一个基础教育网站搜索引擎的设计与实现

陈 权 曹卓文 杨晓江

（南京师范大学教育技术系 南京 210097）

【摘要】 在研究网站元数据的基础上，介绍一个以基础教育网站为检索对象的搜索引擎系统。结合基础教育网站的特点，分析该系统的关键技术，如主题蜘蛛搜索、网站分类、网站信息提取等，并对系统的整体架构、功能模块进行详细描述。

【关键词】 主题蜘蛛 网站分类 信息提取 搜索引擎 【分类号】 TP393

Design and Implementation of a Search Engine for K12 – related Websites

Chen Quan Cao Zhuowen Yang Xiaojiang

(*Department of Education Technology*, *Nanjing Normal University*, *Nanjing 210097*, *China*)

【**Abstract**】 On the basis of some research work on metadata for Website, this paper introduces a search engine system for Websites related to K12 education. Combined with the characters of K12 – related Websites, some key technologies are analyzed, such as topic – spider search, Website classification, information extraction for site etc. Both of the whole architecture of the system and the function modules are described in detail in the paper.

【**Keywords**】 Topic spider Website classification Information extraction Search engine

在互联网教育资源不断增长的同时，资源获取的有效性和便捷性却逐渐成为问题。为此，笔者设计并实现了一个以基础教育网站为检索对象的搜索引擎系统，旨在使教育资源站点能够充分被共享。与通用的网页搜索引擎相比，该系统具有以下特点：

（1）检索对象为基础教育网站，而非通用搜索引擎所检索的网页，针对性强；

（2）提取了网站的学科、学段、区域、单位、地址等相关属性，方便用户检索；

（3）更贴近用户需求，注重对资源质量的管理与评价；

（4）加入了人工校验、用户报错等模块，以确保资源有效性；

（5）提倡共享，支持用户向系统推荐站点，支持用户向其他好友推荐优秀站点，以供其他用户参考。

1 总体设计

1.1 网站元数据

关于教育领域学习资源标准的研究，我国教育信息化技术标准委员会已研制并形成了比较完整的标准体系。

收稿日期：2007 – 04 – 20
收修改稿日期：2007 – 04 – 26

CELTS – 42[1] 对基础教育资源元数据的制定提供了详细的规范说明。该规范从资源内容描述类（包括标题、学科、关键词、描述等 10 个数据元素）、知识产权信息类（包括作者、出版社、其它作者等 5 个核心元素）、外部属性描述类（包括日期、类型、格式等 8 个核心元素）3 个方面对核心元数据元素进行了描述。本系统在此规范的基础上，结合基础教育网站的自身特点，制定了描述基础教育网站的相关元素，详见表 1。

表 1 基础教育网站元数据

元素	含 义
标题	指网站首页标题
学科	指网站的学科属性
关键词	用来标明网站主要特征的词语，供检索时用
描述	包括导航描述与摘要描述
标识	指网站的域名
格式	通常为 text/xml 格式
日期	包括网站的创建时间与收录时间
语种	指网站的语种
类型	指网站的类别
作者	指网站的责任者，包括创建者、单位、地址、联系电话、电子邮件等
适用对象	指网站的主要访问对象

1.2 系统框架

从功能上来看，系统主要由自动采集、自动标引、管理维护、用户交互服务 4 个部分组成，如图 1 所示。

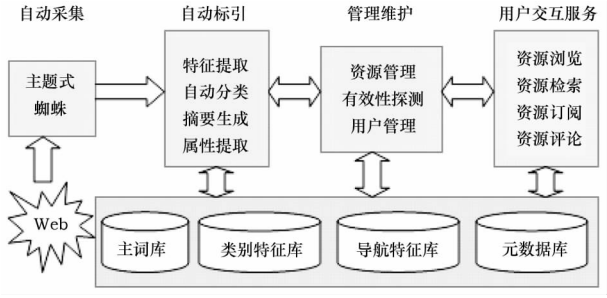

图 1 系统框架图

具体来说,自动采集部分由一个主题蜘蛛程序构成,其功能是通过 HTTP 协议在网络上将与基础教育相关的 Web 页面下载到本地资源库;自动标引部分是对本地资源库中网站资源的相关属性进行自动标引,包括特征提取、自动分类、自动摘要、属性抽取几个模块;管理维护部分由资源管理、有效性探测及用户管理几个模块构成,相应功能包括:对网站属性的准确性进行人工校验、对网站 URL 链接的有效性进行周期性探测、对后台操作用户进行管理;用户交互服务部分提供一个用户界面接口,向用户提供资源检索、浏览、订阅、评论等一系列服务。

2 关键技术

2.1 主题蜘蛛搜索

在通用的搜索引擎系统中,数据的采集工作由网络蜘蛛(Web Spider)完成。蜘蛛程序从一个 URL 种子队列出发,通过 HTTP 协议请求并下载 Web 页面,分析页面并提取链接,加入 URL 队列。蜘蛛程序抓取 Web 页面的策略可以是广度优先或深度优先。通用搜索引擎系统中网络蜘蛛不对队列中的 URL 进行分析,而是全部下载到本地数据库,这种抓取模式并不适合主题领域。本系统采用的是面向特定领域的搜索策略——主题蜘蛛搜索。

通常,主题蜘蛛的搜索策略主要有基于内容评价与基于链接结构评价两种。基于内容评价的搜索策略是一种根据主题与链接文本的相似度来评价链接价值高低的策略,文献[2]介绍了一种基于连续值的相似度函数来计算页面与主题相似度的方法。但这种搜索策略的不足之处是没有考虑到 Web 页面之间的链接关系。相比之下,基于链接结构评价的搜索策略则考虑到了 Web 页面的这种半结构化特征,该策略正是通过对 Web 页面之间相互引用关系的分析来确定链接的重要性,从而决定链接访问的顺序,其代表性方法有 Page - Rank 方法[3] 与 HITS 方法[4]。由于只考虑到链接之间相互引用的关系,这种基于链接的搜索策略存在着“主题漂移”的缺陷。基于此,本系统采取了一种基于综合价值评价的搜索策略,该策略综合了页面文本信息与链接的结构信息两方面因素,对链接的综合价值进行计算。实验结果表明,该搜索策略能有效提高搜索效率和搜索结果的相关性。

2.2 网站分类

网站元数据元素的标引是借助于网站分类技术与网站信息提取技术共同完成的。网站分类是将主题蜘蛛下载的网页以网站为单位进行类别判定。关于网站分类,文献[5]提出以特征向量集(Sets of Feature Vectors)表示网站特征进行分类,文献[6]则提出根据站点的物理与逻辑结构合并网页,从而进行主题识别。与以上算法所适用的普通网站不同的是,基础教育网站有其自身特点,经过调研,笔者发现大部分教育网站的标题、导航具有规范性,有助于分类。另一方面,文献[5]、文献[6]介绍的算法需要分析网站所有网页,执行效率低。因此,本系统采用一种渐进式的分类方法,即渐进地利用网站的标题、导航、网页文本进行分类。这样,部分网站直接通过标题或者导航就可以实现类别判定,从而在保证正确率的同时提高分类效率。下面依次介绍这 3 种分类方法。

(1)基于标题分类

通过对大量基础教育网站的统计分析,笔者发现,相当一部分网站的首页标题包含网站类别特征的信息。以校园网为例,在收集的 445 个校园网中,首页标题为“XX 中学”的有 267 个,为“XX 小学”的有 103 个,共占 83.1%;而教育行政类网站的首页标题则通常包含“教育局”、“教育厅”。由此,笔者建立了通过标题判定网站类别的规则集,满足某条规则的网站即可被判为相应类别。

部分规则如下(以校园网为例):

Has(Title,“实验中学”)

Has(Title,“附属中学”)

Has(Title,“高级中学”)

Ends with(Title,“中学”)

如某网站首页标题为“江苏省扬州中学”,则判为校园网;某网站标题为“仁怀四中高级中学校园网”,也被判为校园网。对于标题分类无法判别的网站继续采用导航分类进行分类。

(2)基于导航分类

网站首页的标题是网站主题的高度概括,而网站导航文本则是网站内容组织和网站结构的特征化概述,导航文本一般是一个词或两个词组成的短语,各类基础教育网站的导航锚文本显示了不同的特征。表 2 中列举了部分类别基础教育网站的导航锚文本。

从表 2 可以看出,不同类别网站的导航文本具有不同的特征,因此,可以利用导航文本进行网站类别判定。

表 2 部分类别基础教育网站导航锚文本

网站类别	导航锚文本
校园网	学校简介、校园动态、德育之窗、教师频道、学生频道、家长频道……
教育行政网	机构设置、教育新闻、政策法规、网上办事、招生考试……
教育门户网	教育新闻、中考、高考、考研、出国留学、招生、考试、外语、培训、求职……
教育资源网	试卷、课件、教案、素材下载……
教育科研网	科研动态、科研成果、课程改革、课题研究、学科教研……

具体分类过程如下：

①根据导航特征库，对导航进行词频统计；

②对词条进行加权计算，得到权重总和。计算公式为：

$W = f_1 * w_1 + f_2 * w_2 + \cdots + f_i * w_i$（$f_i$ 为第 i 个词条的词频，w_i 为第 i 个词条在导航特征库中的权重）；

③选取越过类别阈值的权值，取其中最大值所属的类别为网站类别。若没有越过类别阈值的权值，则表示该网站无法根据导航来进行类别判定，需要继续根据网站内容分类。

（3）基于网站内容分类

基于向量空间模型（VSM）的类中心向量法以其良好的泛化性能被广泛采用，文献[7]指出，此方法在基础教育文本分类中表现出较好的查准率和查全率，因此，本系统采用此算法利用网站内容进行分类。

基于 VSM 的类中心向量法根据既定的分类体系从大量训练文档中统计出每个类别的中心向量，通过计算待分类文档与类中心向量的空间距离获得该文档与各类别的相似度，从而判定类别。

具体分类过程如下：

①计算各类别训练集所有文档向量，得到类别向量集，取其算术平均作为类别中心向量；

②对待分类网站内容进行预处理、分词，采用词频加权统计将其表示为特征向量；

③计算对待分类网站与各类别的相似度，本系统采用余弦距离，计算公式为：

$$\mathrm{Sim}(D_i, D_j) = \frac{\sum_{k=1}^{M} W_{ik} \cdot W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}}$$

其中，$\mathrm{Sim}(D_i, D_j)$ 为待分网站与第 j 类的相似度，D_i 为待分类网站的特征向量，D_j 为第 j 类的中心向量，M 为特征向量的维数，W_k 为向量的第 k 维（即第 k 个词条的权重）；

④对所有越过阈值的相似度值，取其中最大值的类别作为网站类别。

2.3 网站信息提取

除了网站的类别属性需要自动判定外，网站的其他属性（如描述、作者、关键词等）也需要逐一自动标引。下面分别介绍网站摘要提取技术与网站作者信息提取技术。

（1）网站摘要提取

网站的摘要与普通文本摘要有所不同，网页中丰富的标记极大地表现了内容，怎样才能提取出概括网站内容的信息呢？如上文所述，网站导航栏目是整个网站内容组织的高度概括，因此，本系统基于词频加权统计，选取相关度高的导航词条和正文句子构成网站的摘要。

具体步骤如下：

①以句子为单位进行词频统计；

②对词条进行加权计算，得到句子的权值。计算公式为：

$W = (f_1 * w_1 + f_2 * w_2 + \cdots + f_i * w_i)/i$（$f_i$ 为第 i 个词条的词频，w_i 为第 i 个词条在导航特征库中对应权重，i 为句子所包含的词条数量）；

③根据句子的权值对句子进行排序，选择权值高的句子作为文摘句。

本系统对导航词条和正文句子的提取分别进行，最后综合两者构成网站摘要。

（2）网站作者信息提取

根据网页结构的特征来看，大部分网站的作者信息出现在网页底部，并且具有一定的规则性，因此，本系统对网站作者信息的提取采用基于规则的方法。

本系统提取的网站作者信息包括网站作者名称、地址、邮编、电话、传真、E - mail 等。笔者为各元素建立了规则集，如地址信息，前面往往包含“地址”二字，其后又常跟邮编信息；又如联系电话，前面包含“电话”、“Tel”、“热线”等，且电话号码一般都是 7 位或 8 位数字，或者带有区号，这些都是电话元素的特征。实验结果表明，这种基于规则的方法能有效地提取大部分网站的作者信息。

3 系统实现

本系统主要基于 Microsoft Visual Studio. Net 2003 和 SQL Server 2000，采用 C#语言开发。后台主要为控制台应用程序和 Windows 应用程序，用户服务部分则采用 B/S 的 Web 应用架构，下面简要介绍相关部分的实现。

图 2 为本系统分类模块的类图（图中列举了主要类的部分属性和方法），图中 WebSite 类表示网站类，包括网站的首页面（HomePage）、网站相关子页面的集合（Pages）以及学科、学段、责任者等其它属性；WebPage 类表示网页类，包括网页的标题、源码、URL 等其它属性；WordSegmentor 类表示分词类，负责供其它类进行调用；ClassifierLibrary 类表示特征库类，包括网站标题、导航、正文 3 个子特征库，供分类器使用；接口 Classifier 定义了分类器的基本方法，具体实现分别由 TitleClassifier、NavigationClassifier、ContentClassifier 3 个类来完成，从而形成了 3 个独立的分类器类。从类图关系来看，WebSite 类处于核心地位，它与 WebPage 类构成了一种组合关系，同时也依赖分词类、特征库类及分类器类。关于分类算法的具体实现细节，上文已进行了详细介绍，即 WebSite 类渐进地调用 TitleClassifier、NavigationClassifier、ContentClassifier 进行分类，分类过程中依赖分词类 WordSegmentor 和特征库类 ClassifierLibrary。

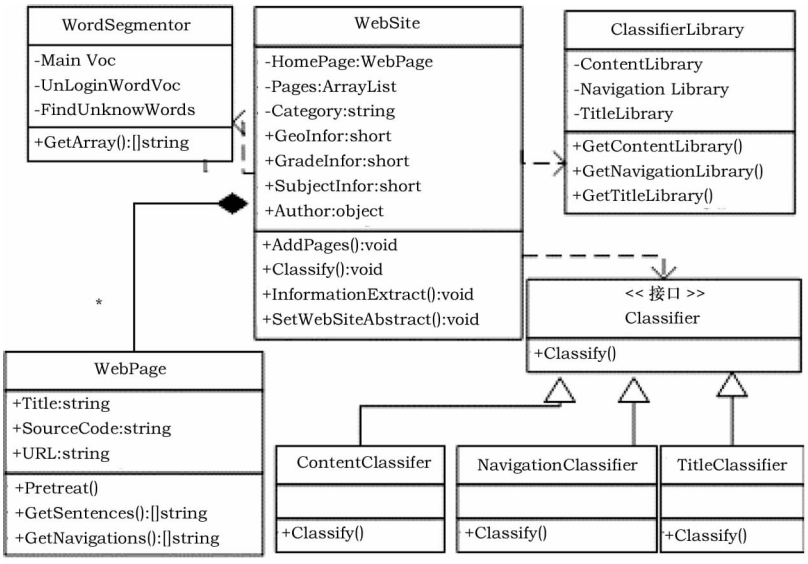

图 2 分类模块类图

在用户服务部分 Web 应用的实现环节上,笔者采用了.Net 环境下的多层架构技术[8]。多层架构的划分,优点是可以使系统更易设计、维护、伸缩性强、灵活性高。本系统将用户交互服务部分从逻辑上分为表现层、业务层、数据访问层、公共层和数据库层等 5 个子层。为了便于说明,这里分别用 Web 层、BLL 层(Business Logic Layer)、DAL 层(Data Access Layer)、Common 层和 DB 层来表示。图 3 展示了各层之间的相互调用机制:用户先通过 Web 浏览器和 Web 层交互,Web 层调用 BLL 层的服务,BLL 层再调用 DAL 层的服务,最后,DAL 层调用 DB 层服务。此外,Common 层还负责为 Web 层、BLL 层和 DAL 层提供公共性的服务。

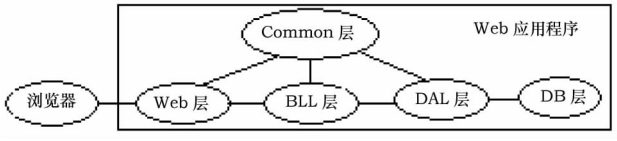

图 3 Web 应用的分层

4 结 语

本文在研究网站元数据的基础上,介绍一个以基础教育网站为检索对象的搜索引擎系统,对系统的总体设计进行详细描述,并着重分析系统的一些关键技术及实现细节。从功能上来看,本系统具备自动发现、获取、标引网站元数据的能力,并能够对已标引的元数据提供用户浏览、检索、订阅、评价等服务。实验结果表明,该搜索引擎系统能够把分散在各地的基础教育网站有效地聚合组织起来,且对于每个网站都给出了详细的描述信息,从而提高了教育资源的共享性。

参考文献:

1 教育部基础教育课程教材发展中心. CELTS - 42 基础教育资源元数据应用规范. http://www.celtsc.edu.cn/680751c665875e93/folder.2006 - 04 - 03.8417036039/celts - 42/celts - 42 - 1 - cd1 - 6.pdf(Accessed Mar.19,2007)

2 Hersovici M, Heydon A, Mitzenmacher M, et al. The Shark - Search Algorithm - An Application: Tailored Web Site Mapping. World Wide Web Conference,1998

3 Cho J, Garcia - Molina H, Page L. Efficient Crawling through URL Ordering. Computer Networks, 1998, 30(1 - 7):161 - 172

4 Kleinberg J M. Authoritative Sources in a Hyperlinked Environment. Association for Computing Machinery,1999,46(5):604 - 632

5 Hans - Peter Kriegel, Matthias Schubert, Classification of Websites as Sets of Feature Vectors. The IASTED International Conference, Austria, 2004

6 余智华. WWW 站点的分析与分类:[学位论文]. 北京:中国科学院,1999

7 田俊华. 基于 Web 的中文文本自动分类研究与实现:[学位论文]. 南京:南京师范大学,2004

8 杨晓江. .Net 环境下 Web 应用的通用设计. 计算机工程与设计,2003(10):46 - 49

(作者 E - mail:xjyang@njnu.edu.cn)