致客户信

尊敬的客户：

我们非常荣幸地向您推荐以下 IBM 企业信息搜索解决方案。IBM® WebSphere® Information Integrator Omnifind 提供了一个企业级搜索中间件的体系架构，是一个灵活、可扩展的系统，使用户可以通过单点的访问即可查询数据众多、形式多样的企业信息。

本方案建议书主要讲述了 IBM WebSphere Information Integrator OmniFind 产品的特点和优势。IBM WebSphere Information Integrator OmniFind 产品是 IBM WebSphere 信息集成家族产品中的重要的基础功能组件。 OmniFind 同时具备高质量搜索、企业级规模、易于管理以及可扩展语义解析的等特性。

IBM 企业搜索解决方案区别其它厂商的突出特点是：

具备的信息集成和内容管理解决方案，提供了对不同种类信息资源进行访问、管理以及整合能力，为用户提供了访问全局、关键性业务信息的统一的信息视图，同时，可以有效利用和保护您现有的投资。

企业信息搜索解决方案是 IBM 强大信息集成基础平台的扩展，该基础平台包括了信息管理、内容管理以及构建企业级应用所需的信息检索技术。

IBM 从2003年9月即使用 OmniFind，通过自身的实践充分证明了 OminiFind 做为企业搜索技术的成功。

基于开放、模块化的解决方案，非常易于与现有的企业应用集成，帮助客户降低构建信息集成基础设施的成本。

IBM 具有广泛、行业专家，可以保证提供的解决方案是高可靠性、可扩展、高安全保证。

感谢您抽出宝贵的时间来浏览这份WebSphere® Information Integrator OmniFind™ Edition 方案建议书。我们期待与您进一步合作，讨论更多的细节。如果需要更多信息，请电话联系IBM 中国 Techline：**800-810-1818 转 5151 或(010)84981188 转 5151 分机**。

敬祝

商祺！

IBM 企业信息搜索解决方案

WebSphere II OmniFind Edition

方案建议书

IBM（中国）有限公司

2006年10月

目　录

1.综述 3

1.1需求分析 3

1.2方案概要 4

1.3用户受益 5

2.方案概览 6

2.1 OMNIFIND设计原则 7

2.2 OMNIFIND技术架构 8

2.3 管理控制台 10

2.4 支持数据源 10

2.5 安全性 11

2.6 门户集成和定制 12

2.7 搜索结果匹配排序（RANKING） 13

2.8 集合，分类以及领域 15

2.9 搜索技术 15

2.10. 可用性和可扩展性 16

3. OMNIFIND应用案例 17

3.1 华为E-SUPPORT项目信息搜索 17

3.2 IBM企业内网信息搜索 18

1. 综述

IBM® WebSphere® Information Integrator OmniFind™ Edition为客户提供一个单点访问、集中查询众多、形式多样的企业信息的服务能力。OminiFind 做为企业搜索中间件平台，能在次秒级的响应并提供相关的结果集，而无需关心业务数据信息处于什么位置，包括Web站点、关系数据库、文件系统、新闻组、门户系统、协作系统、应用程序以及内容管理系统等。

企业搜索技术的最终目标是要满足客户的要求。客户希望迅速、直接地找到所需的信息，而无需在不相关的结果中查找。WebSphere Information Integrator OmniFind Edition 就是为了满足全球查询用户的这个目标而设计的。

1.1. 需求分析

随着企业信息系统地建立和发展，产生了大量的业务信息。其中不仅有数据库中业务交易信息、客户的资料等数据库中存储的结构化信息，而且还有大量产品资料、服务记录、往来邮件、事件处理说明、规章制度手册、工作记录报告等非结构化的文本信息内容。如何从企业纷繁复杂的信息资源中，找到用户所需要的内容是信息管理的一个巨大挑战。

越来越多的客户希望利用搜索技术查找到企业内部相关的内容，就像在互联网上各种常用的搜索网站一样便捷。然而，与互联网搜索相比，企业信息搜索有着以下几点显著的差异：数据来源丰富多样，信息访问控制的安全性，以及与业务处理的集成等。

与互联网不同，在企业内部信息不仅分布在网站上，大量有价值的信息存储在文件系统、内容管理资料库、数据库及邮件系统中。格式可能是文本、XML、Word 文档、PDF 及 PPT 文件等。这些信息可能有不同的安全访问级别、对不同的用户需控制其访问的信息内容，往往都要求做到文档级别的安全性管理。另外企业内部信息搜索应用的目的性更强，往往还要求搜索的结果能够与企业现有的业务处理进行紧密地关联，使搜索能够为更灵活的业务处理流程服务，如减少寻找客户资料的时间、提供客户网上自助服务的快捷查询手段等。

另外，对自然语言的解析能力，高效、高质量的查询结果，很强的系统可扩展性，对海量数据及大并发用户访问的支持，丰富的 API 应用开发接口，易于二次开发及与现有应用的集成能力，也都成为衡量企业信息搜索的关键因素。

随着先进、成熟的信息搜索技术的发展和推广，企业信息搜索正在逐渐成为一种更为普及更为通用的信息访问接入手段。

1.2. 方案概要

用户希望以自己熟悉的方式来搜索信息，就像他们习惯在 Internet 上搜索的体验一样，进行特定的、实时的、随意形式的进行文本搜索。对企业而言，信息过时或信息不可见的成本是影响内部生产成本和直接收入的重要因素。那些隐藏或不可及的数据信息对企业而言不仅仅影响生产效率，同时也会使最终用户对企业的形象产生冲击。

面对企业信息搜索需求，IBM 发布了 IBM WebSphere Information Integrator OmniFind Edition(企业信息搜索引擎，以下简称 OmniFind)，它作为 IBM 总体信息整合平台的一部分专门定位在企业信息搜索领域。OmniFind 具有查询不同类型的数据源、准确分词、快速返回结果以及支持海量数据大并发访问的能力，这将有助于企业更好地洞察它们的运营情况，并更好地利用企业现有的数据资源，快速准确地定位企业中最佳的相关内容。

IBM WebSphere Information Integrator 企业信息搜索解决方案，核心是帮助客户迅速、方便地找到正确的信息，而不用费时、复杂的 IT 解决方案。

WebSphere Information Integrator OmniFind Edition 针对企业信息搜索的要求，提供了信息相关性搜索功能，次秒级的响应速度，关联算法和语言分析的能力。

OmniFind 有能力将客户的整个企业的信息进行分析、建立索引。只需通过简单输入一个关键字或者短语，你就能迅速搜索企业 intranet、企业公共 Web 站点、关系数据库系统、文件系统以及内容知识库。WebSphere Information Integrator OmniFind Edition 有能力搜索你所有的企业信息，不论它们位于何处。

做为验证，IBM 已经在自己的 Intranet 使用该产品达一年之久，IBM 的 Intranet 是世界上最大的公司 Intranet 之一，包括了一万多个不同的 Web 站点和九百万个独立的页面。IBM 的 Intranet 服务器承担着多达三十万人的使用、每天八万多次的查询，同时，保持次秒级的响应速度、准确率和 IBM 严格的 24x7 标准。

1.3. 用户受益

IBM WebSphere Information Integrator OmniFind Edition 产品帮助客户实现企业级信息搜索功能，它将成为您公司全面信息集成平台的一部分。提供了高质量、可扩展、安全性、基于任意文本形式的搜索，并能使员工、供应商、合作伙伴和最终客户找到最相关的企业信息。

WebSphere Information Integrator OmniFind Edition 给客户带来许多技术和业务收益：

- 使用多种语言学分析和搜索排序技术，返回高质量的搜索结果集
- 支持种格式的数据源，包括 Web 内容、内容管理系统、关系数据库、协作系统等等。提供可扩展的架构，使其能连接更多其他的数据源。
- 次秒级响应的高性能特性
- 支持上百万文档和成千上万用户的可扩展规模
- 提供集中式监控管理。 具备透明的分析功能，只需极少的管理工作，即可保证获得高质量的搜索结果。
- 提供灵活的安全保护机制，以防企业资产泄漏。
- 提供开箱即有的门户 Portle 样例，可以方便与门户环境集成，同时，提供 IBM 索引、搜索接口的 Java API。
- 为 IBM WebSphere Portal 搜索引擎的客户提供了迁移的方法，从而提供更广泛的、可扩展的内容范围，获得更具相关性的搜索结果。

IBM WebSphere Information Integrator OmniFind Edition 是 IBM 信息管理和信息整合产品家族的一个部分。该产品的目标是帮助客户实现随需应变的信息服务，展示一个统一、综合的信息视图，而不用关心数据的格式、位置和访问接口。

现有，许多组织还在挣扎着基于那些信息孤岛，来支持决策制定，并依赖这些信息实现业务流程。WebSphere Information Integrator 产品使企业 intranet 功能增强，提供高质量的企业搜索、次秒级响应速度，并将正确的信息在正确的时间传递给正确的人。

WebSphere Information Integrator 产品家族提供了自治特性，可以减少 IT 人员管理复杂数据架构的负担。这些能力还可使客户快速响应市场的变化，实现现有资产更大的价值，更好地进行成本控制，基于有效地访问和使用信息支持商业洞察力、实现创新。只有 IBM 有能力将这些基础技术付诸实现。

2. 方案概览

IBM® WebSphere® Information Integrator 产品为客户提供对各种业务数据信息实时、综合的访问能力 — 信息可以包括：结构化和非结构化的，集中以及分散的，公有以及私有的— 包括企业内或企业外部的信息。

搜索功能是信息基础设施所必备的基本能力，它提供了对文本或者非结构化数据(所占比例高达整个企业数据的 85%) 关键的访问能力。IBM 信息集成平台的重要特性，是支持任意形式的查询，仅需关键字或词组即可完成搜索，可以帮客户更好地利用复杂、综合的信息资产。

WebSphere Information Integrator OmniFind Edition 是满足企业级信息搜索需求的主要产品，属于 Websphere II 家族产品的一个重要部分。它提供了高质量、可扩展、高安全性的的文本搜索服务功能，使企业员工、供应商、合作伙伴以及最终客户可以方便找到最相关的企业信息。可以与客户现有系统无缝集成，所包含的企业搜索组件可以有效地进行多种、不同数据源信息的收集，同时建立索引并实现快速查询。通过对信息进行语言学分析和其他类型的数据分析，WebSphere Information Integrator OmniFind Edition 可以为客户提供高关联性的企业信息结果集，次秒级的响应速度，同时支持多种数据源，而无需关心业务数据位于何处，信息可以存在于：Web 站点、关系数据库、文件系统、门户网站、协同系统、应用程序以及内容管理系统。

WebSphere Information Integrator OmniFind Edition易于集成到安全的企业级Java® 应用程序中，因此，私有信息不会轻易暴露给未授权的用户。此外，OmniFind 的架构支持支持数以百万的文档查询和成千上万的用户使用。事实上，OmniFind 做为IBM intranet的企业搜索引擎，正承担着每天 80,000 次的查询，搜索范围跨越九百多万个独立的页面。

WebSphere Information Integrator OmniFind Edition 为客户提供了更加广泛的信息查询能力，它的优势包括：

- **先进的排序技术**—OmniFind 会依据许多因数来查询相关联的文档，而且可以根据输入的查询类型给出不同的权重因数。例如，有一些是普通的查询，如“401K 计划”；而其它的查询则需要非常具体的信息，比方“更改 intranet 密码”。OmniFind 会考虑不同类型的查询、以及影响的因数来完成它的分析，返回相关度更高和更精确的结果集。

- **国际化语言支持**—OmniFind 分析被索引的文档，识别它的语言，使用高级的语言学处理技术分析文本。支持 61 种语言环境的查询，对其中 25 种语言提供了高级语言学处理功能。索引过程进一步分析词汇结构，查重并删除重复内容，并采用其他的技术以改进整个搜索的质量。
- **参数搜索** — OmniFind 支持根据数值范围进行查询 — 用户可以搜索在$10 和$20 之间的价格，或者更接近一具体日期的结果集。
- **分类**—可以帮助用户找到一组相关的信息，进一步改善搜索体验。OmniFind 提供了基于规则的分类系统，同时还支持 IBM WebSphere Portal 门户系统提供的分类法。
- **动态摘要** — 提供对结果集摘要中的查询条件以高亮方式显示，这对用户来讲是非常有用的。OmniFind 能为每个包括所有查询条件的文档动态生成摘要。这种摘要使用户能轻易地确定哪一个文档可能包含更多他们所需要地信息，而无需阅读每一篇结果集中的文档。

2.1. OmniFind 设计原则

OmniFind 企业信息搜索解决方案的设计，充分考虑了如下因素：

1．技术的先进性和成熟性：采用经实践验证的、成熟的、先进的技术，满足生命周期内具有持续的可维护性和扩展性。

2．安全性：充分考虑系统各个层面的安全性，包括运行安全、信息访问安全、基于角色的权限管理等各个方面。

3．对自然语言的处理：支持多种语言，提供准确分词能力，提高信息搜索效率。

4．易管理性：采用合理的系统体系结构，实现对搜索系统的集中管理和监控，易于管理维护。

5．可扩展性：系统架构在开放的 J2EE 安全应用支撑体系结构之上，具有良好的可扩展能力，适应业务量的变化，支持海量数据，支持大并发访问，适用于多种不同的数据来源。

2.2. OmniFind 技术架构

WebSphere Information Integrator OmniFind Edition 提供了一个企业级的搜索中间件体系架构，包括三个主要的组件：爬行器、解析及索引服务器(Parser, Indexer)、搜索服务器。

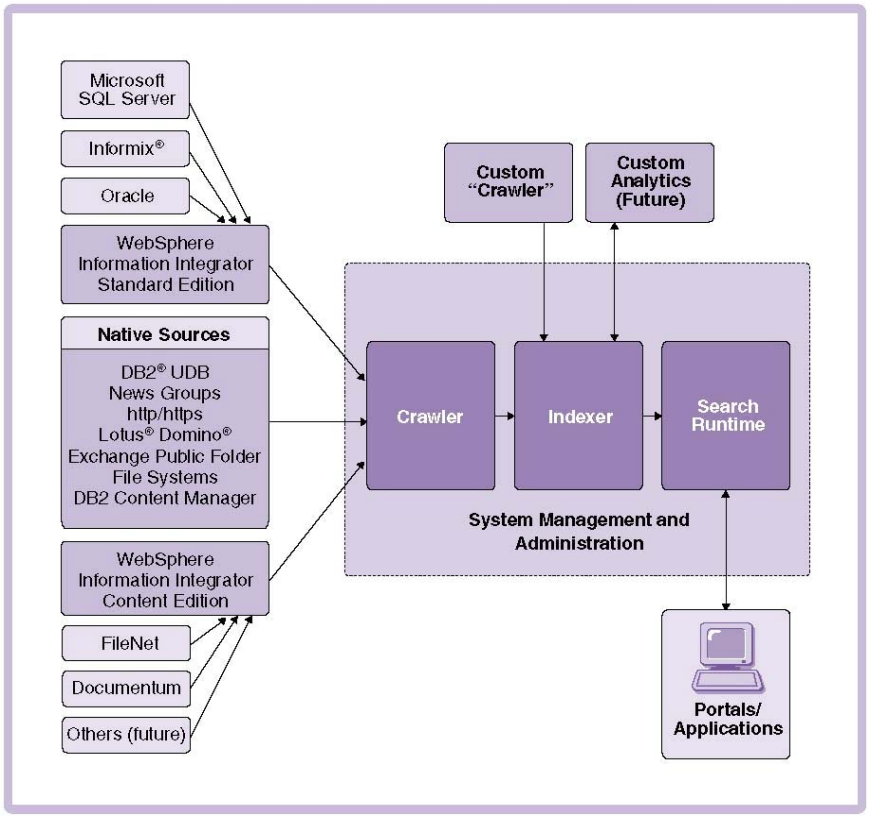

图1. WebSphere Information Integrator OmniFind Edition 技术架构

2.2.1. 爬行器

爬行器会以固定时间间隔，根据管理员的配置，爬行所有数据资源，同时，将从这些数据源提取的内容移入IBM DB2® 数据库中进行存储。管理员可以决定哪些数据源与哪个特定应用相关，同时，还可将这些数据源分成不同的集合。接下来，爬行器就可以从那些数据源收集相关数据及元数据。

爬行器可以定期或 24 小时循环的从各类数据资源中搜集数据。它们了解数据源之间的区别，能提取所有相关信息，包括元数据。被检索的数据可以是数据库中的数据（如一张或多张表的一个或多个字段），支持各类主流数据库，和提供 ODBC3.0 接口的数据源、支持邮件系统数据(包括 Lotus Notes 和 Exchange Server 等)、支持 Word 文档、pdf 文档、支持 Internet 和 Intranet 网站信息、支持 IBM 内容管理的对象数据并通过 IBM 内容集成产品支持 Filenet、

Documentum 等第三方的内容数据资源。并提供用户自定义数据的接口，对用户特有的数据类型进行支持。提供数据代码集的自动转换，在爬行的同时可实现客户化的文档级搜索权限控制。

2.2.2. 解析及索引服务器(Parser, Indexer)

该组件分析由爬行器收集的文档，同时分析他们以创建索引。解析组件分析文档内容和文档元数据信息。它将分析结果以文件系统的方式存储，方便索引组件的访问。

解析和索引服务器，可以按自然语言规则分析文档并构建索引。系统提供多种语言的词条库，自动识别数据资源的语种，并根据词条库中的字典，对数据进行解析分词处理(praser)，支持对简体中文及繁体中文、英文、日文等多种语言的准确分词能力，搜索还支持其他多种语言。OmniFind 将提供分类模型，并可根据行业特点通过服务定制客户化的分类及扩充专业词汇和相应的关联关系，也可然后根据规则由 Indexer 建立相应的索引，索引服务器将进一步处理该信息以分析内部网内容的链接结构、执行复制内容删除，以及对可以增强总体搜索质量的文档集合执行其他处理。索引可根据规则进行分类，构建一类或多类索引。

解析组件支持如下配置：

- **XML 文档的域映射规则 —**使用户能搜索 XML 文档中的结构化和非结构化的内容。如果将 XML 元素映射到搜索域，用户即可在查询中指定具体域名，同时搜索 XML 文档指定的部分。
- **分类 —**能使用户根据文档所属的不同类别来搜索文档。也能在搜索结果中选择文件的类别，浏览属于某些特定类别的文档。

索引服务器可以在固定的时间间隙添加新增、更改的文档信息，同样，以文件系统的形式存储。它分析文档的同时创建索引。在这个过程中，支持多达 25 种语言的语言处理机制进行文本提取和分析。进一步的处理，是由索引服务器分析 intranet 中内容的关联结构、查重并删除重复内容，同时，对文档集合进行其它的处理，全面提高整个搜索质量。索引器的设计可以支持两千万文档的规模。

2.2.3. 搜索服务器

该组件代表搜索应用程序进行工作，负责处理查询、搜索索引，OmniFind 将提供一个或多个搜索引擎，同时将搜索结果返回给搜索应用程序。搜索服务器处理搜索请求，在索引中找到相关度最高的文档，同时，在次秒级的响应并返回结果集, 实现对所有资源的全文检索，最

后将检索的结果返回用户。通过成熟的排序分析功能，搜索服务器可以将最先返回相关度最高的搜索结果（高命中率）。另外，OmniFind 支持多个搜索服务器并行，在部署上支持同时使用两台搜索服务器，提供备份的功能，提供相应的可扩展性及冗余备份能力，确保搜索应用的高效及高可用性。当扩展到数百万文档和数千用户时，OmniFind 可以以次秒级响应时间交付高度相关的搜索结果。

OmniFind 提供二次搜索能力，并通过搜索结果的高速缓存提供系统的响应效率。支持基于属性的检索和全文检索结合。提供专用的搜索描述语言，支持与、或。

此外，整个技术架构是开放的、可扩展的，更易于支持多变的行业或者企业特定搜索应用程序。OmniFind 提供了完善的 Java API 接口，用户可以在此基础上定义自己的搜索应用，可以更方便地将搜索功能集成到客户现有的企业应用程序中。WebSphere Information Integrator OmniFind Edition 提供了制定自定义分析插件的基础，可支持特殊领域的扩展查询功能等。

2.3. 管理控制台

使用管理控制台，客户可以创建、管理信息集合，启动、停止其他的组件（如爬行器，解析器，索引器和搜索服务器），监控整个系统的活动和日志文件，配置管理的用户，以及将搜索应用和相应的信息集合进行关联。针对爬行器、索引器和搜索服务器生成的详细报告，在管理控制台中可以直接查看，也可被发送到邮箱方便以后查阅。

管理员能配置一个警告器，或选择记录消息日志，一旦特定的事件发生，该消息被记录日志，您可以通过配置选项来自动接受 e-mail 电子邮件。当然，对于那些不触发事件的消息被日志记录时，您也可通过特定的配置选项来选择接收 e-mail。

管理员可以监控搜索活动情况，包括性能状态、覆盖率、出错率、查询速率、响应时间的状态、最多的查询及最近的查询，同时，还可以查看、输出一系列的报告，包括每天及每小时的最多查询及响应时间，每秒最多查询数等。

2.4. 支持数据源

WebSphere Information Integrator OmniFind Edition 可以无缝集成现有系统，同时，它的服务组件具备访问不同数据源、收集各种数据信息的能力，并可以创建索引，提高搜索速度。通过对信息进行语言分析以及其他类型的数据分析，OmniFind 提供了高质量、高相关性的搜索结果。因此，对于用户来讲，无需了解对不同类型数据源的搜索技术接口。

WebSphere Information Integrator OmniFind Edition 支持多种数据源类型，包括：文件系统、内容知识库、数据库、协作系统、intranets 网站、extranets 网站以及公众 Web 网站：

- 文件系统
- HTTP/HTTPS
- 新闻组 (NNTP)
- Lotus® Notes®/Domino® 数据库
- Microsoft® Exchange 共享文件夹
- IBM DB2® Content Manager 内容管理系统
- EMC Documentum and FileNet Panagon Content services，通过 WebSphere Information Integrator Content Edition 实现支持
- DB2 UDB for Linux®, UNIX® and Microsoft Windows®
- DB2 UDB for z/OS®, Informix® Dynamic Server, and Oracle databases，通过产品中打包授权限制的WebSphere Information Integrator limited use license
- Microsoft SQL Server 2000
- Hummingbird® Enterprise DM
- 支持其它数据源的增加的连接器，可以通过 Data Listener API 支持。（Additional connectors can be built to other data sources via a Data Listener API.）

2.5. 安全性

数据资源的安全性是构建一个企业级搜索引擎需要考虑的重要因素。WebSphere Information Integrator OmniFind Edition提供了多种搜索安全控制机制，它的设计可以集成现有的认证组件，因此，对最终用户来讲无需另一个独立的登录过程。当OmniFind需要登陆用户的标识时，它将会与主机环境，例如WebSphere® Portal, WebSphere Application Server或者其他的应用程序进行交互，来获得用户的身份凭证。

这种方法可以和一个用户注册库，例如 LDAP 目录服务器协同工作，同时，支持 OmniFind 和企业现有认证策略的无缝集成，而不需要另外地维护用户注册系统。

OmniFind 中的安全机制可以保护数据源，以防未授权的搜索操作，同时，严格限制特殊用户的管理功能。使用 OmniFind，用户可以搜索广泛类型的数据源。OmniFind 在几个层次上

进行了安全性服务的调整和加强，确保只有授权的用户能访问相关的内容，同时，也只有授权的用户能访问、使用管理控制台。

WebSphere Information Integrator OmniFind Edition 提供了 4 个层次的访问控制，它们可以独立或者联合使用以提供多种层次的授权：

- 管理级访问控制，确定哪些用户能设置和维护信息集合
- 信息集合级访问控制，确定哪些搜索应用程序被允许访问所有或某些特殊的集合
- 文档级访问控制，确定哪些用户有访问特殊文档的权限
- 数据加密安全性，加密敏感数据，例如密码

OmniFind 提供了安全令牌（Security Token）插件应用开发接口 API，可以方便实现 OmniFind 部署及与现有安全架构的集成。通过 plug-in 可以为每个文档定义搜索权限，确保用户无法检索到其没有得到查看授权的信息。总的目标是实现文档级别的安全保护，实现的方法是通过在数据信息被收集的时候，指定数据信息可选的相关安全令牌。通过启用搜索应用程序的安全性，您可以使用这些令牌（tokens） 来加强访问控制，同时，确保只有有合适凭证的用户才能查询数据、浏览结果。

对于集合级的访问控制主要与企业应用配合，可以控制某个部门的搜索应用能够搜索的集合。文档级访问控制，可以将用户与可访问的文档直接关联。其授权是通过设置安全性令牌（Token）实现。OmniFind 提供的机制允许在对文档进行抓取(Crawl)的同时，为每个文档设置安全令牌信息。该令牌信息可以是操作系统 ID, 用户 ID, 组 id 等，设置该安全性令牌可以有管理员指定、预定义，通过 API 由用户自定义等多种实现方式。

若需要与现有的安全体系结合，还可通过服务支持开发与特定安全体系的插件，即在 Crawler 环节加入 Security Plug-In 插件。

简而言之，OmniFind 安全模型提供了一种机制，可以在搜索时将安全标记与每个文档相关联，而在查询时将安全标记与用户查询相关联。在查询时，索引可以非常高效地进行文档过滤，所以用户只能查看其具有查看授权的那些文档。另外，OmniFind 的安全控制机制还可以与企业现有的内部安全机制集成使用。

2.6. 门户集成和定制

WebSphere Information Integrator OmniFind Edition 为搜索服务提供了一个独立的 HTML（基于 Struts 的）Web 应用，它是搜索服务器组件的一部分。OmniFind 还提供了一个样本搜

用应用程序，以及一些门户小服务程序样本（Portlets）程序，用户可以此为模板，开发、创建满足自己特殊需求的搜索应用程序。OmniFind 提供了 SIAPI API，可以很容易地集成到企业 JAVA 应用程序，开发客户自己的搜索门户。

WebSphere Information Integrator OmniFind Edition 为 WebSphere Portal 门户系统的客户提供了增强的搜索能力，可以查询更为广泛的内容，具备百万级文档的扩展能力。使用 WebSphere Portal 门户系统搜索引擎的客户可以无缝迁移到 WebSphere Information Integrator OmniFind Edition 平台上，支持对已有门户浏览分类法和种类的导入和重用，可以实现对现有基于规则的分类规则的迁移，同时，提供 WebSphere Portal 门户系统的查询中心（Search Center Portlet）相同的用户体验。

WebSphere Information Integrator 产品还会增强门户系统的功能，我们知道，门户系统是企业范围众多应用程序和信息系统的统一的访问窗口。OmniFind 企业搜索可为门户系统提供更广泛的内容访问，更灵活、高扩展能力，以及丰富的文本语义分析功能，实现在更多的信息中获得更高质量的搜索结果。此外，WebSphere Portal 门户系统的客户，能将现存的分类法和种类迁移到 WebSphere Information Integrator OmniFind Edition 系统中。

举例来说，需要快速索引新内容的客户用例，如新闻信息流或电子交易分类目录，OmniFind 可以处理增加的额外信息，通过使用“数据监听器”（Data Listener）API，可以迅速把更改的内容推进搜索系统中建立索引。

客户的搜索应用程序可以选择以不同的形式展示搜索服务器返回的结果 。客户的搜索应用程序可以是一个 portlet，servlet，或 Java 应用程序。

2.7. 搜索结果匹配排序（Ranking）

WebSphere Information Integrator OmniFind Edition 使用先进的关联算法，由 IBM 的研究中心开发，特别进行了优化，可以为客户 intranet 内容的查询提供高质量、高相关的搜索结果。为了这个目标，OmniFind 将每一个查询分解为许多变量，而且，相关性（高命中率）的确定已远不仅仅是关联性的分析。

当用户在应用程序中输入查询条件，搜索服务器进行相应的处理，并返回与查询内容和条件最匹配的、最相关（高命中率）的结果集。OmniFind 搜索服务器使用如下几种技术来保证结果集的匹配质量：

- 基于文本的评分
- 静态排序结果集

- 动态文档内容摘要
- 对相同 Web 站点汇总收缩结果集（Collapsing）

2.7.1. 基于文本的评分

OmniFind 为每一个匹配查询条件的文档的动态打分、评分。为了计算每个文档匹配查询条件的评分，OmniFind 会考虑许多因数，例如：

- 每个查询条件在整个集合中出现的频率。通常情况下，在大多数文档中都出现的查询条件，对于文档的评分影响，比那些出现在少数文档集合中的查询条件的影响要小。
- 每一个查询条件在匹配文档中出现的次数。通常情况下，查询条件出现越多的文档，所得的评分也越高。
- 查询条件在每个匹配文档中所出现的密集程度。通常情况下，查询条件在文档中出现越密集的文档比查询条件出现非常稀疏的文档评分要高。
- 查询条件在每个匹配文档中显示的前后关联性。

2.7.2. 静态匹配排序

提供静态的匹配 Ranking 能力，高级排名分析有助于确保搜索服务器仅返回高度相关的结果。当创建一个信息集合，您可以指定是否要为这个文档集合关联一个静态的匹配排序因子。关联的静态排序因子将提高相应文档在查询结果集中的重要性。对于 Web 内容，一个文档包含许多其它文档的引用链接，以及这些链接的来源，都会提高该文档在搜索结果中的相关性指数。对于那些包括日期域或日期元数据的文档，您可以用文档的日期条件来提高它在结果集中的相关性度。如，最新 NNTP 新闻组的文章，就会比较早的相关程度高。

2.7.3. 动态摘要

OmniFind 支持动态文档摘要的生成能力。动态摘要是确定结果文档中哪些短语最好地表达了用户搜索内容的技术。OmniFind 的动态摘要功能，可以尽量捕获文档中包含大量搜索条件的句子。一些句子，或者句子的某部分被选入的同时，会返回结果时，可以自动根据搜索串对文档进行动态的摘要处理，搜索条件将以高亮度方式显示在以 HTML 呈现的结果集中，并对搜索的输入请求提供自然语言识别和分析能力。

2.7.4. 对相同 Web 站点汇总收缩结果集

可为来自同一 Web 站点返回的搜索结果集指定选项来分组文档。OmniFind 能进一步组织搜索结果，因此，对来自同一 Web 站点分别得来的的搜索结果能被组合在一起。当结果集被汇总时，排序最高的将代表性地直接以左对齐方式显示出来。一个或更多排序比较低的结果集将被分组排列在其下。如一个企业搜索的应用程序样本，可以只将每个搜索 Web 网站返回的排序最高的两个结果文档显示出来。如果从同一 Web 网站返回地结果文档超过两个，您可以选择去查看其它收缩起来的结果集。

2.8. 集合，分类以及领域

查询可以针对一个特定的集合，一个集合内某个特定的分类，或集合内一个或多个领域，还可以是一个分类与一个或多个领域的组合。分类的方法可被用来将共享同一 URI I(Universal Resource Identifier)模式的文档进行分组，也可被用来将包含或排除某些特定单词或词组的文档分为一组。最终用户可以指定特定的分类做为搜索的目标，来限制其搜索结果。分类方法也可用于创建对于特定文档的“快速连接”（quick links）。OmniFind 的搜索结果集可以包含预定义的链接和快速链接。不论何时用户提交一个包含这些特定的单词和短语的查询，快速链接都将以文档形式做为查询结果集返回。

领域使您能限定用户在结果集能看到的部分。限定用户能搜索的文档范围，确保结果集中返回的内容确实是用户想要寻找的信息。例如，创建一个包括关于技术支持部门的 URI 领域和另一个包括人力资源部门的 URI 领域。

2.9. 搜索技术

当您为一个搜索集合配置搜索服务器时，可以为搜索集合指定其搜索选项，同时，配置一个搜索缓存区来存储最频繁被请求的搜索结果集。当搜索服务器处理这些请求时，最先查找缓存区中是否有需要的结果集。如果搜索服务器找到适合的结果文档，就会为客户迅速返回结果集。

WebSphere Information Integrator OmniFind 还提供了对查询条件进行拼写检查的选项。如果用户在查询条件中出现拼写错误，搜索服务器将对如何正确拼写查询条件，提供一些提示信息。例如，如果您指定“saerch’”作为查询条件，会看到为您指定“search”作为一个可能纠正选项的情况。

WebSphere Information Integrator OmniFind 支持多种查询技术，例如，基于任意文本形式的查询，搜索特定的短语，排除特定的单词或者指定更复杂的查询，来增进查询结果的准确度。

2.10. 可用性和可扩展性

WebSphere Information Integrator OmniFind Edition 的设计，提供了优越的性能，可扩展性以及高质量的结果集，可以快速访问企业中广泛的、典型信息源。为了保证这些特性的实现，它提供了：从原始资源提取文档，进行后台解析，分析相关内容，创建集合（索引），实现优化查询速度和结果准确性。

OmniFind 还在全球最具挑战的 intranets 上证明了它的强大功能和可扩展性 - IBM 的 intranet 有超过 300,000 人员使用搜索服务。 在上面实现了两个截然不同的应用：

- 巨大规模的日常 intranet 搜索。超过 10,000 Web 站点以及 2 千 5 百万 URLs。超过 9 百万独立的文档被索引，同时，每 4 小时完成索引的升级。这个应用从 2003 年 9 月开始，一直持续一天 24 小时一周 7 天不间断的工作。
- 一个员工蓝页的应用，帮助实现相关专家的定位，基于 IBM 公司的规模，搜索覆盖超过 50 万条 XML 记录，同时，每天每 2-3 小时内补充 2 万条更新记录。该应用已于 2004 年 3 月完成并开始使用。

3. OmniFind 应用案例

3.1. 华为 E-Support 项目信息搜索

项目背景：

华为 e-support 系统是面向华为全球技术员工提供的网上技术支持系统。为员工提供各类华为产品的技术支持服务（如：如何解决某类交换机的故障),是电话支持系统的有力补充，通过企业内部有效的知识和经验共享，提高效率降低电话支持的压力。

华为的产品很多，产品相关设计文档、技术支持手册、故障处理记录等产品相关资料数据量大，而且分布在多个文件系统和内容管理系统（DB2 Content Management）中，信息搜索能力是实现 E-support 系统的关键。

E-support 系统所涉及的文档资料数据组织不统一，文档资料本身与描述它的元数据往往分布在不同的系统（如文件系统存储资料本身，而该文件的分类信息、安全信息则存储在其他数据库中），为文档检索增加了很大的难度。

华为的各类资料都要求有很严格的安全控制，不允许员工搜索有安全级别而未经授权的文档。

文档的格式多样，包含 Word、Excel、PPT、PDF 和文本文件等多种格式。

OmniFind 解决方案

经过慎重比较，华为选择 IBM WII OmniFind 作为 E-support 系统的搜索引擎。

通过 OmniFind 的元数据扩展插件，将华为 CM 系统中的文档与 Oracle 中的属性信息在搜索时合并，在搜索时提供完整文档元数据。

通过 Ominfind 的安全插件，将华为对文档搜索的安全要求与 OmniFind 集成，满足 E-support 信息搜索的安全性。

通过 IBM WII 产品屏蔽华为 E-support 系统相关各类数据库的异构性。

通过 IBM Lab-based services,为华为扩充其它产品功能。

解决方案主要特点

数据来源广泛，能够准确搜索华为 E-support 系统的各类数据资源。包括 DB2 CM 系统、Oracle 数据库、文件服务器等。

文档安全性要求高，支持文档级的搜索安全性，能够与华为现有的安全管理体系有效整合。

提供文档元数据的扩展能力，能够有效解决文档和属性分别存储对搜索的挑战。

支持包括中英文在内的二十多种语言的准确分词能力。

很强的系统可扩展性，能够支持海量数据及大并发的用户访问。

提供丰富的 API 应用开发接口，易于二次开发及与现有应用的集成。

3.2. IBM 企业内网信息搜索

OmniFind 的信息搜索具有很好的扩展性，在架构上可支持多个爬行器、多个 parser 和多个 Indexer，以及多个搜索引擎。不仅在性能上可很好的扩展，在高可靠性方面也有很高的价值。在过去两年的时间里该技术在 IBM 内部为 30 多万 IBM 员工提供信息搜索服务。其数据资源包括内部主要的共享数据库、Intranet 网站及其他相关数据源。平均每日支持 8000 多个搜索请求。提供多语言支持，2700 多个资源分类。中心式的部署架构。