泥娃

多语种和密文全文搜索

NewaSoft®

开启您的智能安全搜索之旅

利用NewaSoft®泥娃 实现安全智能全文搜索解决方案

“泥娃搜索提供更便捷、更智能和更安全的全文搜索服务。并提供配套的全方位技术解决方案。”

图 1. 泥娃搜索

信息技术的飞速发展，对信息的安全提出了更高要求，如何实现信息安全，从信息的安全存储，安全传输到信息的安全检索，是云计算时代必须面临的挑战，如何高效的检索这些加密的非结构化数据，还是一个亟待解决的难题。

泥娃多语种和密文全文搜索系统，构建一种基于语义树的全文搜索系统，在此基础上展开加密信息的全文搜索，在信息资源加密存储的前提下，通过对其构建密文全文索引，满足人们对于信息安全的需求。

人工智能可为许多行业带来巨大发展潜力，语义理解和智能搜索技术方兴未艾，借助独有的语义树索引技术实现一体化的多语种的自然语言理解服务，结合知识图谱实现智能化的搜索服务。

泥娃搜索旨在帮助企业、单位和个人，满足信息安全和信息搜索工作需求。统一的语义树索引处理技术，结合自主研发的密文索引算法，以满足人们对于加密信息安全检索的需求。结合NewaSoft® 分布式爬虫服务、文档信息索引服务，构成成套的技术解决方案，提供标准API接口和完善的技术服务，有助于开发、部署和集成智能全文搜索技术。

从多语种文字全文搜索扩展到安全和智能搜索领域

NewaSoft® 泥娃

NewaSoft® 泥娃提供更简单、更灵活的全文搜索平台，以满足人们对于全文搜索的工作需求。通过开放的API接口实现和其他系统的互通，容易集成到企业的OA、ERP、档案管理和云存储服务，提高相关文档的利用率。

系统可帮助企业以更低的风险、更轻松地利用颠覆性的新技术。重要的是，它还可灵活地支持各种文档的导入工作，因此可以更轻松地集成其他业务和技术应用，节省成本，同时减少专门系统之间的大型文档集迁移需求。

下面我们将讨论泥娃搜索最重要组成部分。

语义树全文索引技术

- **基于路径散列的消息摘要技术**。通过信息分组、路径散列计算、结果序列调和散列，结合输出字符串的设定，从而输出消息摘要。
- **语义联想记忆技术**。通过语义标识ID的链式存储，构建语义上下的关系，实现对语句的上下文搜索，从而实现一定程度的语义会话功能。语义联想记忆系统主要用于人工智能领域的语义理解、智能机器人的人机对话、自然语言的语句搜索。

NewaSoft® 高性能全文搜索解决方案

《多语种全文搜索系统》
《密文全文搜索系统》

配套产品和应用

《集群监控系统》
《分布式爬虫服务系统》
《智能交通路径搜索分析和应用系统》

网址：www.newasoft.com
上海泥娃通信科技有限公司
苏州泥娃软件科技有限公司

- **分离编解码技术。**利用数字不同进制的转换结合分离码表，形成信息变换序列和位数序列分离，实现信息的编码；结合分离码表、变换序列和位数序列实现信息的解码。
- **密文索引算法。**密文索引技术主要利用分离编解码的特性，从而保证在密文条件下搜索的结果和原文搜索结果的一致性。

泥娃搜索产品特性

- **针对多语种的强大可扩展性。**实现国际化界面设置，简单的设置配置文件即可实现不同语言的界面设置。统一的语义树索引技术，无需分词和字典，彻底实现和语言无关的全文搜索，便于多语种全文搜索服务的部署和实施。
- **快速、经济的全文搜索网络。**语义树的索引技术，实现索引的体积仅为关键词索引的0.1%，先计算后搜索的技术保证查询时的计算仅为关键词搜索的1%，两者甚至于更低，并且随着数据量的增加，索引体积和文档的占比更低，计算效率更高。多语种全文搜索的无差异化特性，可以更经济部署多语种搜索。
- **真正的可搜索的加密机制。**自主研发的分离码算法消息摘要算法实现了密文的索引，密文的索引无法还原原始的信息，保证安全的同时，还可以提供文的检索服务。

和关键词全文搜索对比

- **索引体积更小：**索引体积仅为0.1%，随着数据量的增大，索引和信息的比值将越来越小，语句不会重复索引，同样的语义树特征节点也不会重复索引。
- **计算占用资源更少：**语义树索引技术采用先计算后查找的方法，即先计算查询语句的语义特征码，后依据计算结果查找语句，检索效率更高。
- **多语种：**独有的算法实现语言统一处理标准，统一搜索方式，无论哪种语系，均同样处理。无需分词和设置简单，无需担心新词索引处理。
- **密文检索：**信息索引采用密文加散列的方式保存，通过密文索引不能还原原始信息，原始信息的加密和密文索引的实现在客户端完成，文档加密支持第三方，安全可靠。
- **智能语义理解：**具有语料自动分析功能，自动的提取语义单元，实现NLP和后续的关联查找；提供多语种的NLP服务，支持语义理解结合最小语义知识库，提供智能化的搜索服务。

通过语义树索引推进到智能语义搜索

NewaSoft® NLP解决方案

NewaSoft® NLP解决方案提供多语种的自然语言理解服务。

系统参数

- 国际化支持：人机界面国际化，全文搜索语言国际化
- Utf8编码的文字信息全文检索

产品规格

- NW-Search-1
  - 单机提供搜索服务，开放的API接口
- NW-Search-2
  - 双机高可用部署，数据库主从服务，实现数据的读写分离，开放的API接口
- NW-Search-n
- 多级集群服务，数据存储采用多副本方案，数据的多种形式的备份，可选集群监控系统，开放的API接口。

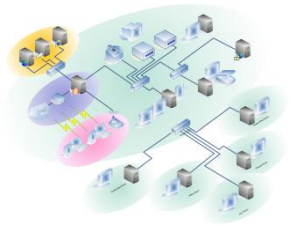

图 2 集群服务示意图

运行环境

- 操作系统：Windows或者Linux64位版本
- 数 据 库：MongoDB

关键技术指标

- 语义树索引支持2的96次方。
- 语义特征编码的散列算法支持：zy6、sm3、sha256；散列效果测试：语义树特征节点、语句特征节点和文档特征节点ID数目一致；同样计算条件下，循环测试，前一次散列结果作为下一次的输入：100万个384位信息，zy6，2540ms ；100万个256位信息，sm3，5760ms；100万个256位信息，sha256，5860ms。
- 密文索引关键算法flcode，密钥长度为K，安全解空间为K的阶乘；支持多次不同密钥的多次加密，加密后的结果支持语义树全文搜索。
- 支持所有utf8编码文字的全文索引。

泥娃搜索系统集成

提供基于http的API接口，轻松实现和第三方系统的集成。

图 3 系统集成

NewaSoft®