



摩卡搜索引擎

Mocha Search

产品白皮书

公 司：摩卡软件有限公司(Mocha Software Co., Ltd.)

地 址：北京市西城区宣武门西大街 127 号大成大厦 15 层

联系我们：400-611-5522

Email: Marketing@mochasoft.com.cn

目 录

1	引言	1
2	理论基础	2
3	产品优势 (Product)	3
3.1	解决顾客的痛	3
3.2	做得更多 (DO MORE WITH LESS)	3
3.3	产品通用性	3
3.4	开放式架构与标准接口	3
4	产品亮点	4
4.1	部署 (DEPLOY)	4
4.2	搜索 (SEARCH)	4
4.3	安全 (SECURITY)	5
4.4	管理 (MANAGE)	6
5	Mocha Search产品描述	7
5.1	产品定位	7
5.2	整体架构	7
6	成功案例	8
6.1	案例背景介绍	8
6.2	现状总结与方案实施	8
6.3	客户评价	8
7	系统运行环境	9
7.1	服务器	9
7.2	客户端	9
8	致谢	10
9	联系我们	10

1 引言

在当今飞速发展的信息社会里，企业信息化建设经历了一个建设、发展、丰富的历程。在经过长时间的使用后，企业内部系统（特别是办公管理系统）产生了大量的由不同部门、不同个人处理过的文档。原有的基于目录结构的检索、查询变得繁琐、费时，不能帮助最终用户进行快速、有效的文档定位。

随着办公自动化系统在国内的不断推广，根据行业性质的不同，业界诞生了很多基于不同行业应用的 OA 系统。这些 OA 厂家以 IBM Lotus Domino 作为支撑平台，开发了很多与 OA 相关的功能模块。Lotus Domino 平台是一个很好的开发平台，它为开发人员提供了很多可配置的功能，如权限管理，人员组织，开发人员节省了大量的开发时间，实现了办公自动化系统的快速安装与部署。但是，Lotus Domino 本身有些局限性，随着用户数据的不断增加（当单个数据库超过 2GB 时），在 Lotus Domino 平台中，搜索变成了一件很困难的事情。由于很多 OA 文档都储存在 Lotus NSF 库而不是关系型数据库，搜索遇到了瓶颈—搜索速度变成很慢，或在某些情况地下，造成宕机。**Mocha Search 是少数深入了解 OA 应用基础上的搜索引擎。**

当前状况下，很多企业也开始重视自己的 IT 建设，也通过国际知名引擎 Google 的成功，了解搜索引擎对企业信息建设的重要性。但是，经过性能价格对比，您会发现，一方面是很多国际品牌的搜索引擎价格昂贵，另一方面国内搜索引擎虽然价格合理但是有性能问题。

这些都不符合多数企业的需求。Mocha Search 本着“做得更多” (Do More With Less) 的理念，从用户的需求角度出发，更好地解决了这些问题。

摩卡搜索 (Mocha Search) 是集培训、咨询、产品定制、服务于一体的，浓缩的，系统化的基于文档搜索的产品，完全基于 J2EE 技术研发。Mocha Search 的应用可以为企业信息化建设提供基于关键字的强有力的查询、搜索的功能，并完全结合数据源系统的权限，保障系统文档的查询 100% 等同于数据源系统的权限控制。

2 理论基础

任何一个系统产生的所有文档都可以看做一个基本的“文件空间”，每一个文档都是由多个词组组成。下图是一个文件空间的示意图：

	W1	W2	W3	...	Wm
D1	√	√			√
D2		√	√		√
D3	√		√		√
...					
Dn		√			√

文件空间示意图

可以看到，文件空间中包括总共 m 个词组，即关键字，文件空间中包括总共 n 个文件。每一个文件都是由一个或多个词组组成。例如，文件 D2 是由词组 W2，W3，和 Wm 组成的。

传统的根据文件目录结构的检索技术是面向文件的。假设用户搜索的条件为 $Q = \{W_x, W_y, W_z\}$ ，根据文件目录结构的检索技术即遍历所有文件 D1—Dn，并通过比对得到包含 Q 的相关文件做为检索的结果。检索结果的列表次序是根据遍历文件的次序决定的。

基于关键字的检索是面向关键字的，它使用的是“反向索引”的技术。首先会根据所有的关键字 W1—Wm 建立反向索引，即以关键字为基础建立对应每一个关键字的文件列表。用户发出搜索条件 Q，搜索的过程即为根据 Q 中的关键字、通过查找反向索引中对应的文件得到检索结果。

搜索的同时，通过计算搜索条件中关键字在相关文件中出现的次数、以及 Q 中关键字的在所有文件中出现的次数可以给出结果集中每个文件跟搜索条件的相关度的值。检索结果的列表次序为根据相关度由高到低排列。

基于关键字搜索技术的优点是在文件数据量巨大的情况下，搜索的性能相比传统检索技术大大提高。同时，此技术提供根据相关度排序的功能，方便用户迅速定位查找的文件。

Mocha Search 使用的就是面向关键字搜索的技术。

3 产品优势

3.1 解决顾客的痛

Mocha Search 很有效地解决 OA (Office Automation) 在 Lotus Domino 平台检索的问题。Mocha Search 可以避开繁忙的工作时间, 在用户不使用系统时建立索引。当 Mocha Search 做搜索的时候, 是通过 Mocha Search 的索引找到相关的文档, 避免直接从海量的业务系统 (OA) 数据库里寻找需要的关键字。

3.2 做得更多 (Do More With Less)

- Mocha Search 是低成本、高效率的搜索引擎。它不需要其他国际知名品牌的搜索引擎的成本, 但是却提供相同的性能, 为我们的顾客提供了更高的性价比。
- 在性能方面, Mocha Search 提供了以下数据:
 - 提供一个 100: 8 的数据对索引比例。那就是如果用户有 100G 的文档量, 索引只需要 8GB 的磁盘空间。
 - 无论数据量大小, Mocha Search 提供了等于 1 秒的简单检索反应时间。
- 提供增量式索引, 大大降低了索引的时间。

3.3 产品通用性

- 支持多语言
支持多语言的搜索、查询, 比如中文, 英文等语言。
- 支持多平台
支持 Domino、数据库应用等多种应用平台。
- 支持多文档格式
纯文本、Domino 文档、HTML、Office 文件 (如 MS Word、MS Excel 等)。

3.4 开放式架构与标准接口

- 开放式技术—J2EE
Mocha Search 是完全基于 J2EE 架构开发, 可以在任何符合 J2EE 标准的平台上运行。
- 开放式技术—HTTP 协议
通过 HTTP 协议传输 XML 格式数据。
使用标准 HTTP 协议, 通用性强。
- 开放式技术—XML
通过 XML 格式传输数据, 便于不同系统间接口定义。核心搜索引擎模块提供 XML 格式的 API 接口, 方便不同应用系统的二次开发和无缝的应用集成。

4 产品亮点

Mocha Search 的功能亮点，主要体现在以下几个方面：

- 部署 (Deploy)
- 搜索 (Search)
- 安全 (Security)
- 管理 (Manage)

4.1 部署 (Deploy)

在部署方面，Mocha Search 通过简单易用的 Wizard 式的安装界面，一步一步引导用户。与安装任何 Windows 软件一样容易，不需要任何的培训。

4.2 搜索 (Search)

4.2.1 提供 Google 式的用户感受

Mocha Search 为顾客提供了一种既熟悉又方便易用的 Google 式的用户感受。

■ 搜索界面



■ 高级搜索界面



■ 搜索结果界面



4.2.2 支持不同应用搜索

Mocha Search 可同时支持不同应用的搜索，主要包括：

- **Lotus Domino**
- **提供 API 搜索数据库，支持的数据库有：**
 - MS SQL
 - IBM DB2
 - Oracle
- **网站**
 - 支持 HTML 和 JSP 格式文件。
 - 支持附件搜索，Mocha Search 实现了 MS Word、Excel、PowerPoint、RTF 文件和 PDF 文件等格式搜索。

Mocha Search 的重要特点就是完全支持基于 Domino 平台开发的 B/S 架构的应用系统。Mocha Search 为 Lotus Domino 提供了特制的 Domino 适配器。通过搜索引擎的 API，它将 Domino 应用中的文档数据，配以文件的访问权限，同搜索引擎进行交互，由产品权限管理模块根据 Domino 适配器提供的权限信息，可以保证查询结果集 100% 符合原应用的权限控制。

Mocha Search 通过数据源定义模块，可以方便用户灵活定制在 Domino 应用中需要建立索引的域。一方面对于用户搜索体验更有实际意义，另外一方面减少索引过程对于系统资源的消耗、提高系统性能。

Mocha Search 能够下载网站的内容，快速搜索企业网站内容。

4.2.3 支持索引检索并行（即边建边搜）

- 在索引库用于检索的同时，可以追加索引记录，即增量索引。
- 在只采用一份索引库的前提下，实现了索引和检索的并行，因此对索引库的操作维护更加稳定方便。

4.3 安全 (Security)

4.3.1 继承了数据源文档权限

目前不同厂商提供的搜索引擎都是基于关键字搜索的技术。不过，在企业用户内部系统的使用中，不同文档的访问权限是根据用户权限控制信息来决定的。这就要求搜索引擎在建立索引、搜索时，必须按照数据源系统定义的权限信息严格控制搜索结果，必须防止最终用户通过搜索查找到其无权限访问的文档。

Mocha Search 通过权限控制管理控制组件实现根据数据源系统文件权限的访问控制。其核心组件搜索引擎提供带有权限信息的 API。部署在应用系统的适配器在建立索引的同时，提供应用系统中每一个文件的权限访问信息，并记录在索引中。

这样，在搜索时，搜索引擎就能够根据发出搜索的用户权限信息控制返回搜索结果集的文件列表，来实现权限的控制。

4.4 管理 (Manage)

4.4.1 简单易用的 B/S 管理方式

■ 管理配置界面

Mocha Search 基本上不需要任何维护工作。一般用户只在数据源上改动。通过简单易用的浏览器管理, 管理员能增、删、改系统配置。

■ Mocha Search 管理配置界面



■ Domino Adapter 管理配置界面



■ 管理配置界面数据源定义界面



5 Mocha Search 产品描述

5.1 产品定位

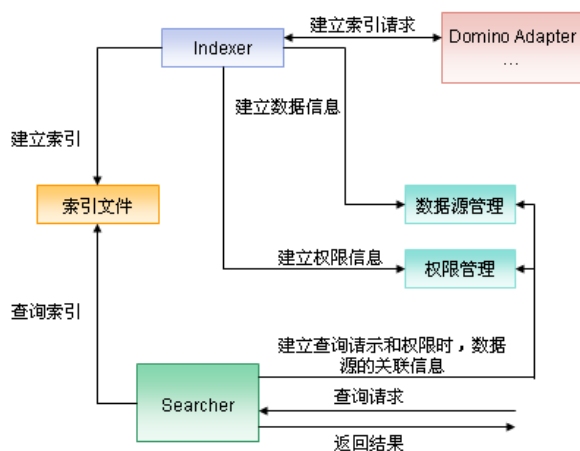
作为一家为用户实现企业信息现代化为目标的、具有影响力的 IT 公司, 摩卡软件有限公司所提供的解决方案以及产品是全方位的。采用最先进、最合适的技术, 提供满足需求的解决方案, 实现用户投资收益最大化, 是摩卡软件有限公司努力的源泉。

摩卡软件有限公司解决方案的整体架构如下图所示, 其中, Mocha Search 产品是摩卡软件有限公司总体解决方案中应用层的核心应用的组件之一。



Mocha Search 产品处于摩卡服务平台 (Mocha Service Platform) 的位置图

5.2 整体架构



Mocha Search 整体架构图

5.2.1 设计要点

Mocha Search 设计主要分为三层:

- **核心搜索引擎:** 提供基于关键字搜索技术的搜索引擎模块, 包括权限控制模块以及数据源定义模块。
- **访问层:** 提供对于外部访问的外围接口。
- **适配器:** 基于各种应用的适配器, Mocha Search 包括支持 Domino 应用等的适配器。

5.2.2 设计优势

Mocha Search 产品逻辑架构设计中考虑的关键因素:

- **关键字搜索**
关键字搜索是 Mocha Search 的基本功能。
- **权限控制**
基于数据源应用系统的文件权限控制。搜索结果集完全符合应用的权限, 用于保证信息的安全。
- **接入可扩展**
提供基于 XML 标准的接口 API, 对于需要建立搜索功能的新应用系统, 根据 API 进行二次开发、集成即可。
- **开放性、平台无关性**
J2EE 标准, 能集成任何第三方产品。
- **模块化、灵活的配置**
成功产品化的体现。
- **分布式**
B/S 结构。
- **性能可扩展**
可以通过配置来增加服务器。

6 成功案例

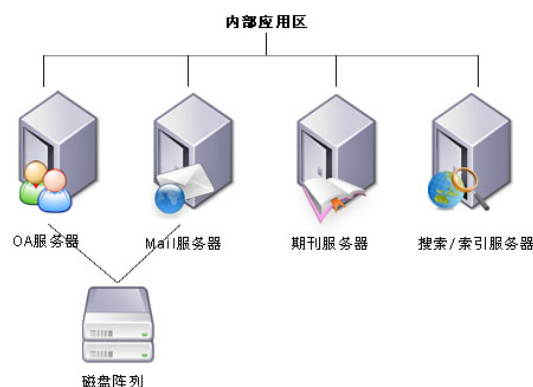
6.1 案例背景介绍

中国移动通信集团公司是中国最大的移动通信运营商，目前内部管理系统的注册用户约 1000 左右，过去的 5 年里，企业信息化建设取得了长足进步，以 Mocha 工作平台为基础，形成了以办公自动化系统为核心的整体规划，但是，随着企业信息化的逐步深入发展，企业的知识积累达到了一定水平，也逐渐产生了新的现实问题：

- 电子文档的数量每年以几何级数递增，目前，EMIS 系统（企业管理信息系统）本身的全文检索功能已经不能满足用户的需求。
- 目前的检索功能受到平台（Domino）以及数据量的限制，检索速度相当慢。
- 目前的检索只能根据时间排序，不能提供根据搜索条件相关度的检索。
- 没有一个能够结合 EMIS 系统权限的搜索引擎。

6.2 现状总结与方案实施

中国移动需要一个搜索工具，实现已有系统文档的快速、提供权限控制的搜索功能。因此，采用了以 Mocha Search 4.0 为核心的搜索引擎工具：



中国移动 Mocha Search 网络拓扑图

6.3 客户评价

“很好解决了原来的检索问题，尤其是性能方面的问题”。

“简单易用，便于查询和系统管理”。

7 系统运行环境

7.1 服务器

- 服务：PC 服务器
- CPU：主频最低 1.2GHz 以上，建议主频 2.2GHz
- 内存：最低 512MB
- 磁盘空间：40GB 以上
- 操作系统支持：Windows 平台

7.2 客户端

- CPU：PIII 或以上
- 内存：128MB 或以上
- IE 5.5 SP2 或更高版本
- 操作系统：Windows 2000 或以上

8 致谢

搜索引擎从技术研究，到成为一个功能完善的、设计精良的产品，摩卡软件有限公司走过了一段艰辛而充满挑战的里程，多年的行业积累使我们深感：产品中有来自于广大用户的思想和智慧，也凝聚了来自摩卡软件有限公司的可敬的广大工程师和技术骨干的辛劳。

随着技术的不断发展，以及 Mocha Search 产品在企业中的不断实施，摩卡软件有限公司将继续对产品的后续提升提供强有力的支持，我们将履行我们的承诺，继续提供“高质量的产品，优质的服务”！

9 联系我们

摩卡软件有限公司

地址：北京西城区宣武门西大街 127 号大成大厦 15 层

联系电话：400-611-5522

传真：(8622) 87341661

网址：<http://www.mochasoft.com.cn>

电子邮件：Marketing@mochasoft.com.cn