: 2
  as400service.ibm.com: 1
  atlas.bocaraton.ibm.com: 1
```

```
autoproxy.ibm.com: 1
cer.si.ibm.com: 1
commerce.www.ibm.com: 1
crmweb.boulder.ibm.com: 3
d02ntcl01.ibm.com: 1
dacs.endicott.ibm.com: 1
duke.toraix.can.ibm.com: 1
ebcweb.austin.ibm.com: 1
ecspubs.ibmus2.ibm.com: 5
edaw3.fishkill.ibm.com: 1
endwww.endicott.ibm.com: 1
gcs.stl.ibm.com: 1
gustwick.austin.ibm.com: 1
ibmfnsys.somers.hqregion.ibm.com: 1
ibmpnyi1.somers.hqregion.ibm.com: 2
ifw-www.mul.ie.ibm.com: 1
iplswww.nas.ibm.com: 2
itirc.ibm.com: 1
logosite.services.ibm.com: 1
lt.lahulpe.ibm.com: 17
messaging.ibm.com: 1
mrsmrn04.leeds.uk.ibm.com: 1
online.lahulpe.ibm.com: 1
page.sg.ibm.com: 1
procure.sby1.ibm.com: 1
reso.somers.hqregion.ibm.com: 1
riscta1.leipzig.de.ibm.com: 1
rrhhar.argentina.ibm.com: 1
seashore.stl.ibm.com: 1
secureway.raleigh.ibm.com: 15
service.software.ibm.com: 1
software.ibmus2.ibm.com: 1
techcenter.austin.ibm.com: 1
tr2.fishkill.ibm.com:8080: 1
ucd.torolab.ibm.com: 1
usmweb.boulder.ibm.com: 1
w3-1.ibm.com: 32
w3-2.ibm.com: 3
w3-3.ibm.com: 108
w3-5.ibm.com: 4
w3.a-nz.au.ibm.com: 1
w3.academy.ibm.com: 1
w3.almaden.ibm.com: 2
w3.alphaworks.ibm.com: 1
w3.ap.ibm.com: 1
w3.asca.ibm.com: 7
w3.austin.ibm.com: 3
w3.boulder.ibm.com: 1
w3.br.ibm.com: 1
w3.btv.ibm.com: 1
w3.can.ibm.com: 40
w3.chq.ibm.com: 4
w3.coc.ibm.com: 1
w3.corporatetechnology.ibm.com: 1
w3.cupertino.ibm.com: 1
w3.dds.dfw.ibm.com: 17
w3.demopkg.ibm.com: 4
w3.design.ibm.com: 1
w3.developer.ibm.com: 3
w3.education.ibm.com: 1
w3.emea.ibm.com: 14
w3.enterlib.ibm.com: 7
w3.finsys.ibm.com: 1
w3.gcg.ibm.com: 1
w3.globalfinancing.de.ibm.com: 1
w3.hakozaki.ibm.com: 1
w3.houston.ibm.com: 1
```

```
  w3.hursley.ibm.com: 5
  w3.iabc.ibm.com: 1
  w3.ibm.com: 180
  w3.ibmfax.ibm.com: 1
  w3.ibmla.ibm.com: 14
  w3.isicc.de.ibm.com: 1
  w3.itso.ibm.com: 1
  w3.japan.ibm.com: 1
  w3.knowledge.raleigh.ibm.com: 1
  w3.linux.ibm.com: 1
  w3.marketiq.ibm.com: 1
  w3.micro.ibm.com: 2
  w3.mtlisc.can.ibm.com: 1
  w3.munich.ibm.com: 1
  w3.ode.raleigh.ibm.com: 1
  w3.paylink.au.ibm.com: 1
  w3.pisc.uk.ibm.com: 1
  w3.pl.ibm.com: 1
  w3.printers.ibm.com: 1
  w3.pssc.mop.ibm.com: 1
  w3.pssed.au.ibm.com: 1
  w3.raleigh.ibm.com: 3
  w3.rchland.ibm.com: 1
  w3.research.ibm.com: 3
  w3.reserve.ibm.com: 1
  w3.rs6000.ibm.com: 1
  w3.security.ibm.com: 1
  w3.software.ibm.com: 6
  w3.ssd.ibm.com: 1
  w3.stl.ibm.com: 1
  w3.techline.ibm.com: 1
  w3.techsupp.yamato.ibm.com: 1
  w3.torolab.ibm.com: 2
  w3.usergroup.ibm.com: 1
  w3.vendor.pok.ibm.com: 1
  w3.viewblue.ibm.com: 1
  w3.watson.ibm.com: 2
  w3.wdg.uk.ibm.com: 1
  w3.ytal.yasu.ibm.com: 1
  w3.zurich.ibm.com: 1
  w3chq.disbursements.ibm.com: 1
  w3is.lagaude.ibm.com: 1
  w3md.btv.ibm.com: 1
  w3ssd.mainz.de.ibm.com: 1
  w3vm.demopkg.ibm.com: 1
  widweb.raleigh.ibm.com: 1
  wtscpok.itso.ibm.com: 1
  wwas.raleigh.ibm.com: 1
  www-1.ibm.com: 63
  www-3.ibm.com: 4
  www-4.ibm.com: 86
  www.almaden.ibm.com: 1
  www.as400.ibm.com: 1
  www.chips.ibm.com: 1
  www.ibm.com: 52
  www.ieg.ibm.com: 1
  www.patents.ibm.com: 1
  www.pc.ibm.com: 2
  www.rs6000.ibm.com: 23
  www.software.ibm.com: 9
  www.storage.ibm.com: 1
  www.watson.ibm.com: 1

GCS timed out 1 times:
    w3-3.ibm.com: 1

GCS ignored 42 URL prohibited by robots.txt:
```

```
  reso.somers.hqregion.ibm.com: 1
  w3.education.ibm.com: 1
  w3.rchland.ibm.com: 34
  w3.zurich.ibm.com: 1
  www.ibm.com: 5

GCS skipped 3846 URL (requires gcs.url logging)
  59 specified an unsupported protocol:
    protocol not supported  gopher: 6
    protocol not supported  mailto: 53
  1206 had content-types (lower or UPPER case, > 10) that were not included
    .2: 12
    .faq: 13
    .1: 14
    .asp: 16
    .cgi: 21
    .shtml: 90
    .pl: 92
    .gif: 157
    .nsf: 160
    .jpg: 214
    .css: 240
  516 URL were on servers and/or paths that were not included
  2065 were excluded for these reasons:
    URL longer than 254: 1
    excluded by rule 1: 1210
    excluded by rule 2: 854
```

第 8 章 使用文本搜索和 QBIC®

此附录的第一部分解释了如何配置并使用文本搜索和按图像内容查询（QBIC），这两个功能仅当您在 EIP 安装期间选择 Content Manager V7.1 连接器时可用。此附录的第二部分提供了装入样本应用程序使用的样本文本和图像数据的信息。

使用文本搜索引擎搜索文档

文本搜索可以与 Content Manager V7.1 服务器集成，这样您就可以自动对存储在 Content Manager 中的文档建立索引、搜索以及检索了。用户可以通过搜索单词或词组查找文档。文本搜索服务器支持单字节和双字节字符集并且可以运行于 AIX 和 Windows 上。

文本搜索包括 XML、HTML 结构化文档支持和标记的 ASCII 文档，这使检索词搜索可以在文档的指定段内进行。您可以在嵌套的段内搜索数据。搜索可以根据完整的 XML 上下文进行，例如，在标题中搜索 IBM，也可以在特定部分中的某个标题中搜索 IBM。指定 DTD 路径时，假定对 DTD 的引用存储为文档的元数据，文本搜索可动态使用对应于每个文档的 DTD。

请参阅 *Planning and Installing Enterprise Information Portal* 获取关于使用文本搜索规划和安装 EIP 系统的信息。

启用文本搜索服务器

要使用文本搜索服务器，您必须在启动 IBM Content Manager for Multiplatforms 管理客户机之前启用对服务器的管理。要启用管理:

1. 启动 IBM Content Manager for Multiplatforms 库服务器。

   允许库服务器完成构建索引类。
2. 通过输入以下内容在它所安装的工作站上启动文本搜索服务器:

   ```
   imlss -start dlinst
   ```

   其中，*dlinst* 是在安装时或使用 *imlcfgsv* 命令实用程序时选择的文本搜索服务器实例的名称。

使用按图像内容查询（QBIC）搜索图像

本节介绍了按图像内容查询（QBIC）并解释了如何配置和使用 QBIC。QBIC 功能仅在安装了 Content Manager V7.1 连接器的情况下可用。QBIC 与 Windows 和 AIX 操作系统兼容。

介绍图像搜索

图像搜索服务器使用 IBM 的 QBIC（按图像内容查询）技术帮助您根据某些可视属性（如颜色和纹理）搜索对象。图像搜索服务器分析图像并在数据库中存储图像信息。然后，用户运行图像查询时可使用可视图像属性来匹配颜色、纹理以及它们的位置，而无须使用文字描述。您可以把文本和关键字搜索和基于内容的查询组合起来，提供更强大的图像和多媒体数据检索。

每个图像搜索服务器具有一个数据目录，它包含一个或多个存有图像搜索目录的图像搜索数据库。图像搜索目录存储关于图像集合可视特性的数据。实际的图像对象存储在 IBM Content Manager for Multiplatforms 系统中的对象服务器上。图像搜索服务器在 AIX 和 Windows 上运行。

请参阅 *Planning and Installing Enterprise Information Portal* 获取关于安装图像搜索的信息。

设置图像搜索

这些指导仅在安装图像搜索之后适用，如果选择 Content Manager V7.1 连接器，则会自动安装图像搜索。设置图像搜索由以下几部分组成:

1. 设置环境
2. 配置图像搜索服务器
3. 配置图像搜索客户机
4. 装入样本图像数据

如果在 AIX 上使用安装向导: 您不必运行配置和设置脚本或发出服务器配置命令。向导可以帮您完成这些任务。

如果在 Windows 上安装: 必须完成这些任务。

设置环境

本节介绍在服务器和客户机机器上完成环境设置任务。图像搜索服务器需要以下环境变量:

QBICTOP
在图像搜索配置期间解析文件名

QbicImagePath
解析服务器图像文件的文件名

QbicMaskPath
解析服务器掩码文件的文件名

QbicSketchPath
解析服务器略图文件的文件名

QbicTextPath
解析服务器文本文件的文件名

图像搜索客户机仅需要 QBICTOP 环境变量。

AIX 示例: 在 AIX 上，运行生成设置脚本的配置脚本，然后运行设置脚本以设置环境。

1. 运行以下配置脚本:

   ```
   /usr/lpp/cmb/bin/frnconfg.iss QBICTOP
   ```

   其中 `QBICTOP` 是控制文件（`*.ini`）的目录路径。将 `QBICTOP` 设置为 `/user1/cmb/qbic`，其中 `/user1` 是图像搜索管理用户标识的主目录。图像搜索用户标识必须对该目录有读 / 写访问权限。

   此脚本将生成设置脚本: `frnsetup.iss`。

2. 从图像搜索用户标识的主目录运行:

   ```
   . ./frnsetup.iss
   ```

   该脚本为图像搜索服务器和客户机填充环境变量。

Windows 示例: 执行以下操作设置环境变量:

1. 选择**开始 → 设置 → 控制面板**。
2. 双击**系统**。
3. 单击“环境”选项卡。
4. 通过在适当的字段输入变量和值并单击**设置**来设置在表 8 中所示的变量和值。

   要求: 图像搜索客户机仅需要 QBICTOP 变量。在客户机环境下，只需要设置 QBICTOP 变量。

表 8. 图像搜索环境变量

变量	值
QBICTOP	d:\cmbroot\iss
QbicImagePath	d:\cmbroot\iss
QbicMaskPath	d:\cmbroot\iss
QbicSketchPath	d:\cmbroot\iss
QbicTextPath	d:\cmbroot\iss
其中，d: 是用于安装图像搜索的驱动器。	

配置图像搜索服务器

在启动图像搜索服务器之前，必须先配置图像搜索服务器。配置服务器包括初始配置和验证连接。

执行以下步骤配置服务器:

1. 输入 qbicadm 启动命令解释器
2. 输入 **config server** 命令。例如,

   ```
   config server LIBSRVRN FRNADMIN PASSWORD 9999
   ```

   其中 LIBSRVRN 是库服务器名称、FRNADMIN 是 Content Manager 用户标识、PASSWORD 是 Content Manager 密码、9999 是图像搜索服务器的端口号。

请参阅第 100 页的『验证连接』获取更多信息。

配置图像搜索客户机

在启动图像搜索客户机（包括图像搜索系统管理程序）之前，必须配置它。执行 Content Manager 系统管理需要您指定一个别名。请验证连接以测试您的配置。

指定别名: 在使用图像搜索 Content Manager 系统管理（它充当图像搜索客户机）之前，必须至少指定一个服务器别名。

执行以下步骤指定别名:

1. 输入 qbicadm 启动命令解释器
2. 输入 **add alias** 命令。例如,

   ```
   add alias QBICSRV HOSTNAME 9999
   ```

其中 QBICSRV 为别名、HOSTNAME 是图像搜索服务器的主机名、9999 是图像搜索服务器的端口号。

验证连接:

注意事项:

1. 要连接到图像搜索服务器，库服务器必须在运行。
2. 图像搜索系统管理需要现有的 Content Manager 用户标识。只有图像搜索用户标识和库服务器用户标识相同时，才能成功地连接到库服务器。此用户标识的缺省值为 frnadmin。如果更改此值，则请确保标识匹配。

执行以下步骤验证连接:

1. 配置图像搜索服务器并添加别名之后，请通过从服务器命令行输入 commsrv 启动服务器。
2. 要启动命令解释器，请输入: qbicadm。
3. 在命令解释器中，输入 connect 命令。

   ```
   connect QBICSRV FRNADMIN PASSWORD
   ```

   其中 QBICSRV 是别名、FRNADMIN 是 Content Manager 用户标识、PASSWORD 是 Content Manager 密码。

   成功连接之后，会显示消息库服务器是 LIBSRVRN。
4. 要从服务器断开连接，请输入: disconnect。
5. 要退出命令解释器，请输入: quit。

装入样本数据并对其建立索引

本节说明如何装入并索引样本文本和图像数据，这些文本和图像数据可以作为样本应用程序使用。本节仅在您安装了 Content Manager V7.1 连接器并选择了文本搜索选项时适用。

在 Enterprise Information Portal CD 上提供了几个样本装入器。本节描述使用示例装入器 LoadSampleTSQBICDL 如何装入图像和文本数据。用户可以分别装入文本和图像搜索数据，确保各功能部件正常工作。

装入样本数据之前

在运行装入程序之前，必须:

1. 登录到 EIP 管理客户机。单击 “开始→程序→**Enterprise Information Portal** for Multiplatforms 8.2→管理”。
2. 选择一个数据库并使用正确的用户标识和密码登录。如果选择缺省数据库 icmnlsdb，则输入 **icmadmin** 作为用户标识，并在密码字段中输入 password。如果使用另一个数据库，则输入适用的用户标识。
3. 使用 Content Manager 系统管理程序创建库服务器配置。请参阅系统管理程序联机帮助获得此任务的辅助信息。
4. 通过完成以下步骤，更改库服务器配置的**访问**特性:
   a. 右键单击新配置并单击**属性**以打开 “属性” 笔记本。
   b. 单击 “访问” 选项卡。

c. 单击**来自任何工作站的无限制会话**单选按钮。

创建文本搜索索引

在装入数据前，必须创建空文本搜索索引，用于索引文本样本。**技巧**: 只能在 Content Manager V6.1 或V7.1 服务器上创建文本搜索索引。

要创建文本搜索索引:

1. 使用以下命令在安装了文本搜索服务器的工作站上启动文本搜索服务器:

   ```
   imlss -start dlinst
   ```

   其中，`dlinst` 是在安装时或使用 `imlcfgsv` 命令实用程序时选择的文本搜索服务器实例的名称。
2. 启动并登录到 Content Manager 系统管理程序。
3. 从左上部窗格中的列表选择**文本搜索**。
4. 双击左窗格中的**搜索服务器**。
5. 双击 **TM**。TM 是文本搜索服务器的搜索服务器别名。
6. 双击左窗格中的 **Indexes** 文件夹。如果显示消息 `RC_EMPTY_LIST`，则从菜单栏单击**选定项 ⟶ 新建**以创建索引。
7. 在“新建索引”窗口，定义索引。单击**帮助**可以获取每个字段的详细描述。

   例如:

   对于 Windows:

   名称 `TMINDEX`

   类型 精确

   索引文件

   `x:\cmbroot\ts\index\tmindex` 其中 `x` 是安装驱动器；如果路径不存在，则创建它。

   对工作文件做索引

   `x:\cmbroot\ts\work\tmindex` 其中 `x` 是安装驱动器；如果路径不存在，则创建它。

   信息条目

   Content Manager 库服务器的名称。

   不要更改客户机和服务器的缺省 DLL 名称。

   对于 AIX:

   名称 `TMINDEX`

   类型 精确

   索引文件

   `/home/cltadmin/tsindex/index/tmindex`；如果路径不存在，则创建它。

   对工作文件做索引

   `/home/cltadmin/tsindex/work/tmindex`；如果路径不存在，则创建它。请确保用户对该目录有写权限。

信息条目

Content Manager 库服务器的名称。

8. 单击**确定**。
9. 双击 **TMINDEX** 打开 TMINDEX 管理笔记本。

创建图像搜索数据库、目录和功能部件

在为样本文本搜索数据创建文本搜索索引后，必须创建图像搜索数据库，并为样本图像数据编目。

要创建图像搜索数据库、目录和功能部件:

1. 输入以下命令，启动安装在工作站上的图像搜索服务器:

   ```
   commsrv
   ```

2. 启动并登录到 Content Manager 系统管理程序。
3. 从左上部的窗格中选择**图像搜索**。
4. 在左窗格中单击**图像搜索服务器**。
5. 单击 **QBICSRV**。

   其中 QBICSRV 是安装期间指定的图像搜索服务器名称。
6. 在左窗格中右键单击**数据库**并选择**新建数据库**。
7. 在“新建数据库”窗口的**名称**字段中输入 SAMPLEDB 并单击**确定**。
8. 在左窗格中，单击**数据库**以显示左窗格中的 **SAMPLEDB** 图标。
9. 单击 **SAMPLEDB**。
10. 在左窗格中，右键单击**目录**并单击**新建目录**。
11. 在“新建目录”中的**名称**字段中输入 SAMPLECAT 并单击**确定**。
12. 在左窗格中，单击**目录**以显示 **SAMPLECAT** 图标。
13. 单击 **SAMPLECAT**。
14. 在左窗格中，右键单击**功能部件**并单击**新建功能部件**。
15. 在“新建功能部件”窗口的**名称**字段中选择每个功能部件并单击**应用**。当选定了所有四个功能部件后显示以下消息:

    所有可能的功能部件都已经添加到目录中。
16. 单击**确定**。
17. 单击**取消**。

运行装入程序

您可以装入样本数据来测试文本和图像搜索。

样本图像数据在以下文件中:

在 Windows 上:

```
x:\cmbroot\samples\java\dl\samples.jar
```

在 AIX 上:

```
/usr/lpp/cmb/samples/java/dl/samples.jar
```

样本装入程序将数据装入 Content Manager 并为它建立索引。阅读源程序的前言，获取关于运行该程序的语法指导。样本装入程序是:

在 Windows 上:

```
x:\cmbroot\samples\java\dl\LoadSampleTSQBICDL.jar
```

在 AIX 上:

```
/usr/lpp/cmb/samples/java/dl/LoadSampleTSQBICDL.jar
```

要运行样本数据装入程序:

1. 通过输入以下命令解压 .jar 文件:

   ```
   jar -xvf samples.jar
   ```

   文件将解压到正确的目录。
2. 通过执行以下任务，设置工作站的环境变量以编译样本装入程序:

   在 Windows 上:

   a. 在文本编辑器中打开 x:\cmbroot\cmbenv71.bat 并更改前三行以设置工作站环境变量:

      ```
      set CMBROOT = e:\cmbroot
      set DB2HOME = e:\sqllib
      set JAVAHOME = d:\jdk117
      ```

   b. 保存 cmbenv71.bat 并通过输入以下命令设置环境变量:

      ```
      cmbenv71
      ```

   在 AIX 上:

   a. 转至 /usr/lpp/cmb/bin/ 并通过输入以下命令运行安装:

      ```
      . /cmbenv71.sh
      ```

   b. 确保所有用户对 /usr/lpp/cmb/samples/java/dl/ 下的子目录和样本文件有写权限。
3. 输入以下区别大小写的命令来编译装入程序:

   ```
   javac  LoadSampleTSQBICDL.java
   ```
4. **要求**: 在启动装入程序之前，以下服务器必须正在运行:
   - 库服务器
   - 对象服务器
   - 文本搜索服务器
   - 图像搜索服务器

   如果要运行 Content Manager 的本地语言版本，请在运行装入程序之前将 FRNDEFLANG 变量设置为 ENU。设置环境变量的 AIX 命令为: export FRNDEFLANG=ENU
5. 输入以下命令，使用装入程序装入样本数据:

   ```
   java LoadSampleTSQBICDL sampleQBIC.dat load.log frnadmin password LIBSRVRN
   ```

   其中用户标识为 frnadmin、密码为 password、库服务器为 LIBSRVRN。
6. 检查 load.log 以确保样本数据已成功装入。

在完成装入样本数据后，请使用 Content Manager 系统管理程序或文本搜索命令行工具索引样本文本数据。

建立样本文本数据索引

要建立样本数据索引:

1. 启动并登录到 Content Manager 系统管理程序。
2. 从左上部窗格中的列表选择**文本搜索**。
3. 双击**搜索服务器**。
4. 双击 **TM**。TM 是文本搜索服务器的搜索服务器别名。
5. 右键单击**新建文本索引**并单击**特性**。
6. 在“特性”笔记本的“显式”页面上单击**刷新**。
7. **索引计数**字段应该显示装入程序装入的文件的数量。
8. 单击**索引**以对这些文件进行索引。
9. 稍后请单击**刷新**，在**主文档索引**字段查看成功索引的文档数。

对数据建立索引之后，可以使用样本 Java 应用程序查询集合或通过使用 `imlsrch` 命令行工具来运行简单查询。

第 9 章 文档格式

信息发掘文档格式

本附录描述了信息发掘支持的文档格式。

字处理: 一般

ANSI 文本（7 和 8 位）	所有版本
ASCII 文本（7 和 8 位版本可用）	所有版本
HTML	V3.0 及以下（存在一些限制）
IBM FFT	所有版本
IBM 可修订格式文本	所有版本
Microsoft Rich Text Format（RTF）	所有版本
Unicode 文本	所有版本

字处理: DOS

DEC WPS Plus（DX）	V4.0 及以下
DEC WPS Plus（WPL）	V4.1 及以下
DisplayWrite® 2 & 3（TXT）	所有版本
DisplayWrite 4 & 5	R2.0 及以下版本
Enable	V3.0、4.0 和 4.5
First Choice	V3.0 及以下
Framework	V3.0
IBM Writing Assistant	V1.01
Lotus Manuscript	V2.0 及以下
MASS11	V8.0 及以下
Microsoft Word	V6.0 及以下
Microsoft Works	V2.0 及以下
MultiMate	V4.0 及以下
Navy DIF	所有版本
Nota Bene	V3.0
Office Writer	V4.0 到 6.0

PC-File Letter	V5.0 及以下
PC-File+ Letter	V3.0 及以下
PFS:Write	VA、B 和 C
Professional Write	V2.1 及以下
Q&A	V2.0
Samna Word	V4.0 及以下
SmartWare II	V1.02
Sprint	V1.0
Total Word	V1.2
Volkswriter 3 & 4	V1.0 及以下
Wang PC（IWP）	V2.6 及以下
WordMARC	版本 Composer Plus 及以下
WordPerfect	V6.1 及以下
WordStar	V7.0 及以下
WordStar 2000	V3.0 及以下
XyWrite	VIII Plus 及以下

字处理: 国际

JustSystems Ichitaro	V5.0、6.0、8.0、9.0 和 10.0

字处理: **Windows**

AMI/AMI Professional	V3.1 及以下
Corel WordPerfect for Windows	V9.0 及以下
JustWrite	V3.0 及以下
Legacy	V1.1 及以下
Lotus WordPro（Win32 / Intel 平台）	SmartSuite® 96、97 和 Millennium
Lotus WordPro（Unix 平台 — 仅文本）	SmartSuite 97 和 Millennium
Microsoft Windows Works	V4.0 及以下
Microsoft Windows Write	V3.0 及以下
Microsoft Word 97	Word 97
Microsoft Word 2000	Word 2000
Microsoft Word for Windows	V7.0 及以下
Microsoft WordPad	所有版本

Novell Perfect Works	V2.0
Novell WordPerfect for Windows	V7.0 及以下
Professional Write Plus	V1.0
Q&A Write for Windows	V3.0
WordStar for Windows	V1.0

字处理: **Macintosh**

Microsoft Word	V4.0 到 6.0
Microsoft Word 98	Word 98
WordPerfect	V1.02 到 3.0
Microsoft Works	V2.0 及以下
MacWrite II	V1.1

电子数据表格式

VP Planner 3D	V1.0
Enable	V3.0、4.0 和 4.5
First Choice	V3.0 及以下
Framework	V3.0
Lotus 1-2-3®（DOS 和 Windows）	V5.0 及以下
Lotus 1-2-3 for SmartSuite	SmartSuite 97 和 Millennium
Lotus 1-2-3 Charts（DOS 和 Windows）	V5.0 及以下
Lotus 1-2-3（OS/2®）	V2.0 及以下
Lotus 1-2-3 Charts（OS/2）	V2.0 及以下
Lotus Symphony	V1.0、1.1 和 2.0
Microsoft Excel 97	Excel 97
Microsoft Excel 2000	Excel 2000
Microsoft Excel Macintosh	V3.0 到 4.0、98
Microsoft Excel Windows	V2.2 到 7.0
Microsoft Excel Charts	V2.x 到 7.0
Microsoft Multiplan	V4.0
Microsoft Windows Works	V4.0 及以下
Microsoft Works（DOS）	V2.0 及以下
Microsoft Works（Mac）	V2.0 及以下
Mosaic Twin	V2.5

Novell Perfect Works	V2.0
QuattroPro for DOS	V5.0 及以下
QuattroPro for Windows	V9.0 及以下
PFS:Professional Plan	V1.0
SuperCalc 5	V4.0
SmartWare II	V1.02

数据库格式

SmartWare II	V1.02
Access	V2.0 及以下
dBase	V5.0 及以下
DataEase	V4.x
dBXL	V1.3
Enable	V3.0、4.0 和 4.5
First Choice	V3.0 及以下
FoxBase	V2.1
Framework	V3.0
Microsoft Windows Works	V4.0 及以下
Microsoft Works（DOS）	V2.0 及以下
Microsoft Works（Mac）	V2.0 及以下
Paradox（DOS）	V4.0 及以下
Paradox（Windows）	V1.0 及以下
Personal R:BASE	V1.0
R:BASE 5000	V3.1 及以下
R:BASE System V	V1.0
Q & A	V2.0 及以下
Reflex	V2.0

标准图形格式

PNG — 可移植网络图形因特网格式	V1.0
Binary Group 3 Fax	所有版本
BMP（包括 RLE、ICO、CUR 和 os/2 DIB）	Windows
CDR（如果 TIFF 图像已嵌入其中）	Coral Draw V2.0 到 9.0
CGM — 计算机图形元文件	ANSI、CALS、NIST, V3.0

CMX - Corel Clip Art 格式	V5 到 6
DCX（multi-page PCX）	Microsoft Fax
DRW - Micrografx Designer	V3.1
DRW - Micrografx Draw	V4.0 及以下
DXF（二进制和 ASCII）AutoCAD 绘图交换格式	V14 及以下
EMF	Windows Enhanced Metafile
EPS Encapsulated PostScript	如果 TIFF 图像已嵌入其中
FMV – FrameMaker 图形	向量和光栅格式，版本 5.0 及以下
FPX - Kodak Flash Pix	无特定格式
GDF - IBM 图形数据格式	V1.0
GEM — 图形环境管理器元文件	位图和向量
GIF — 图形交换格式	Compuserve
GP4 - Group 4 CALS 格式	类型 I 和类型 II
HPGL - Hewlett Packard 图形语言	V2.0
IMG - GEM Paint	无特定版本
JFIF（不属于 TIFF 格式的 JPEG 格式）	所有版本
JPEG — 联合图像专家组格式	所有版本
MET - OS/2 PM 元文件	V3.0
PBM — 可移植位图	无特定版本
Kodak Photo CD	V1.0
PCD - PCX 位图	PC Paintbrush
Perfect Works（Draw）	Novell V2.0
PGM - Portable Graymap	无特定版本
PIC - Lotus 1-2-3 图片文件格式	无特定版本
PICT1 和 PICT2（光栅）	Macintosh 标准
PIF - IBM 图片交换格式	V1.0
PNTG	MacPaint
PPM — 可移植像素图	无特定版本
Progressive JPEG	无特定版本
PSP - Paintshop Pro（仅 Win32）	V5.0、5.0.1

RND - AutoShade Rendering 文件格式
V2.0

SDW Ami Draw Snapshot（Lotus）
所有版本

SRS - Sun Raster 文件格式 无特定版本

Targa Truevision

TIFF V6 及以下

TIFF CCITT Group 3 & 4 传真系统

VISO（页面预览方式仅针对版本 4）Visio 4、5、2000
Visio 4、5、2000

WMF Windows 元文件

WordPerfect 图形 [WPG 和 WPG2]
V2.0 及以下

XBM - X-Windows 位图 x10 兼容

XPM - X-Windows 像素图 x10 兼容

XWD - X-Windows 转储 x10 兼容

高端图形格式

PSD - Adobe Photoshop 文件格式
V4.0

AI - Adobe Illustrator 文件格式
V7.0 及以下

CDR - Corel Draw V8.0 及以下

DSF - Micrografx Designer Windows 95, V6.0

DWG - AutoCAD 本机绘图格式
V12 到 14

IGES — 初始图形交换规范 V5.1

PDF — 可移植文档格式 Acrobat V2.1、3.0、4.0, 包括日文 PDF

PS - Postscript 级别 2

演示格式

Microsoft PowerPoint for Macintosh
V4.0, 98

Corel Presentations V8.0 和 9.0

Novell Presentations V3.0 和 7.0

Harvard Graphics for DOS V2.x 和 3.x

Harvard Graphics Windows 版本

Freelance 96 Freelance 96

Freelance for Windows 95	SmartSuite 97 和 Millennium
Freelance for Windows	V1.0 和 2.0
Freelance for OS/2	V2.0 及以下
Microsoft PowerPoint for Windows	V7.0 及以下
Microsoft PowerPoint 97	PowerPoint 97
Microsoft PowerPoint 2000	PowerPoint 2000

压缩和编码格式

ZIP PKWARE	V2.0g 及以下
GZIP	无特定版本
LZA Self Extracting Compress	无特定版本
LZH Compress	无特定版本
Microsoft Binder	V7.0, Binder 97
MIME（文本邮件）	无特定版本
UUEncode	无特定版本
UNIX® Compress	无特定版本
UNIX TAR	无特定版本

其它

vCard Electronic Business Card	V2.1
可执行文件（**EXE**、**DLL**）	无特定版本
Windows NT 的可执行文件	无特定版本
MSG（仅文本）	Microsoft Outlook 邮件格式
Microsoft Project（仅文本）	Project 98

第 10 章 权限管理

本章介绍了 EIP 权限管理功能并描述了权限管理概念。它解释了您可以用于保护产权的标记技术。

保护您的知识产权

以数字格式存储的多媒体资源即为您的知识产权。对这些对象的保护对于您的商业目标很关键，尤其是在这些对象处于万维网（相对容易复制的位置）上时。您可以使用 Content Manager 提供的标记技术，通过标记多媒体数字对象以进行保护或者通过对对象进行指纹识别以进行辨认来阻止知识产权的未授权使用。

您可以将标记应用于有价值的对象以实现:

- 标识源，以阻止未授权复制或重复使用。这被称为水印且通常可见。
- 标识内容接收方，以阻止未授权复制或重复使用。这被称为指纹且通常不可见。
- 提供获取附加信息的联系方式。
- 给出信息（如时间和日期），以在增值分发链中使用。

您可以在将数字对象交付给顾客之前标记那些对象。交付前可同时应用水印和指纹。但是，交付之前应用指纹意味着接收方是已知的且可能需要在交付过程中动态应用标记。交付之前从您自己控制的环境应用标记会增加安全性，因为篡改的风险较小。

您可以在管理和交付过程中的多个阶段应用标记。您的情形会对适用的内容有所影响。标记可以在过程中的以下时间点应用:

- 存储对象之前

  如果公共标记将用于该对象（例如，标识所有者的可视水印），则您可以在存储对象之前或存储对象时应用标记。您可以同时将原始、未标记对象和已标记对象存储在 Content Manager 系统中。您也可以仅存储已标记对象并将未标记对象存储在一个独立资源库中。
- 存储对象之后

  如果希望标记存储在 Content Manager 系统中的对象，您可以检索对象，标记它，并用已标记版本替换未标记对象，或者将已标记对象存储为一个新项。
- 检索对象时

  如果要应用的标记根据接收方有所不同，则可以在检索对象之后动态应用标记。然后，交付已标记对象以取代原始对象。

  如果系统中有很多未标记的旧对象且您不希望花时间或使用资源返回并标记每个对象，则您可以在检索对象时标记它们。

使用标记技术

您拥有多种可供选择的内容标记技术。每种技术都针对某个特定的问题，并且在对除去和修改的阻止能力方面有所不同。

标记具有如下特征:

- 传送的信息

  水印 标识内容的源。它可以包含诸如对象的所有者和版本的信息。

  指纹 标识内容的接收方。它可以包含诸如交付对象的地点和对象的信息。

- 可见性

  可见 标记可见且会被注意到。

  不可见 标记隐藏在图像中。

- 完整性

  脆弱的 标记已为某次修改所破坏。

  健壮的 标记阻止了对已标记对象的修改（如调整大小、压缩、旋转和裁剪）。

- 应用的时间
  - 捕捉对象时
  - 存储对象时
  - 检索对象以对顾客分发时
  - 在接收方工作站接收对象时
- 位置
  - 如果使用可见标记的目的在于威慑非法重复使用，则可以将其应用于图像的大部分。可以移动标记使其覆盖图像中更多有纹理的区域，从而使其更加难以除去。
  - 如果使用不可见标记，则图像的纹理区域允许在对图像造成最小影响的情况下嵌入数据。
  - 如果使用可见标记的目的在于表示所有者，则可以将其不显著地放置在图像的角落。
  - 如果同时使用可见和不可见标记，则应当先应用可见标记。
- 格式

  二进制 标记可以是遍布图像重复的随机位序列。此随机序列是可用于对图像进行标记或去除标记的密钥。

  标记也可以是图像。

  结构化数据

  标记可以是嵌入的文本数据。

可见标记

可见标记是放置在图像上的透明标记，在这种方式下标记和图像都可见。难以除去的标记是对盗用您对象的行为的有效震慑。

以下情形请使用可见标记:

- 希望提供图像供顾客复查，但是阻止顾客重复使用复查副本时
- 希望将图像用于万维网上的广告时

不可见标记

不可见标记是隐藏在图像中的数据，在该方式下图像显示为不可更改。需要应用程序以应用、检测和加密标记。

以下情形下请使用不可见标记:

- 希望嵌入信息以标识所有权并阻止已标记对象的非法副本时（水印识别）
- 希望嵌入信息以跟踪分发路径时（指纹识别）
- 希望在图像中嵌入注释或标题时

第 11 章 辅助选项功能部件

本产品包含很多功能部件，使其对于残疾人员来说更容易访问。这些功能部件包括:

- 使用键盘代替鼠标操作所有功能部件的能力。
- 支持增强的显示特性
- 与辅助技术兼容
- 与操作系统辅助选项功能部件兼容
- 可访问的文档格式

键盘输入和导航

以下功能部件可用于键盘输入和导航:

键盘输入

您可以使用键盘代替鼠标操作产品。

菜单项和控件提供了访问键，使您可以从键盘直接激活控件或选择菜单项。这些键是自记录的；访问键在它们出现的控键或菜单上加有下划线。

键盘焦点

在基于 Windows 的系统中，键盘焦点的位置是突出显示的，指示窗口的哪些区域是活动的以及击键会在哪里产生作用。

响应时间调整

在基于 Windows 的系统中，您可以通过控制面板调节响应时间。

可访问显示器的功能部件

客户机具有很多功能部件，可以增强用户界面、改善对视力不佳用户的辅助选项。这些增强包括对高对比度设置和可定制字体特性的支持。

高对比度方式

客户机支持操作系统提供的高对比度方式选项。此功能部件支持背景和前景色之前的较高对比度。

字体设置

在基于 Windows 的系统中，您可以指定显示设置，从而确定菜单和对话窗口中文本的颜色、大小和字体。客户机允许您选择文档列表的字体。

独立于颜色

使用本产品的任何功能，您都不需要在颜色之间进行区分。

与辅助技术兼容

客户机与屏幕阅读器应用程序（如 Narrator 和 Via Voice）兼容。客户机具有这些辅助选项应用程序所必需的特性，从而使得在屏信息对于视力不佳的用户是可见的。

可访问文档

本产品的文档以 PDF 格式提供。您可以使用来自 Adobe（access.adobe.com）的免费工具将 PDF 文件转换为 HTML 或文本。这就使得用户可以根据他们浏览器中设置的显示首选项来查看文档。它还允许使用屏幕阅读器和其它辅助技术。

声明

本信息是为在美国提供的产品和服务编写的。

IBM 可能在其它国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务，则由用户自行负责。

IBM 公司可能已拥有或正在申请与本文档中所描述的内容有关的各项专利。提供本文档并未授予用户使用这些专利的任何许可证。您可以用书面方式将许可证查询寄往:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节（DBCS）信息的许可证查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106, Japan

本条款不适用联合王国或任何这样的条款与当地法律不一致的国家或地区: 国际商业机器公司以“按现状”的基础提供本出版物，不附有任何形式的（无论是明示的，还是默示的）保证，包括（但不限于）对非侵权性、适销性和适用于某特定用途的默示保证。某些国家或地区在某些交易中不允许免除明示或默示的保证。因此本条款可能不适用于您。

本信息中可能包含技术方面不够准确的地方或印刷错误。此处的信息将定期更改；这些更改将编入本出版物的新版本中。IBM 可以随时对本出版物中描述的产品和／或程序进行改进和／或更改，而不另行通知。

本信息中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的: （i）允许在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及（ii）允许对已经交换的信息进行相互使用，请与下列地址联系:

IBM Corporation
J46A/G4

555 Bailey Avenue
San Jose, CA 95141-1003
U.S.A.

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可证协议或任何同等协议中的条款提供。

此处包含的任何性能数据都是在受控环境中测得的。因此，在其它操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。此外，有些测量是通过推算而估计的，实际结果可能会有差异。本文档的用户应当验证其特定环境的适用数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其它可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其它关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本资料包括日常业务运作中的数据和报告示例。为尽可能表述完整，这些示例包含人名及公司、品牌和产品的名称。所有这些人名或名称均系虚构，如有实际的人名或企业名称和地址与此雷同，纯属巧合。

版权许可证:

本资料包括以源语言编写的样本应用程序，这些样本应用程序说明不同操作平台上的编程技术。如果目的是为了开发、使用、经销或分发这样的应用程序，即符合为其编写本样本程序的操作平台的应用程序编程接口的应用程序，则可以任何形式复制、修改、分发这些样本程序，而无须向 IBM 付费。这些程序没有在所有情况下进行过彻底测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。如果目的是为了开发、使用、经销或分发这样的应用程序，即符合 IBM 的应用程序编程接口的应用程序，则可以任何形式复制、修改、分发这些样本程序，而无须向 IBM 付费。

商标

以下术语是国际商业机器公司在美国和 / 或其它国家或地区的商标:

IBM
400
Advanced Peer-to-Peer Networking
AIX
AIXwindows
APPN
AS/400
C Set ++
CICS
DATABASE 2
DataJoiner
DisplayWrite
e-business
HotMedia
Hummingbird
ImagePlus
IMS
Micro Channel
MQSeries
MVS/ESA
NetView
OS/2
PowerPC
PTX
QBIC
RS/6000
SecureWay
SP
VideoCharger
Visual Warehouse
VisualAge
VisualInfo
WebSphere

DB2
DB2 Universal Database
OS/390
PAL

Approach、Domino、Lotus、Lotus 1-2-3、Lotus Notes 和 SmartSuite 是 Lotus Develpoment Corporation 在美国和／或其它国家或地区的商标或注册商标。

Intel 和 Pentium 是 Intel Corporation 在美国和／或其它国家或地区的商标或注册商标。

Microsoft、Windows 和 Windows NT 是 Microsoft Corporation 在美国和／或其它国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国和／或其它国家或地区的商标或注册商标。

UNIX 是 The Open Group 在美国和其它国家或地区的注册商标。

其它公司、产品和服务名称可能是其它公司的商标或服务标记。

词汇表

本词汇表定义了特定于此系统的术语和缩写。以*斜体*显示的术语在本词汇表的其它地方有所定义。

〔A〕

按图像内容查询（query by image content，QBIC）：一种查询技术，使搜索可以基于可视内容（称为特征），而不是纯文本。使用 QBIC，您可以根据对象的可视特征（如颜色和纹理）搜索对象。

〔B〕

本机实体（native entity）：一种*对象*，在特定的*内容服务器*上管理，由*本机属性*构成。例如，Content Manager *索引类*是由 Content Manager *关键字段*组成的本机实体。

本机属性（native attribute）：对象的特征，在特定的*内容服务器*上管理，并且特定于该内容服务器。例如，*关键字段*策略编号可能是 Content Manager 内容服务器上的本机属性，而字段策略标识则可能是 Content Manager OnDemand 内容服务器上的本机属性。

本机文本索引（native text index）：在特定*内容服务器*上管理的文本*项*的索引。例如，Content Manager 内容服务器上的单个文本搜索索引。

部件（part）：请参阅*对象（object）*。

〔C〕

参数搜索（parametric search）：一种基于对象*特性*的*对象*查询。

操作列表（action list）：一个经核准的操作列表，由系统管理员或一些其它*工作流协调者*定义，用户可以在*工作流*或文档转送过程中执行。

查询字符串（query string）：指定查询的特性或特性值的字符串。您可以在应用程序中创建查询字符串并将其传递到查询。

超类（superclass）：从其中派生出类的类。在类和超类之间可能有一个或多个类。

超文本标记语言（Hypertext Markup Language，HTML）：符合 SGML 标准的标记语言，主要为支持包括超文本链接的文本和图形信息的联机显示而设计。

持久标识（persistent identifier，PID）：唯一标识*对象*的标识，无论此对象存储在什么位置。PID 由项标识和位置组成。

抽象类（abstract class）：一种面向对象编程类，它表示一种概念；从它派生的那些类表示这种概念的实现。您不能构造抽象类的对象，即不可将其实例化。

〔D〕

登台区域（staging area）：*资源管理器*的工作存储区域。又称为*资源管理器高速缓存*。

登台（staging）：将已存储*对象*从脱机或低优先级设备移回联机或高优先级设备的过程，通常在系统要求或用户请求时发生。用户请求存储在永久存储器中的对象时，一个工作副本写入*登台区域*。

迭代器（iterator）：一个类或者结构，您可以用它在一系列对象中步进，每次一个。

动态数据对象（dynamic data object，DDO）：应用程序中已存储对象的一般表示，用于将对象移入或移出存储器。

对象服务器高速缓存（object server cache）：请参阅*资源管理器高速缓存（resource manager cache）*。

对象服务器（object server）：请参阅*资源管理器（resource manager）*。

对象链接与嵌入（Object Linking and Embedding，OLE）：Microsoft® 公司用于链接和嵌入应用程序的规范使其可以在其它应用程序中激活的规范。

对象（object）：用户可以将其作为一个单独单元存储、检索和操作的所有数字内容，例如，*JPEG* 图像、MP3 音频、*AVI* 视频和来自书籍的文本块。

多媒体文件系统（multimedia file system）：一个为存储和传递视频和音频而优化的*文件系统*。

多媒体（multimedia）：从计算机将不同媒体元素（文本、图形、音频、静止图像、视频、动画）结合起来以进行显示和控制。

多用途的网际邮件扩充（Multipurpose Internet Mail Extensions，MIME）：请参阅 *MIME 类型（MIME type）*。

〔E〕

二进制大型对象（binary large object，BLOB）： 大小从 0 字节到 2 吉字节的字节序列。此字符串不具有关联的代码页和字符集。图像、音频和视频对象可以存储在 BLOB 中。

〔F〕

方法（method）： 在 Java 设计或编程中，实现操作指定行为的软件。与 C++ 中的成员函数同义。

访问控制表（access control list）： 由一个或多个用户标识或用户组及其关联*特权*组成的列表。您可以使用访问控制表控制用户对 Enterprise Information Portal 系统中*搜索模版*的访问。

访问控制（access control）： 确保某些功能和已存储*对象*仅可由已授权用户通过已授权方式访问的过程。

服务器定义（server definition）： 特定*内容服务器*的特征，将其唯一标识给 Enterprise Information Portal。

服务器类型定义（server type definition）： 由管理员标识的特征列表，用于将特定类型的定制服务器唯一标识给 Enterprise Information Portal。

服务器目录清单（server inventory）： 来自指定*内容服务器*的*本机实体*和*本机属性*的综合列表。

〔G〕

高速缓存（cache）： 一种特殊用途的缓冲区，比主存储器小但比主存储器快，用于保存可以频繁访问的数据副本。高速缓存的使用减少了访问时间，但可能会增加内存需求。

根组件（root component）： 分层*项类型*的第一级别或唯一级别，由相关系统或用户定义的*属性*组成。

公共网关接口（Common Gateway Interface，CGI）： 用于 Web 服务器和其外程序之间信息交换的标准。外部程序可以用任何编程语言编写，只要 Web 服务器运行于其上的操作系统支持该编程语言即可。请参阅 *CGI 脚本（CGI script）*。

公用交换单元（common interchange unit，CIU）： 传送公用交换文件（CIF）的独立单元。它是 CIF 中用于标识与接收数据库之间的关系的那一部分。一个 CIF 可以包含多个 CIU。

公用交换文件（common interchange file，CIF）： 包含一个 ImagePlus Interchange Architecture（IPIA）数据流的文件。

工作包（work packet）： 在 Enterprise Information Portal V7.1 中，指从一个位置到另一个位置传送的*文档*的集合。用户通过*工作列表*访问和使用工作包。

工作步骤（work step）： *工作流*或*文档转送过程*中的离散点，单独的*工作项*、*文档*或*文件夹*必须它。

工作列表（worklist）： 指定给用户的*工作项*、*文档*或*文件夹*的集合。

工作流状态（workflow state）： 整个*工作流*的状态。

工作流（workflow）： Enterprise Information Portal 中的一系列*工作步骤*和管理那些步骤的规则，*工作篮*、*文档*或*文件夹*在处理时通过它们。

例如，索赔核准将描述独立保险索赔核准时必须遵循的过程。

工作项（work item）： 早期 Content Manager 工作流和 Enterprise Information Portal 高级工作流中，*工作流*中活动的工作活动。

工作状态（work state）： 单独的*工作项*、*文档*或*文件夹*的状态。

构造器（constructor）： 编程语言中的一种方法，与类有相同的名称，用于创建和初始化该类的对象。

关键字段（key field）： 请参阅*属性（attribute）*。

〔H〕

绘制（render）： 把通常不是面向图像的数据描绘或显示成为图像。在 Content Manager 中为了能够显示，通常把字处理文档绘制成图像。

〔J〕

基数（cardinality）： 数据库表中的行数。

集合（collection）： 具有一系列相似管理规则的一组对象。

吉字节（gigabyte，GB）： (1) 对于处理器存储、真实和虚拟存储和通道容量，指 2^{30} 或 1 073 741 824 字节。 (2) 对于磁盘存储容量和通信容量，指 1 000 000 000 字节。

建立索引（index）： 添加或编辑标识特定*项*或*对象*的属性值，以便以后可检索它。

交错式音频 / 视频（Audio/Video Interleaved，AVI）： 一种 RIFF（*资源交换文件格式*）文件规范，允许音频和视频数据在文件中交错。可以在维护文件设备上顺序访问的同时在交替块中访问（以便回放和录制）的独立磁道。

交换（interchange）： 使用*公共交换文件*或*公共交换单元*将图像与其索引一同从一个 Content Manager ImagePlus for OS/390 系统导入或导出到另一个 ImagePlus 系统。

接口类（connector class）： 面向对象编程类，它提供对那些对特定*内容服务器*来说为本机内容的 API 的标准访问。

局域网（local area network，LAN）： 一种网络，一组设备在其中彼此连接以进行通信，且该网络可以连接到一个更大的网络。

句柄（handle）： 表示对象并用于检索对象的字符串。

卷（volume）： 系统中用于存储对象的实际物理存储设备或部件的一种表示方法。

〔K〕

可扩展标记语言（Extensible Markup Language，XML）： 一种标准元语言，用于定义由 SGML 派生且为其子集的标记语言。XML 省略了 SGML 中较复杂且较少使用的部分，并且使得编写应用程序以处理文档类型和作者，管理结构化信息以及通过各种计算系统传输和共享结构化信息变得更加容易。XML 的使用不需要 SGML 所必需的健壮应用程序和处理。XML 正在万维网协会（W3C）的协助下进行开发。

客户机应用程序（client application）： 利用面向对象或因特网 API 编写的一种应用程序，用于从 Enterprise Information Portal 访问*内容服务器*。

客户机／服务器（client/server）： 在通信中，指分布式数据处理的交互模型，程序在其中将请求从一个位置的程序发送到另一个位置的程序并等待响应。请求程序称为客户机；应答程序称为服务器。

库服务器（library server）： Content Manager 系统的组件，它存储、管理并处理*项*上的查询。

库客户机（library client）： Content Manager 系统的组件，它为库系统提供底层的编程接口。库客户机包含为软件开发人员工具箱一部分的 API。

扩展数据对象（extended data object，XDO）： 应用程序中已存储的复杂多媒体*对象*（用于将对象移入或移出存储器）的一般表示。XDO 通常包含在 *DDO* 中。

〔L〕

类（class）： 面向对象设计或编程中，可被实例化以创建具有公共定义，并因此而具有公共特性、操作和行为的对象的模型或模板。一个对象就是一个类的实例。

历史记录（history log）： 保存*工作流*活动记录的文件。

联合集合（federated collection）： 是进行*联合搜索*后得出的一组对象。

联合实体（federated entity）： Enterprise Information Portal 元数据对象，由*联合属性*组成，可以选择与一个或多个*联合文本索引*关联。

联合数据存储器（federated datastore）： 任意数量的特定*内容服务器*（如 Content Manager）的虚拟表示。

联合属性（federated attribute）： 一种 Enterprise Information Portal 元数据类别，可映射为一个或多个*内容服务器*上的*本机属性*。例如，关联属性策略编号可以在 Content Manager 中映射为*属性*、策略编号 ，也可以在 Content Manager ImagePlus for OS/390 中映射为属性、策略标识。

联合搜索（federated search）： Enterprise Information Portal 发出的查询，同时在一个或多个*内容服务器*上搜索数据，并且这些服务器可以是不同种类的。

联合图像专家组（Joint Photographic Experts Group，JPEG）： (1) 一个致力于建立压缩数字化连续调和图像的标准的小组。 (2) 此小组开发的静态图像标准。

联合文本索引（federated text index）： Enterprise Information Portal 元数据对象，可在一个或多个*内容服务器*上映射为一个或多个*本机文本索引*。

链接（link）： 两*项*（源和目标）之间的导向性关系。您可以使用一系列链接建立一到多关联的模型。请对照*引用（reference）*。

流式数据（streamed data）： 以指定的速率通过网络连接发送的任何数据。流可以为一种数据类型或类型组合。数据速率（以位／秒表达）根据流和网络的不同而有所不同。

〔M〕

媒体服务器（media server）： Content Manager 系统的一种基于 AIX 的组件，用于存储和访问视频文件。

媒体归档程序（media archiver）： 一种用于存储音频和视频流式数据的物理设备。VideoCharger 是一种媒体归档程序。

〔N〕

内容服务器（content server）： 一种软件系统，用于存储多媒体和商务数据及用户在处理这些数据时所需的相关元数据。Content Manager 和 Content Manager ImagePlus for OS/390 是内容服务器的示例。

〔P〕

匹配度（rank）： 一个整数值，表示给定部分与查询结果的关系。较高的匹配度表示更接近一些的匹配。

〔S〕

释放（release）： 将暂挂条件从*项*除去。满足条件时，暂挂项就被释放，或者具有适当权限的用户覆盖条件时，就可手工释放它。

瘦客户机（thin client）： 一种客户机，安装很少的软件或不安装软件，但对由连接到它的网络服务器管理和传递的软件拥有访问权。瘦客户机是全功能客户机（如工作站）的替代项。

数据包（package）： 提供访问保护和名称空间管理的相关*类*和接口的集合。

数据存储器（datastore）： (1) 存储数据的位置（诸如数据库系统、文件或目录）的通称。 (2) 在应用程序中，它是*内容服务器*的虚拟表示。

数据格式（data format）： 请参阅 *MIME 类型（MIME type）*。

属性（attribute）： 一种数据单元，描述项的某些特征或特性（例如，名称、地址、年龄等），也可用于定位项。属性具有类型，指示了由该属性存储的信息范围以及该范围中的值。例如，关于多媒体文件系统中某个文件的信息，如标题、运行时或编码类型（MPEG1、H.263 等等）。对于 Enterprise Information Portal，另见*联合属性（federated attribute）*和*本机属性（native attribute）*。

搜索模板（search template）： 一种格式，由管理员设计的*搜索条件*组成，用于特定的联合搜索类型。管理员还标识了可以访问每个搜索模板的*用户*和*用户组*。

搜索条件（search criteria）： Enterprise Information Portal 中的特定字段，管理员为*搜索模板*定义，用于限制或更进一步定义*用户*可用的选择。

索引类视图（index class view）： 在早期 Content Manager 中，API 中用于*索引类子集*的术语。

索引类子集（index class subset）： 早期 Content Manager 中*索引类*的一个视图，应用程序用它存储、检索和显示文件夹和对象。

索引类（index class）： 请参阅*项类型（item type）*。

〔T〕

套印版面（overlay）： 指一个预定义的数据（例如线、阴影、文本、框或徽标）集合，在打印期间可以与页面中的可变数据相结合。

特权集（privilege set）： 用于处理系统组件和功能的一组*特权*。管理员将特权集指定给用户（用户标识）和*用户组*。

特权（privilege）： 以特定方式访问特定*对象*的权限。特权包括如在系统中创建、删除和选择存储对象等的权限。特权由管理员指定。

特性（property）： 描述对象的*对象*特征。特性可以更改或修改。字形就是特性的一个示例。

特征（feature）： 存储在图像搜索服务器中的可视内容信息。也指图像搜索应用程序用以确定匹配与否的可视特性。四种 *QBIC* 特征是: 平均颜色、直方图颜色、位置颜色和纹理。

通配符（wildcard character）： 一种诸如星号（*）或问号（?）的特殊字符，可以用以表示一个或多个字符。任何字符或字符集都可以替换通配符。

统一资源定位器（Uniform Resource Locator）： 表示计算机上或网络（如因特网）中信息资源的一系列字符。此字符序列包含用于访问信息资源的协议的缩写名称和协议用于定位信息资源的信息。例如，在因特网环境中，这些是用以访问各种信息资源的协议的缩写名称: http、ftp、gopher、telnet 和 news。

图像对象内容体系结构（Image Object Content Architecture，IOCA）： 用于交换和表示图像的结构集合。

〔W〕

网关（gateway）： 使具有不同网络体系结构的两个计算机网络互相连接的功能性部件。网关连接不同体系结构的网络或系统。桥连接具有相同或相似体系结构的网络或系统。

网络表文件（network table file）： 一个文本文件，其中包含 Content Manager 系统中每个节点的特定于系统的配置信息。系统中的每个节点必须具有标识节点本身和需要连接到的节点列表的网络表文件。

网络表的名称是 FRNOLINT.TBL。

文档类型定义（document type definition，DTD）： 为 XML 文档的特殊类指定结构的规则。DTD 定义了元素、属性和符号的结构，并且建立了每个元素、属性和符号可

以在文档的特殊类中如何使用的约束。DTD 与数据库模式很相似，因为 DTD 完整地描述了特殊标记语言的结构。

文档（document）： 可以作为一个独立单位在 Content Manager 系统和用户之间进行存储、检索和交换的*项*。具有文档*语义类型*的项，它应当包含形成文档的信息，但没有必要暗示它是一种 Content Manager 文档模型的实现。

从文档已分类项类型（Content Manager 文档模型的特定实现）创建的项必须包含文档部分。可以使用文档已分类项类型来创建类型为文档或文件夹语义的项。

文档部分可包括各种类型的内容，例如包括文本、图像和电子数据表。

文件夹管理器（folder manager）： 将数据作为联机文档和文件夹管理的 Content Manager 模型。您可以将文件夹管理器 API 用作应用程序和 Content Manager 内容服务器之间的主要接口。

文件夹（folder）： 任何*项类型*的*项*（不管分类）的类型均为文件夹*语义*。除了所有非资源项功能和可从项类型分类（例如*文档*或资源项）使用的任何其它功能，类型为文件夹语义的任何项还包含由 Content Manager 提供的特定文件夹功能。文件夹可以包含任意类型的任意个项，包括文档和子文件夹。文件夹按*属性*建立索引。

文件系统（file system）： 在 AIX 中，对硬盘驱动器分区以进行存储的方法。

〔X〕

项类型分类（item type classification）： *项类型*中的归类，进一步标识了该项类型的*项*。相同项类型的所有项都具有相同的项类型分类。

Content Manager 提供以下项类型分类: *文件夹*、*文档*、对象、视频、图像和文本; 用户还可以定义他们自己的项类型分类。

项类型（item type）： 用于定义和以后定位相似*项*（由*根组件*、零个或多个*子组件*及分类组成）的模板。

项（item）： Enterprise Information Portal 管理的最小信息单位的通称。每个项都有一个标识。例如，项可以是*文件夹*或*文档*。

信息发掘（Information Mining）： 从文本中抽取关键信息（摘要）、在文档集合中查找显著的主题（归类）并使用强大、灵活的查询搜索相关文档的自动化处理过程。

〔Y〕

引用（reference）： 根组件或*子组件*和另一个*根组件*之间的单向、一对一关联。请对照*链接（link）*。

应用程序编程接口（application programming interface，API）： 使应用程序之间能够通信的软件接口。API 是编程语言构造或语句的集合，这些构造或语句可以编写为应用程序，以得到潜在许可程序提供的特定功能和服务。

用户出口例程（user exit routine）： 在预定义*用户出口*接收控制的用户编写的例程。

用户出口（user exit）： 是 IBM 提供的程序中的点，用户出口例程可在此获得控制。

用户映射（user mapping）： 将 Enterprise Information Portal 用户标识和密码关联到一个或多个内容服务器中相应的用户标识和密码。用户映射使用户仅需登录一次，就可登录到 Enterprise Information Portal 和多个*内容服务器*上。

用户组（user group）： 由一个或多个已定义的独立*用户*组成的组，以同一个组名称标识。

用户（user）： 在 Enterprise Information Portal 中，指 Enterprise Information Portal 管理程序标识的所有用户。

游标（cursor）： 应用程序使用的一种已命名的控制结构，用于指向某些已排序行的集合中特定的行。游标用于从集合中检索行。

语义类型（semantic type）： *项*的用法或规则。Content Manager 提供的语义类型是基本、注释和注意; 用户也可以定义他们自己的语义类型。

远程方法调用（Remote Method Invocation，RMI）： 支持分布式编程的 API 集合。一个 Java 虚拟机（JVM）中的对象可以调用其它 JVM 中的对象上的方法。

〔Z〕

暂挂（suspend）： 将*对象*从其*工作流*除去，并定义激活它所需的暂挂条件。随后激活此对象的目的是使它可以被继续处理。

资源管理器: 管理*对象*的 Content Manager 系统的组件。这些对象由存储在*库服务器*上的*项*引用。

资源管理器高速缓存（resource manager cache）： *资源管理器*的工作存储区域。又称为*登台区域*。

资源交换文件格式（Resource Interchange File Format，RIFF）： 用于存储声音或图形以在不同类型的计算机设备上回放。

子类（subclass）： 一种从其它类派生出的类。在类和子类之间可能有一个或多个类。

子组件（child component）： 分层*项类型*的可选第二或较低级别。每个子组件与其上一级别直接关联。

自述文件（README file）： 安装或运行与其关联的程序之前应当查看的文件。自述文件通常包含最终产品信息、安装信息或使用产品的技巧。

组合搜索（combined search）： 组合了以下一种或一种以上搜索类型的查询: *参数*、文本或图像。

组件（component）： *根组件*或*子组件*的通称。

A

ADSM: 请参阅 *Tivoli® Storage Manager*。

API: 请参阅*应用程序编程接口（application programming interface）*。

AVI: 请参阅*交错式音频／视频（Audio/Video Interleaved）*。

B

BLOB: 请参阅*二进制大型对象（binary large object）*。

C

CGI: 请参阅*公共网关接口（Common Gateway Interface）*。

CGI 脚本（CGI script）： 一种计算机程序，在 Web 服务器上运行并使用*公共网关接口（CGI）*执行通常不由 Web 服务器执行的任务（例如，数据库访问和表单处理）。CGI 脚本是一种 CGI 程序，它是用诸如 Perl 的脚本语言编写的。

CIF: 请参阅*公共交换文件（common interchange file）*。

CIU: 请参阅*公共交换单元（common interchange unit）*。

D

DDO: 请参阅*动态数据对象（dynamic data object）*。

DTD: 请参阅*文档类型定义（document type definition）*。

G

GB: 请参阅*吉字节（gigabyte）*。

H

HTML: 请参阅*超文本标记语言（Hypertext Markup Language）*。

I

IOCA: 请参阅*图像对象内容体系结构（Image Object Content Architecture）*。

J

JavaBeans: 一种独立于平台的软件组件技术，用于构建称为 “bean” 的可重复使用的 Java 组件。构建之后，这些 bean 可以由其它软件工程师或在 Java 应用程序中使用。使用 JavaBeans，软件工程师可以操作 bean 并将其聚集在图形化的拖放开发环境中。

JPEG: 请参阅*联合图像专家组（Joint Photographic Experts Group）*。

L

LAN: 请参阅*局域网（local area network）*。

M

MIME 类型（MIME type）： 标识正在通过因特网传输的对象类型的因特网标准。MIME 类型包括音频、图像和视频的多种变量。每个对象都具有 MIME 类型。

O

OLE: 请参阅*对象链接与嵌入（Object Linking and Embedding）*。

P

PID: 请参阅*持久标识（persistent identifier）*。

Q

QBIC: 请参阅*按图像内容查询（query by image content）*。

R

RIFF: 请参阅*资源交换文件格式（Resource Interchange File Format）*。

RMI 服务器（RMI server）： 实现 Java *远程方法调用（RMI）*分布式对象模型的一种服务器。

T

Tivoli Storage Manager（TSM）： 在异构环境中提供存储管理和数据访问服务的*客户机／服务器*产品。它支持各

种通信方法，提供管理文件备份和文件存储的管理工具，并且还提供用于调度备份操作的工具。

TSM: 请参阅 *Tivoli Storage Manager*。

TSM 卷（TSM volume）: 由 *Tivoli Storage Manager 管理的逻辑存储区域*。

X

XDO: 请参阅*扩展数据对象（extended data object）*。

XML: 请参阅*可扩展标记语言（Extensible Markup Language）*。

索引

C

E

I

L

M

程序号: 5724-B43

中国印刷

S152-0232-01