DOI:10.16180/j.cnki.issn1007-7820.2014.02.038

电子科技 2014 年第 27 卷第 2 期
Electronic Sci. & Tech. /Feb. 15, 2014

学术期刊电子论文检索系统设计

蒋从文，李隐峰，齐　鹏，杨志英
(西安电子科技大学 电子工程学院，陕西 西安　710071)

摘　要　设计了一种能将各个学术期刊网站上的电子论文信息采集到一个统一的数据库中并提供检索的系统。系统分为数据采集、数据分析和存储、数据检索 3 个模块。前两个模块负责将互联网上电子论文的内容结构化存储到本地数据库，最后一个模块负责对数据库内容生成索引并提供查询。目前，该系统已存有 150 万篇中文期刊论文。

关键词　数据采集；数据检索；Sphinx；全文索引
中图分类号　TP274⁺.2　　**文献标识码**　A　　**文章编号**　1007－7820(2014)02－122－03

The Design of Academic Journal Electronic Papers Retrieval System

JIANG Congwen, LI Yinfeng, QI Peng, YANG Zhiying
(School of Electronic Engineering, Xidian University, Xi'an 710071, China)

Abstract　Many academic journal have website on internet, so more people can search papers from it. This paper designs a system for collecting the electronic papers on websites to a database and providing retrieval service. The system has three modules. They are data acquisition, data analysis and storage, data retrieval. The first two modules are responsible for the storage of structured electronic paper on the Internet to the local database, The last one is responsible for the generation of database index and providing retrieval service. There have 1.5 million electronic papers in this system.

Keywords　data acquisition; data retrieval; Sphinx; full-text index

互联网上散落着海量的电子论文，它们分布在不同的期刊站点，要在最短的时间内查询到最多的期刊论文并不容易。本系统目的就是将分散在各处的电子论文整合到一个数据库中，并提供统一的查询接口，方便用户在更大的范围内查找所需内容，提高查询效率，同时也增加了电子论文潜在的读者。

1　系统设计

1.1　系统总体结构

整个系统是基于 B/S 架构的，分为数据采集、数据分析和存储、数据检索 3 个模块，符合软件设计低耦合原则。这 3 个模块可以工作在不同的计算机上，形成一个分布式系统。3 个模块的结构如图 1 所示。

系统工作流程：首先通过数据采集模块将期刊站点服务器上的 HTML 页面获取到本地，然后数据分析模块会对 HTML 网页进行解析，提取 HTML 网页中需要的论文基本信息，之后存入数据库。由于存储的数据量规模比较大，纯粹利用数据库的索引加快查询速度已不现实，因此增加数据检索模块对数据库建立单独的索引，这样用户输入查找关键字后不会直接去查询数据库，而是去查询数据检索模块建立的索引，再由单独的索引得到查询结果返回给用户。

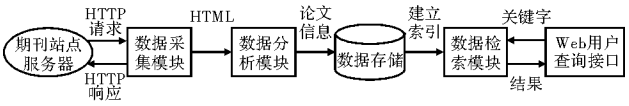

图 1　系统结构图

1.2　数据采集模块

数据采集模块也叫网路爬虫，是系统中关键且基础的构件[1]。它要将网页 HTML 数据下载到本地以供之后的进一步处理。本系统要采集的具体目标有两类：由电子期刊站点自带检索接口查询得到的结果页面和结果页面里每篇电子论文的详情页面。网络爬虫会采集结果页面里的每篇论文，然后转到下一个结果页面继续采集，直到所有结果页面采集完毕，则该站点采集完毕，转到下一个期刊站点。整个采集过程使用的是一种深度优先的采集策略。采集目标的树状图如图 2 所示。

收稿日期：2013-01-28
作者简介：蒋从文(1990—)，男，硕士研究生。研究方向：网络信息系统开发。E-mial：jiangcongwen110@163.com。李隐峰(1974—)，男，副教授。研究方向：Web 信息系统，网络安全。齐鹏(1987—)，男，硕士研究生。研究方向：网络信息系统开发。杨志英(1991—)，男，硕士研究生。研究方向：网络信息系统开发。

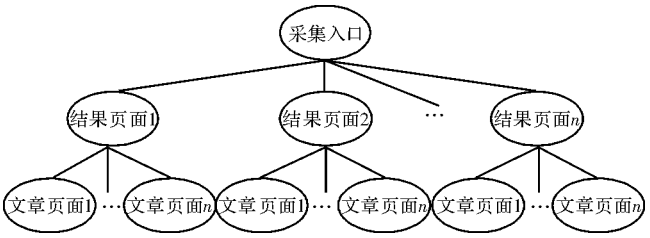

图2　采集页面树状图

本模块向期刊站点发送 HTTP 请求时可能要向其提交数据，这些数据可以利用专用 HTTP 分析工具看到，如 Live HTTP headers。对于期刊站点来说，它实际上是利用某种论文采编系统生成的，要向搜索结果页面提交的数据一般是固定或是有规律的，只需要找到提交数据的规律性就可采集到所有的搜索结果页面。

1.3　数据分析和存储模块

数据分析和存储模块目的是从采集到的 HTML 数据中提取出论文的基本信息，并存储到数据库中。这是一个将非结构化的数据转换成结构化数据的过程。

HTML 文档对象模型（HTML Document Object Model，DOM）定义了访问和处理 HTML 文档的标准方法。在 DOM 中，HTML 文档中的每个成分都是一个节点（Node）。DOM 是这样规定的：整个文档是一个文档节点；每个 HTML 标签是一个元素节点；包含在 HTML 元素中的文本是文本节点；每一个 HTML 属性是一个属性节点；注释属于注释节点。

节点彼此都有等级关系。HTML 文档中的所有节点组成了一个文档树（或节点数）。HTML 文档中的每个元素、属性、文本等都代表着树中的一个节点。树起始于文档节点，并由此继续伸出枝条，直到处于这棵树最低级别的所有文本节点为止。

可以将这种等级关系概括为两类：包含（嵌套、父子、继承）关系和邻居（并列、相邻、兄弟）关系，如图 3 所示是一个文档树（节点数），可以清晰地看到各个节点之间的层次关系，如图 3 所示。

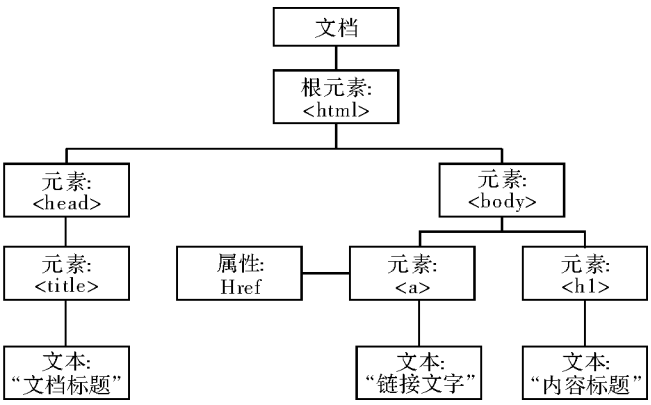

图3　HTML 文档树模型

数据分析模块就是查找电子论文基本信息所在的节点并从节点中提取出所需要的信息，然后存储到结构化的数据库中。

1.4　数据检索模块

本系统使用的是 MySQL 数据库，需要对存储的论文信息建立全文索引，如论文的标题、作者、摘要、关键字等。但 MySQL 全文索引对中文支持并不理想，且当数据库中记录规模越来越大时，多列索引并不能满足查询速度的要求。

基于以上考虑，系统使用 Sphinx 作为一个单独的检索模块。Sphinx 是一个独立的全文检索引擎，其全文检索查询速度要远快于 MySQL，且 Sphinx 可以轻易地与 SQL 数据库和脚本语言集成。Sphinx 主要包括 Indexer、Search、Searchd、Sphinxapi 这 4 个部分。

整个 Sphinx 的功能主要分为两个部分：索引模块和检索模块。结构示意图如 4 所示。

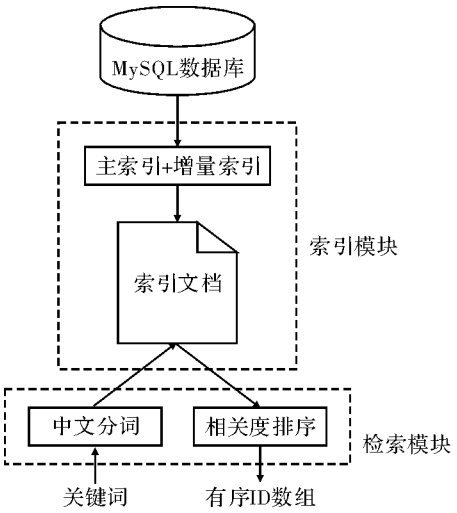

图4　Sphinx 功能结构示意图

索引模块要做的工作就是对数据源的内容进行分析，提取出所有的关键词，同时记录下每个关键词出现的次数和位置等，形成一个倒排索引文件。该倒排索引文件有如下特点：对于其中出现的每个关键词，都可以知道其在数据源中出现的次数和每一次出现的位置在哪里，并可以根据它对原始数据源进行访问。之所以称为倒排索引文件，是因为一般索引文件都是从给定内容中查找相应的关键词，而倒排索引文件恰好相反，是通过给定的关键词去查找这些关键字出现的内容的位置。

Sphinx 创建索引的一次流程如图 5 所示。

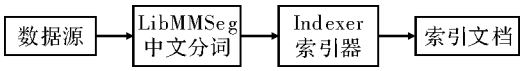

图5　Sphinx 创建索引流程图

首先需要将数据库中要建立索引字段中的内容按关键词进行分词。Sphinx 索引器 Indexer 不支持中文分词，因此通过 LibMMseg 中文分词软件包对数据库中的中文数据进行分词，然后由 Indexer 分析并生成索引文件[2]。

检索模块是 Sphinx 的另一核心部分。这一部分目的就是对搜索结果进行排序，而排序的最主要依据就是相关度。Sphinx 的相关度排序是对词组评分和统计学评分进行权值计算得出检索结果排序。词组评分是根据检索词与检索文档的精确匹配程度，以及出现在文档中不同位置的信息给予不同的权重。统计学评分则基于经典的 BM25 函数，主要根据检索词在检索文档中的出现频率和在整个索引文件中的出现频率来计算权重[3]。Sphinx 通过守护进程 Searchd 对外提供搜索服务[4-5]。

2　数据统计

目前，系统已经采集了 1 004 个期刊站点，按学科分为 9 个大类，统计数据如表 1 所示。

表 1　期刊站点按学科统计数量表

学科类型	站点数量
电子通讯	54
数学与物理	55
地球与环境	45
经济与管理	40
生命科学	50
工程技术	300
医药卫生	320
人文社科	60
化学与材料	80

系统搜索结果会给出论文标题，摘要和关键词，搜索关键词高亮显示，提供下载的会给出下载链接，搜索结果界面如图 6 所示[6-7]。

图 6　搜索结果页界

3　结束语

本文对电子论文检索系统的各个模块进行了相应的介绍，目前系统已经采集到 150 万篇以上论文，数据库数据大小约 3 GB，检索一个关键字花费时间在几十 ms，检索范围包括论文标题、作者、摘要、关键词。只需一次查询，就可得到来自多个期刊的相关内容，且结果已经排好序，大幅提高了用户的查询效率。

参考文献

[1] 张俊林. 这就是搜索引擎：核心技术详解[M]. 北京：电子工业出版社，2012.
[2] 刘清明，彭宇扬，彭自成. 基于 Sphinx 的 Web 站内搜索引擎的设计与实现[J]. 微计算机信息，2010，26(5)：116－118.
[3] 曾湛伟. 基于 Sphinx 的特色数据库全文检索系统的设计与实现[J]. 现代图书情报技术，2010(6)：78－82.
[4] 许天亮，王义峰，曾平. 个性化元搜索引擎技术研究[J]. 电子科技，2008(1)：56－59.
[5] 林欢欢，庄福振，王文杰，等. 一种新型网络信息采集器的研究[J]. 计算机仿真，2009(5)：129－133.
[6] 刘华. 多通道数据采集系统设计[J]. 电子科技，2012，25(6)：24－26.
[7] 谭德坤，王力红. 基于模糊语言方法的信息检索系统的研究[J]. 计算机仿真，2005，22(2)：152－155，177.