

# 站内全文检索系统 技术白皮书

北京词网科技有限公司

第一版 2008.5

## 目录

1 概述.....	3
2 应用范围.....	3
2.1 大型门户网站.....	3
2.2 分类信息、点评类网站.....	3
2.3 社区、论坛.....	4
2.4 博客、社交类网站.....	4
2.5 网校、教育类网站.....	4
2.5 企业网站.....	5
3 主要特性.....	5
4 系统结构图.....	6
5 系统对接方式.....	6
6 性能与可扩展性.....	6
6.1 系统性能.....	6
6.2 可扩展性.....	7
7 维护与更新.....	7

# 1 概述

本系统主要面向大中型网站、企业网站，为其提供一站式的全文、数据检索解决方案。经过简单配置，本系统可支持对数据库内信息的快速检索，也可配合根据客户网站定制的采集系统，对客户网站的页面内容进行检索。

本系统以独立服务的形式提供，客户拥有独立的维护界面，与客户现有网站采用接口形式对接，便于二次开发和后续维护。与客户应用逻辑无关的检索服务，可为各种需要检索功能的网站提供有效的帮助。

本系统采用 Linux 操作系统平台，以 C 为主要开发工具，运行效率和稳定性居同类产品前列，并为客户提供培训和远程支持服务。

## 2 应用范围

### 2.1 大型门户网站

大型门户网站的内容信息繁多，用户通过传统的频道形式逐层访问、查找信息非常困难。此时站内全文检索功能的作用就可以展现出来，通过检索功能，用户可以方便的进行全文检索，从而使网站内容被充分的利用。有效的提升了门户网站的用户停留时间和 pageviews

通过可定制的抓取引擎，本系统可以做到比通用搜索引擎更细致的抓取，此前无法被搜索引擎抓取的动态页、次级页，也可以被本系统检索到，且可以分类索引、查找，大大提高了用户体验。

本系统在此应用范围的典型客户：南方报业集团、21 世纪经济报道

## 2.2 分类信息、点评类网站

分类信息、点评类网站的重要特征是结构化信息较多、内容更新速度快。针对此类网站，本系统提供了基于数据库的检索模块，现在我们支持包括mysql、SQL Server、Oracle 在内的多种数据库接口，客户只需将要检索的内容放置于数据库，简单配置或由我们的售后服务人员帮助配置后，马上就可以利用本系统为用户提供检索服务。

针对分类信息类网站内容更新特别是新增速度快的特点，本系统支持快速索引服务，通过专门优化的算法和内存调度方案，可以确保内容发布后，5 分钟内即可被本系统检索到。<sup>1</sup>

本系统在此应用范围的典型客户：赶集网

## 2.3 社区、论坛

社区、BBS 类网站一般采用论坛软件内置的检索模块，此类模块为确保最大兼容性，一般采用数据库查询的形式进行全文搜索，在内容量达到一定程度之后，每次搜索均会占用大量的系统资源，严重的会导致程序、服务器挂起。所以，大多数论坛类网站均会对用户的检索行为进行一定的限制，此种限制虽然保证了正常运行，但却严重的影响了用户体验，使用户不得不对一些常见问题频繁发问，影响论坛的交流秩序。

本系统针对社区、论坛类网站提供了论坛检索解决方案，对论坛内的帖子、回帖进行索引，并提供基于插件形式的模块，客户只需用本系统提供的插件替换掉原有的论坛检索，即可马上解决检索难题。本系统目前提供了 Discuz!、phpwind、动网等多种论坛的检索插件，对于不常见或客户自行开

---

<sup>1</sup> 视数据量而定，需较高性能的设备支持

发的论坛，我们也有可用的解决方案。

本系统在此应用范围的典型客户：程序员社区（csdn.net）论坛检索

## 2.4 博客、社交类网站

博客类网站内容范围宽泛、话题随意性强，而且时常出现违规、违法内容，如网站监管不到，难免出现不必要的问题。

针对这一问题，本系统在提供面向用户的检索服务之外，还支持特定关键词的检索和屏蔽功能，通过网络更新，会自动更新最新的关键词表，然后做出如下两项技术限制：

- 1、对特定关键词的查询请求返回无结果提示，避免可能存在的违规内容被检索到。
- 2、对特定关键词进行内部检索，帮助网站管理员及时发现违规内容，并在第一时间予以有效处理。

本系统在此应用范围的典型客户：程序员社区（csdn.net）博客检索

## 2.5 网校、教育类网站

网校类网站的特点是存在大量视频内容，此前的检索技术无法对视频本身进行索引，只能对视频的标题、作者等索引，导致信息含量大大流失，用户无法查找到所需的视频内容。

本系统利用目前国内先进的语音识别技术，可对视频、音频内容进行索引<sup>2</sup>，

---

<sup>2</sup> 需额外付费购买视频索引模块

通过对语音数据的识别，可以做到对视频内容的搜索——只要视频中提及，就可以被搜索到。

本系统在此应用范围的典型客户：人大附中网校站内检索

## 2.5 企业网站

大型企业特别是服务型企业网站内容繁杂、各种自助服务的入口繁多，用户不容易上手使用。导致用户学习成本高、进而不愿意利用网上自助来完成服务，加大了企业呼叫中心、售后服务人员的投入。

本系统可以为企业定制一键式搜索系统，不仅可以通过搜索框搜索网站内的内容，还可以输入特定内容实现特定的服务功能。以我公司典型客户——北京移动网站为例，用户在搜索框内可以输入彩铃的名字，进行彩铃定制，还可以直接输入手机号，查询话费、变更套餐等。

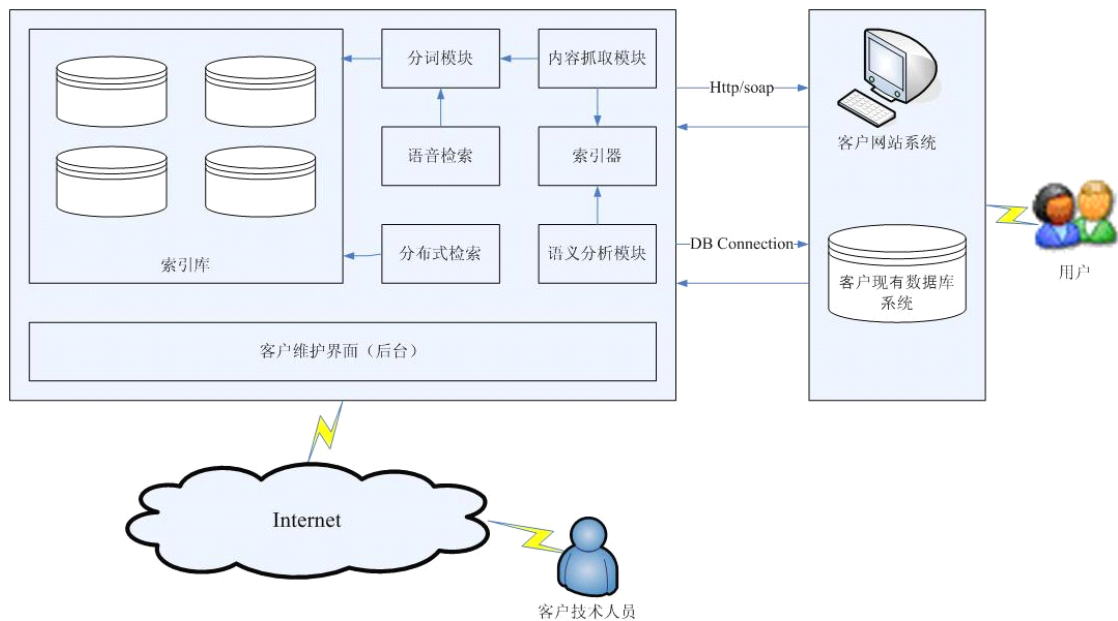
## 3 主要特性

综上所述，本系统的主要特性包括以下几点：

- 针对客户网站及内容结构进行针对性的开发，尤其在抓取引擎的开发上，可以照顾到客户网站的特性，做到最少遗漏。
- 支持直接对数据库表进行索引，并拥有方便使用的后台维护界面，普通技术人员即可很快上手进行操作。
- 大数据量的快速索引支持、分布式检索支持。
- 优越的安全防范措施，可避免越权访问，并可在企业内网中建立服务，无需访问公网

- 与现有系统、软件的无缝对接，提供插件、接口等各种对接形式，并配有售后培训服务，确保客户技术人员很快上手。
- 符合中国国情的关键词实时过滤及有害信息查询。
- 语音识别技术的应用，可以对视频内容进行有效的检索。
- One-Box 搜索技术，可在搜索框中实现更多的服务性功能。

## 4 系统结构图



## 5 系统对接方式

本系统与客户现有系统采用 HTTP POST 接口的形式对接，无论客户网站采取何种技术开发，均可以方便的与本系统对接，很容易的实现整个检索功能。

在此基础上，我们提供 ASP、PHP、JSP 等各种主流编程语言的库函数，开发

人员只需调用库函数，即可使用本系统提供的各种 API。

为确保数据安全，在接口协议中，系统规定了相应的鉴权方式，避免未经授权的检索或数据抓取。

API 接口细节请参见：《词网全文检索系统 技术开发接口说明书》

## 6 性能与可扩展性

### 6.1 系统性能

本公司产品可达到如下性能指标：

- 海量数据的处理能力：系统将能够顺利处理百万文档以上的数据集合，数据量在 T 级以上；
- 离线分析功能的快速准确：对几万样例数据的分类训练等功能的运行时间在以小时为量级的时间范围内；
- 批量数据处理高速：对百万数据的批量处理，每天的数据处理量为百万量级；
- 对分词等基础功能其准确率达到 98%以上，一般功能的准确性在 90%左右，并均提供对可疑数据的人工确认与机器学习手段，以便不断提高准确度。

### 6.2 可扩展性

本系统既可以单机运行，也可以部署在多台设备上实现集群，以提高大规模数据的索引、查询速度，理论上讲，本系统容量无特定上限，只需不断新增集群设备即可满足数据增长的需求。

本系统采取模块化开发，内容抓取、语音识别、快速检索、语义关联等模块均可为客户定制开发，并可以方便的替换或增删，在取得较强的配置伸缩性的同时，也为客户降低了许多没必要付出的成本。

## **7 支持与维护**

### **7.1 人员培训**

本系统包括 20 课时的维护培训，为客户培训可以承担基本维护和开发任务的人员 2 人次。

### **7.2 售后技术支持**

产品售出后，即为客户配备指定的售后技术支持人员，协助客户进行系统配置和接口开发，对于有特定需求的客户，还可为客户对系统进行有针对性的定制。

### **7.3 远程协助**

利用远程维护功能，售后技术支持人员可以远程登入产品维护后台，帮助用户解决常见问题、协助配置或通过日志定位、排除故障。

### **7.4 自动更新**

系统内置自动更新功能，经授权的客户可以无需人工干预，系统会自动更新最新的分词规则、关键词及修补必要的系统漏洞、升级软件。