

EKG 内网文档搜索引擎系统

Enterprise Knowledge Gateway (EKG) System

系统介绍

System Introduction



灵玖中科软件（北京）有限公司

LING-JOIN Zhongke Software (Beijing) Co. Ltd.

Document Information

Document ID	LINGJOIN-ELINT-WHITEPAPER	Version	V1.0
Security level	Confidential 秘密★	Status	Creation and first draft for comment
Author	灵玖中科软件（北京）有限公司	Date	2010/9/21
Publisher	/	Approved by	灵玖软件

Version History

Note: The first version is "v0.1". Each subsequent version will add 0.1 to the exiting version. The version number should be updated only when there are significant changes, for example, changes made to reflect reviews. The first figure in the version 1.x denotes current review status by 1. x denotes review process has passed round 1 etc .Anyone who create, review or modify the document should describe his action.

Version	Author/Reviewer	Date	Description
V0.1	王毅伟	2010/9/15	Creation and first draft for comment.
V1.0	王毅伟	2010/9/21	First complete draft for comment.

目 录

一、内网文档搜索引擎系统简介	3
二、内网文档搜索引擎系统主要功能	4
三、体系架构(SYSTEM ARCHITECTURE)	5
四、内网文档搜索引擎系统的特色	6
五、内网文档搜索引擎系统主要技术指标	6

一、内网文档搜索引擎系统简介

内网文档搜索引擎系统是一款针对政府、企业或机构局域网内部文档管理、搜索与挖掘的软件产品，凝聚了灵玖中科软件（北京）有限公司多年自然语言理解与精准搜索引擎的技术积累。

随着信息化的高速发展，社会各界、政府及企业中的电子文档（word、excel、ppt、txt、pdf 等）也变得越来越。大量政府和企事业单位的文档是包含着客户关系、产品信息、市场情报、策划思想等软资产的载体。文档作为一种信息资源，作为企业生产、技术、科研和经营等活动的真实记录和一项基础性工作，同时作为与企业同步发展的无形资产，在企业管理等方面积极地发挥着重要作用。

政府/企事业单位的电子文档日益增多，文档格式的多样化、文档内容分散存储于各个不同的电脑上等问题，文档的管理存在三大隐患：

1、文档总量大，单篇文档的篇幅较长，检索与管理的效率极其低下

据统计，一台工作电脑上，大小、版本不同的电子文档一般都在数千到数万个以上，这样庞大的文档量，即使浏览一遍标题也需要花费大量的时间，靠人工有效管理几乎不可能。据统计，用资源管理器等常规手段查找文档，每篇文档平均需要 3 分钟。按用户每天需要查找 10 篇计算，一个人每年需花费 100 小时以上。而对于拥有成百上千电脑的单位，对这些文档的管理已经成为“食之无味弃之可惜”的鸡肋。

而内网文档搜索引擎在上前万文档中查找文件的速度是 0.1 秒钟以内。

2、文档分散存储，共享困难，协同工作往往靠上级或者个人协调

目前，除了少量文档通过共享目录、邮件、网络即时通讯软件被少部分员工共享外，大多数电子文档属于某个员工的“私产”，传统共享方式的效率低下使员工共享同事的文档比上互联网查找文档更为困难。

而通过内网文档搜索引擎，大家可以比 Google, Baidu 更快更便利。

3、分散存储，缺乏必要的安全保障措施，安全隐患极大

安全隐患主要表现在：

1) 大量文档分散成各台电脑上，一旦有员工离职，相关的大量工作文档被直接带走丢失，或沉落于电脑中，成为“不为人知”的秘密，由此带来大量文档资产的流失。

2) 由于大量文档处于不被管理的无序状态，文档被病毒感染、意外丢失、越权使用等情况屡屡发生，使文档的安全性无法得到保障。

3) 由于工作文档经常会被不断地修订，有些文档多达数十，乃至上百个不同的版本，大量的版本进一步增加了文档管理的难度。为此，很多用户会用新版本覆盖老版本，但是，这种情况下，一旦用户需要恢复老版本，已经很难找回来。

而通过内网文档搜索引擎，大家可以避免以上隐患。

二、内网文档搜索引擎系统主要功能

内网文档搜索引擎系统包含四大功能：

1 自动备份：

每天定时自动采集局域网内部共享的电子文档信息，存储在指定的服务器文件夹内，并保留同一文档的不同版本，基本信息采用数据库的形式进行存储。用户可以通过浏览器随时检索并提取任意版本的文档。

2 精准搜索：

内网搜索引擎提供了多种精准搜索手段，其中包括：

- 1) 关键词搜索：搜索任意关键词；
- 2) 空间范围搜索：可以搜索特定员工或特定部门的文档资源；
- 3) 时间范围搜索：可以搜索任意时间段的文档资源；
- 4) 文档格式搜索：支持 Word,ppt,excel,pdf,rtx,html 以及程序代码等主流的文档格式；
- 5) 指定位置搜索：支持标题、正文等不同位置进行搜索；

3 知识分享：

内网搜索引擎提供了多种知识分享的方式，用户单位可以根据知识分享记录实行奖惩措施，其中包括：

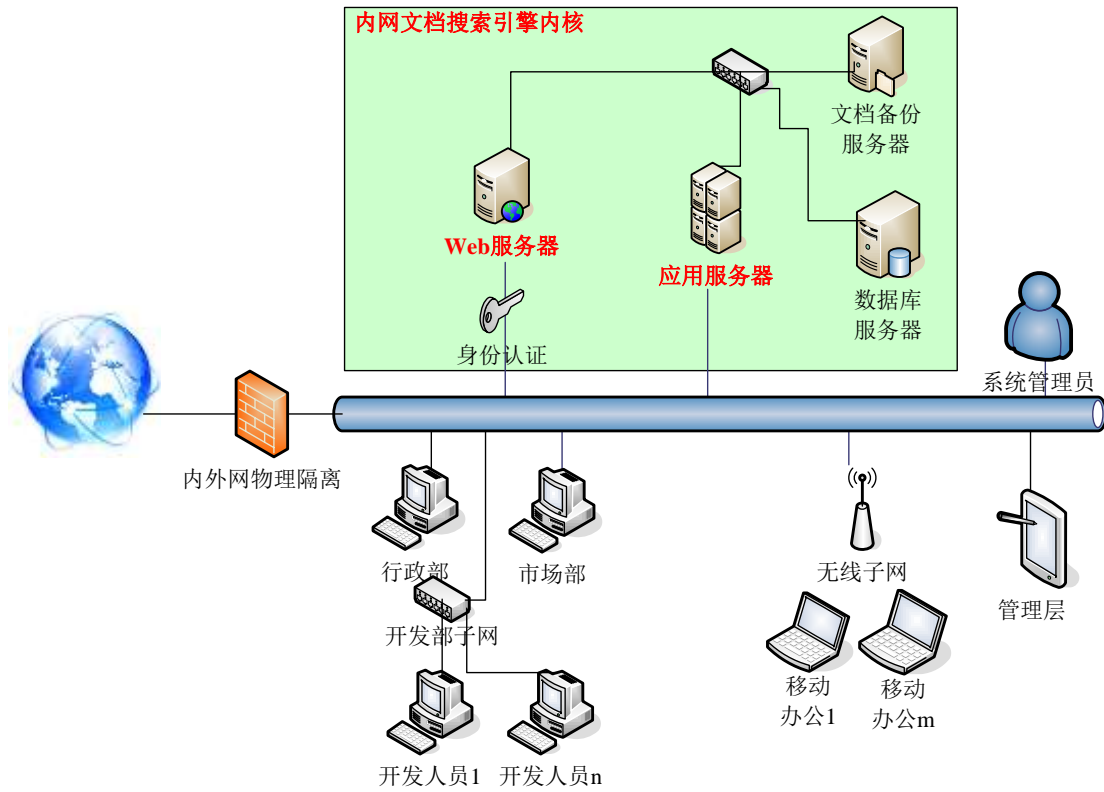
- 1) 搜索并阅读或者下载任何同事的文献资料，系统自动记录每一次分享的时间、分享者与阅读者；
- 2) 部门统计：统计各部门的知识分享记录，鼓励部门间进行知识分享；
- 3) 人员统计：统计每个人分享出的文档总数，给出排行榜，鼓励个人分享文档；
- 4) 文档 Top10 推荐：推荐关注度最高 Top 10 的文档；
- 5) 贡献人员 Top10 推荐：推荐贡献度最高 Top 10 的同事；
- 6) 阅读 Top10 推荐：推荐阅读文档数量最大的 Top 10 同事；

4 安全防护：

内网搜索引擎提供的安全防护包括：

- 1) 针对每个分享的文件夹严格实行权限管理（由所有者自行设定），权限分为三级：个人（仅限于分享者个人阅读下载）、部门（限于本部门分享）、所有（局域网内所有人员分享）；
- 2) 内网隔离措施：内网搜索引擎只能由授权用户在内网进行访问，外网无法登录内网服务，真正物理隔离，确保文档资源不被窃取流失；
- 3) 文档全过程版本备份：文档修改过程中的版本完整备份，实时按需读取恢复数据；

三、体系架构(System Architecture)



内网文档搜索引擎系统体系架构图

其中，内网文档搜索引擎系统内核中的应用服务器，Web 服务器，文档备份服务器，数据库服务器是在逻辑功能上相互独立的，在物理具体实现的时候，可以分置于多台服务器，也可以由一台普通的 PC Server 实现所有内网文档搜索引擎系统的全部功能。

四、内网文档搜索引擎系统的特色

内网文档搜索引擎系统包含三大特色

1 高效

内网文档搜索引擎系统可以管理 TB 级别的数据，数亿级别的文档，毫秒级别的搜索响应速度；系统效率是传统资源管理器性能的万倍以上。

2 简便

内网文档搜索引擎系统的简便主要体现在：

1) 普通用户无需安装任何客户端或者插件，像使用 Baidu, Google 一样方便地搜索自己以及单位内部的文档资料；

2) 在搜索和索引过程中，不占用客户机器的 CPU 和内存，不影响正常的办公；

3) 后台配置只需要设置或者导入部门人员的信息和共享的目录名称，系统会自动备份，自动搜索与自我管理；接入局域网，即可一劳永逸地实现对文档的管理和搜索、分享与下载。

3 安全：

内网搜索引擎提供的安全性体现在：

1) 内网搜索引擎只在内网运行，与外网物理隔离，不会收集用户资料，也不可能往外网传递任何文档资源；

2) 内网用户也采用了文档分级管理措施，保障信息共享的同时确保信息不能滥用；

3) 文档修改过程中的版本完整备份，实时按需读取恢复数据；

五、内网文档搜索引擎系统主要技术指标

1. 采集：在 10M 网络带宽环境下，每小时平均可以索引 10 万篇文档。

2. 检索：毫秒级别。

3. 系统能力：普通 PC Server 可以支持 20 人以内的局域网络；一般服务器可以同时支持 100 人规模的企业；5 台服务器集群可以支持万人规模的企业。