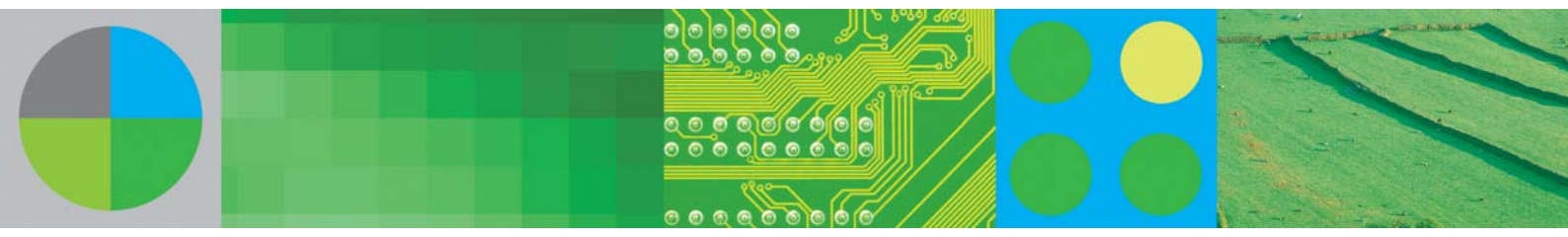
DB2 Information Management Software

信息搜索与 WebSphere Information Integrator OmniFind Edition

ON 随需应变的业务

目录 Content



信息搜索需求概述

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。这些信息部分可能存储在数据库中，大量的则保存在文件服务器、邮件系统、网站的网页、内容管理服务器、流程引擎等中。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。

以互联网为例，对于浩如烟海的咨询，其中绝大我们所需的信息，我们并不知道其所在的位置，今天我们所采用的最常用也是最有效的手段是使用Google、Yahoo一类的搜索引擎，动态搜索相关信息。在企业内部也面临同样的情况，大量企业的信息资源分散在各处，以不同的格式存在、按不同的分类组织，受不同的安全机制控制。而最终使用其内容的用户不可能去掌握这些复杂性，希望能够通过输入简单的关键词的组合，由系统自动从各类信息资源中搜索到相应的内容。从而提供对信息访问的最简单、最直接的途径。应而企业信息搜索技术应运而生。

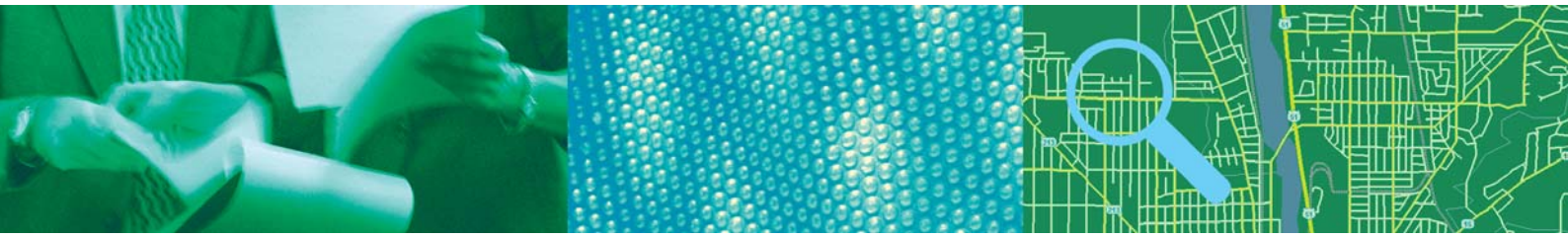
但是相对于互联网的简单信息搜索，在企业内部实现信息搜索有更大的复杂性。主要体现在信息分布的多样性、信息访问控制的安全性、及与业务处理的集成性等多个方面。与互联网不同，在企业内部信息不仅分布在网站上，大量有价值的信息是存储在文件系统、内容资料库、数据库及邮件系统中。格式可能是文本、XML、Word 文档、PDF 及 PPT 文件等。这些信息可能有不同的安全访问级别、对不同的用户需控制其访问的信息内容，往往都要求做到文档级的安全性管理。另外企业内部信息搜索应用的目的性更强，往往还要求搜索的结果能够与企业现有的业务处理进行紧密地关联，使搜索能够为更灵活的业务处理流程服务，如减少寻找客户资料的时间、提供客户网上自助服务的快捷查询手段等。当然对自然语言的处理能力是两者都必须具有的功能。从而在传统的基于流程的信息服务之外，信息搜索正逐渐成为一种更为普及更为通用的信息访问接入手段。

企业搜索不同于单纯的数据库查询，查询适合于结构化数据，而搜索则更适合于非结构化技术。企业搜索是用于从文件系统、内容存储库、数据库、协作系统、应用程序和公司内部网中存储的大量企业信息中查找最相关的信息。企业搜索必须整合文本搜索和传统的数据库查询技术使企业具备从数据库记录等结构性数据和文件系统等非结构性数据中获得搜索结果的能力。使用搜索技术，企业无需局限于预先定义的查询方式。

基于企业内容的搜索和基于 Internet 的内容搜索之间存在着很大的区别。企业搜索中，不同的内容源需要不同的技术来确定其文档相关性，同时必须使用不同的安全和访问模型，而且还要满足高质量搜索的不同用户需求。但是，即使是有些最成功的 Web 搜索技术 (如网页分级) 还没有实现针对企业环境的优化，对企业环境中的文档还没有像 Internet 上的文档那样相互链接起来。这就是为什么在企业中找到正确相关的信息是如此的费时费力。

最新推出的IBM满足企业级信息搜索需求的主要产品是 WebSphere Information Integrator (WebSphere II) OmniFind Edition。 它属于Websphere II家族产品的一个重要部分，以下简称Omnifind。

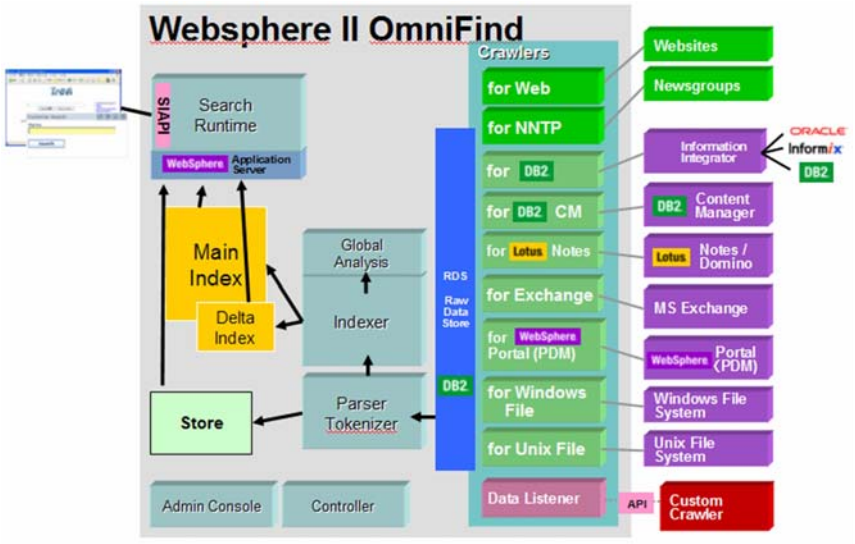
作为IBM总体信息整合平台的一部分实现了企业搜索的功能。OmniFind 具有查询不同类型的数据源和立即返回结果的能力，这将有助于企业更好地洞察它们的运营情况，并更好地利用企业现有的数据资源，快速准确地定位企业中最佳的相关内容。



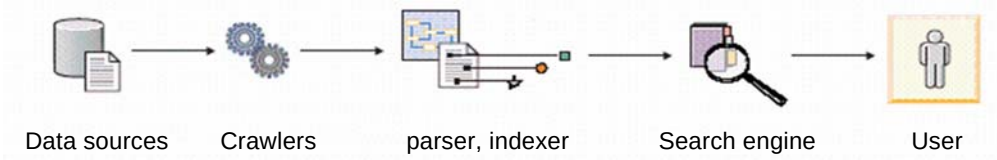
IBM 信息搜索方案 —— OmniFind

体系架构

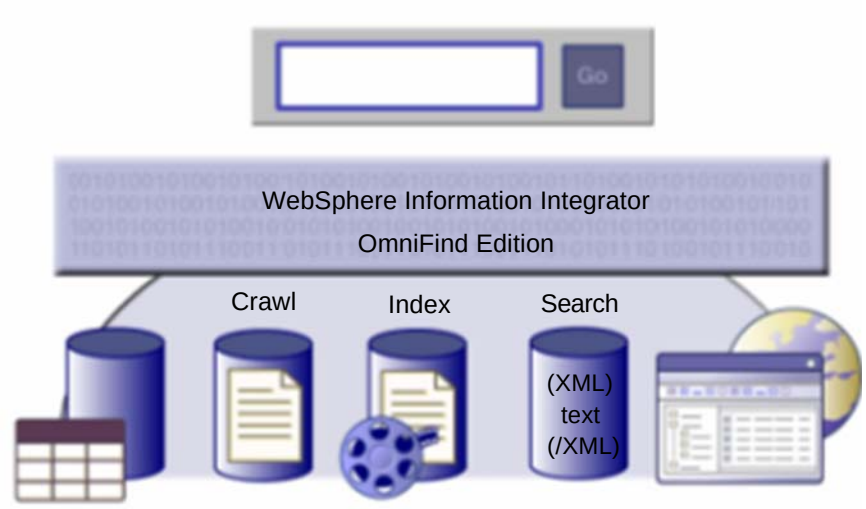
Omnifind 提供了一个企业级搜索中间件的体系架构；该架构提供多种爬行器(crawler)能够快速访问企业的各类业务信息，爬行器返回的信息通过分词处理实现按自然语言的分词，然后对分词后的信息建立专用索引，基于此索引在前端提供强大的搜索引擎，实现对各类信息的高质量快速搜索，并提供相应的 API 与用户的各类应用集成。通过此架构满足企业级的信息搜索需求。其架构如下图所示:



OmniFind 设计的目标是无论数据在那里、以何种形式存在，能够对其快速地访问，通过准确分词建立索引，提供完整的管理和安全控制机制，实现次秒级高质量的搜索查询。



对应于上述的搜索处理过程，OmniFind的体系架构中设计了三个主要组件: Crawler、索引服务器和搜索服务器，如图所示:



爬行器 (Crawler) :

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。

OmniFind可进行搜索的企业信息源可以是: HTTP、HTTPS、新闻组 (NNTP) 、文件系统、IBM Lotus? Domino? 数据库、Microsoft? Exchange 公共文件夹、IBM DB2 Content Manager、IBM DB2 Universal Database (UDB)、UNIX 和 Microsoft Windows 文件系统、IBM DB2 UDB for z/OS、IBM Informix Dynamic Server、Oracle 数据库、Documentum 和 FileNet 存储库等。

分词及索引服务器(Parser, Indexer):

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，对数据进行分词处理(praser)，支持对简体中文及繁体中文、英文、日文等20多种

语言的准确分词能力，搜索还支持30种其他语言。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。索引服务器在设计上可扩展到两千万文档。

搜索服务器:

负责处理搜索请求，基于索引信息，OmniFind将提供一个或多个搜索引擎，在索引中查找最相关文档并以次秒级响应时间返回结果，实现对所有资源的全文检索，最后将检索的结果返回用户。

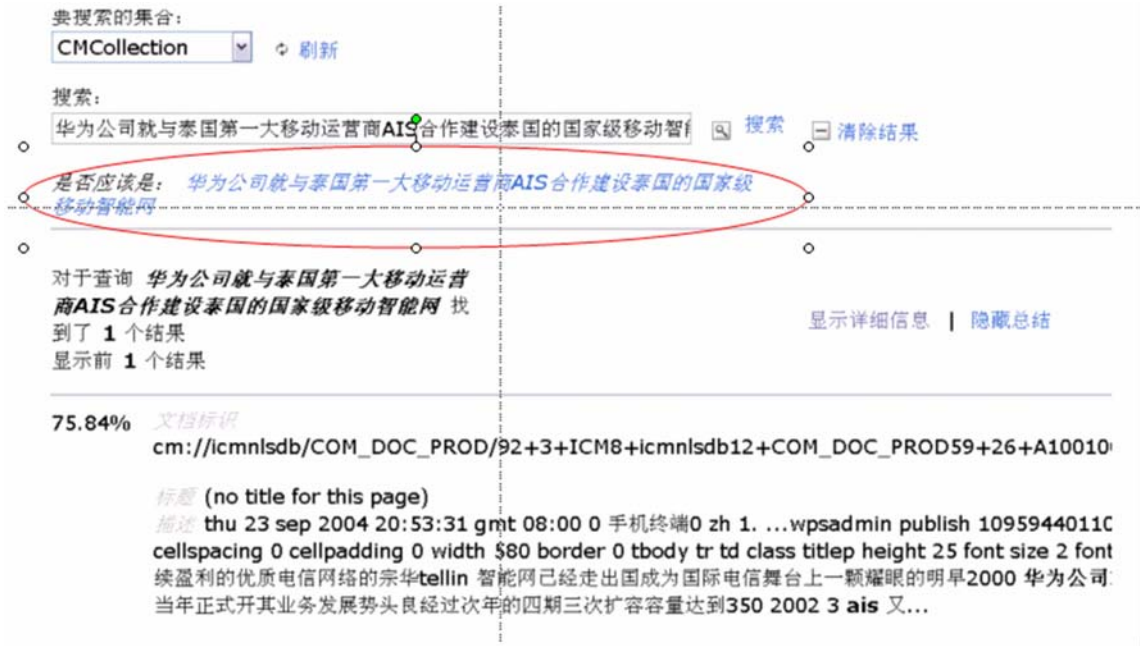
OmniFind提供动态和静态的匹配Ranking能力，高级排名分析有助于确保搜索服务器仅返回高度相关的结果。

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、。

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.

OmniFind支持多个搜索服务器并行，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 可以以亚秒级响应时间交付高度相关的搜索结果。

OmniFind支持动态文档摘要的生成能力。OmniFind在返回结果时，可以自动根据搜索串对文档进行动态的摘要处理。并对匹配的字进行高亮显示。对搜索的输入请求提供自然语言识别和分析能力。其搜索结果样例如下:



目前 OmniFind 所支持的平台包括:

Red Hat Enterprise Linux 3

Microsoft Windows 2000

IBM AIX5L Version 5.2

SUSE LINUX Enterprise Server 8

安全性控制

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。

OmniFind 提供了多种搜索安全控制机制，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。OmniFind 可以提供支持文档或集合级别安全性的基础设施。

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。

简而言之，OmniFind 安全模型提供了一种机制，可以在搜索时间将安全标记与每个文档相关联，而在查询时间将安全标记与用户查询相关联。在查询时间，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。

系统管理

OmniFind 非常便于安装和管理，所以使用很短的时间即可建立和运行企业搜索应用程序。管理员仅需指定搜索从何处开始到何处结束和刷新索引的频度。OmniFind 设计用于减少 IT 人员的管理需求，其分析特性是透明的，可以最大限度地减少完成高质量搜索结果所需的管理任务。使用 OmniFind，可以方便地定义合适的安全性、监控系统活动并解决发生的任何问题。

OmniFind 非常适用于企业 Java 应用程序。此外，它还包括一个示例搜索应用程序，用户可以将该应用程序用于创建满足组织独特需求的搜索应用程序的模板。

应用特性

OmniFind 支持对搜索要求的多国语言自动分词能力。OmniFind 不仅提供自动分词能力，而且基于分词功能提供选项支持对用户输入的查询请求进行拼写较正，基于校验后的结果进行搜索。

OmniFind 支持动态文档摘要的生成能力。OmniFind 在返回结果时，可以自动根据搜索串对文档进行动态的摘要处理。

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，OmniFind 提供专门的搜索语言，支持全文检索基础上对属性信息的过滤和与或等多种组合匹配方式，通过客户化编程，可实现独立存储的元数据属性信息和全文信息结合的搜索。应用可灵活定义搜索需求。搜索结果

提供动态摘要及匹配度信息，对命中的词汇进行高亮显示。

OmniFind 提供了对结果的排序能力。其排序方式支持两种。一种为 Text based scoring，动态计算匹配度评分。另外还支持 Static Ranking，能够根据文档本身的因素 (如: 文档被引用的计数值或文档的时间戳) 对范围结果的排序产生影响。

OmniFind 还提供了二次检索能力。OmniFind 所返回的结果中包含搜索的查询串信息。二次查询实现时，是将在第一次搜索的查询串基础上添加新的查询要求。OmniFind 二次查询时，将首先从 cache 中提取信息，这种方式无疑将大幅缩短查询时间，提高查询效率。

与 Portal 整合

OmniFind 向 IBM WebSphere Portal 软件客户机提供了增强的大型内容搜索功能，可以扩展到数百万文档和更丰富的文本分析框架。IBM WebSphere Portal Search 客户可以向 OmniFind 无缝过渡，它导入和重用了导航和分类的现有门户分类法，为基于规则的类别迁移规则，并提供与 WebSphere Portal 软件的 Search Center portlet 相同的用户体验。除此之外，管理员还可以使用熟悉的控制台可以快速建立和运行程序。因此，用户可以方便地将 OmniFind 集成到自己的业务中，不管是将其用于专门应用程序还是用于公司内部网或外部网。

应用案例

OmniFind 的数据检索技术具有很好的扩展性，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。在过去一年多的时间里该技术在 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。数据资源量约 2000 万份文档，每日 8000 多个检索请求。提供多语言支持，2700 多个资源分类。中心式的部署架构。

国际商业机器中国有限公司

北京总公司
北京朝阳区工体北路甲二号
盈科中心 IBM 大厦 25 层
邮政编码: 100027
电话: (010)65391188
传真: (010)65391688

上海分公司
上海市淮海中路 333 号
瑞安广场 10 楼
邮政编码: 200021
电话: (021)63262288
传真: (021)63261177

广州分公司
广州市天河北路 183 号
大都会广场 18-20 层
邮政编码: 510620
电话: (020)87553828
传真: (020)87550182

沈阳分公司
沈阳市沈河区青年大街 219 号
华新国际大厦 19 层
邮政编码: 110015
电话: (024)23962288
传真: (024)23961040

武汉分公司
武汉市汉口建设大道 700 号
武汉香格里拉大饭店 302 室
邮政编码: 430015
电话: (027)85805588
传真: (027)85800088

深圳分公司
深圳市深南中路 333 号
信兴广场地王商业大厦
38 层 3805, 3806
邮政编码: 518008
电话: (0755)82462193
传真: (0755)82462186

南京分公司
南京市新街口街金陵饭店
世界贸易中心 16 楼
邮政编码: 210005
电话: (025)84716677
传真: (025)84729054

成都分公司
成都市人民南路 2 段 18 号
川信大厦 27 层
邮政编码: 610016
电话: (028)86199888
传真: (028)86199500

西安分公司
西安市高新区科技路 48 号
创业广场 B 座 1202 室
邮政编码: 710075
电话: (029)88316868
传真: (029)88323777

杭州分公司
杭州市杭大路 15 号
嘉华国际商务中心 1506 室
邮政编码: 310007
电话: (0571)28896988
传真: (0571)28891128

昆明办事处
昆明市洪化桥 20 号
海逸酒店 512, 513 室
邮政编码: 650031
电话: (0871)5388555
传真: (0871)5380199

福州办事处
福州市五四路 73 号
福建外贸中心酒店 9925 室
邮政编码: 350001
电话: (0591)87523388-9925/9938
(0591)87600122
传真: (0591)87541814

重庆办事处
重庆市渝中区邹容路 68 号
大都会商厦 21 楼 2105 房
邮政编码: 400010
电话: (023)63830503
传真: (023)63830513

长沙办事处
长沙市解放东路 380 号
华天大酒店贵宾楼 1008 室
邮政编码: 410001
电话: (0731)4169188
传真: (0731)4116845

乌鲁木齐办事处
乌鲁木齐市东风路 1 号
海德酒店 17 楼 B 座
邮政编码: 830002
电话: (0991)2338911
传真: (0991)2831805

哈尔滨办事处
哈尔滨市道里区友谊路 555 号
哈尔滨香格里拉大饭店 4 层
邮政编码: 150018
电话: (0451)87606688
传真: (0451)84899988