

Future 互联网内容采集和分析系统

采集、分析、检索互联网内容

概述

当今互联网已经成为传播信息最快最方便的途径，每天在互联网上都有无数的网站和网页正在产生。用户可以通过搜索引擎查找信息，但是由于搜索引擎是为所有互联网用户服务的，所以用户无法通过搜索引擎方便的查找到自己指定的一组网站的内容，搜索引擎也不会主动推送这些网站的内容给用户。

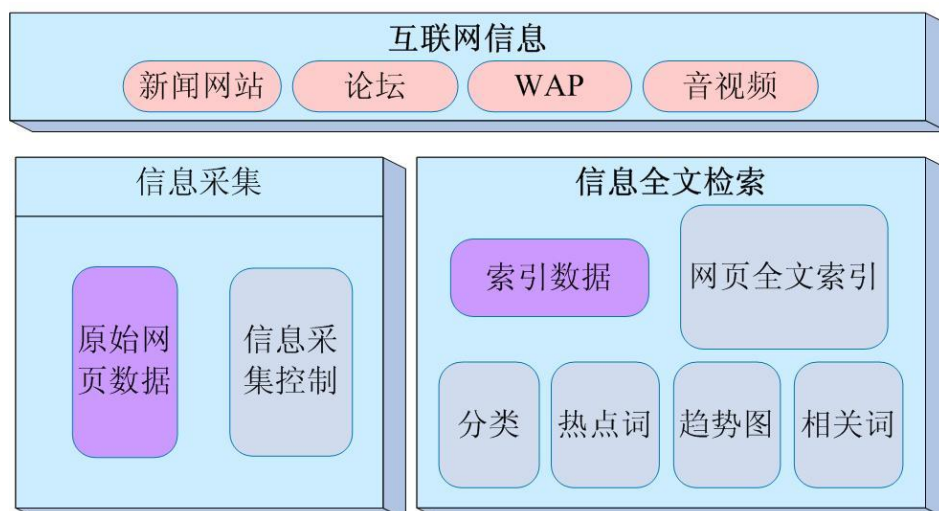
智信远景软件技术有限公司多年来一直对互联网信息分析进行深入研究，依托自身研发的中文自然语言信息处理技术，推出了 Future 互联网内容采集与分析系统，这套系统很好的解决了这一问题。用户可以通过图形化界面输入自己需要采集的网站网址等信息，该系统会定时采集用户指定网站，保存在数据库中，同时用户可以通过该系统的搜索引擎搜索网页。该系统也可以将互联网中热门信息或用户定制的关键字信息推送给用户，极大的方便了用户对特定网站内容的监控。

软件体系结构

系统结构

该系统主要分为采集和信息索引及全文检索两个模块。

采集模块负责网页的抓取，并对整个采集过程进行控制和监视。爬虫程序自动采集的网站包括新闻网站，论坛，博客，WAP 网站和音视频页面信息，并且在指定的时间段内自动下载网站更新。网页被下载后，保存在指定的磁盘阵列。信息全文检索模块负责建立全文索引，提供全文检索服务。其中一台索引服务器建立索引并提供全文检索服务，另一台服务器通过对信息的智能语言分析后，提供分类，热点词，趋势图分析，相关词等服务。



逻辑结构

该系统设计的逻辑结构图如下。

最上层为需要采集的互联网数据，系统会对采集到的互联网数据进行 HTML 分析，元数据提取以及数据入库的操作。

中间层是 URL Table 数据，系统将这些数据保存在设计好的 MySQL 数据库中，同时采用集群的方式处理数据，保证整个数据处理的高效率。

第三层是索引数据库，系统在进行全文索引，分类索引，NLP 处理后，会将所有相关数据保留到这里。这里的设计也同样采用了集群的方式，保证整个数据处理和检索的高效性。



产品特点及优点

1.实时性

- 网页爬虫实时采集互联网数据。
- 信息分析模块实时处理采集到的互联网数据。

2.稳定性

- 实现 7*24 小时不间断采集网络数据。
- 集群式的设计方式保证系统稳定。

3.高效性

- 软件自动发现互联网热点信息，及时呈现给用户。
- 用户可以通过内嵌的搜索引擎快速发现信息。
- 多线程爬虫高效采集网页数据。

4.安全性

- 访问软件查看信息，需要用户名和密码。
- 内嵌数据库的访问受密码保护，数据集中存储和备份。

5.智能化

- 中文自然语言技术处理互联网网页数据，自动将分析结果呈现到网页上，供用户浏览。
- 智能分析网页数据，提取有用文字及元数据，过滤广告等杂乱信息。
- 智能增量采集网页，增量处理数据，保证系统性能。

系统需求

1. 软件需求

本软件的技术架构决定了运行环境的灵活性和可扩充性

- 服务器端操作系统：Windows 2000 以上；
- 客户端要求：IE5.5 以上；
- 网络环境：Intranet 与 Internet；

2. 采集网站数量与服务器配置对应表

采集网站数量	服务器配置建议		
	CPU	内存	硬盘空间
0-100	英特尔(R) 至强(R) 双核处理器 E3065 2.33G Hz	1GB DDR	视采集网站及索引数据量而定。
100-500	英特尔(R) 至强(R) 双核处理器 E3110 3.0G Hz	2GB DDR	
500-1000	英特尔(R) 至强(R) 四核处理器 X3220 2.4G Hz	4GB DDR	
1000 以上	英特尔(R) 至强(R) 双核处理器 X3230 2.66G Hz	8GB DDR	

注：

- 采集、索引、检索均需使用独立的服务器。
- 500 个网站以上建议服务器采用负载均衡设计。
- 此服务器配置建议，并非服务器最低配置要求。

更多信息请访问

www.finfosoft.com

联系方式：

北京智信远景软件技术有限公司

电话：(010) 85997746

传真：(010) 85997745

Email: sales@finfosoft.com