第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·软件技术与数据库· 文章编号：1000—3428(2008)13—0054—02 文献标识码：A 中图分类号：TP311.132

Web权威信息自动提取技术的研究及应用

李 净，袁小华，沈晓晶
(1. 上海水产大学信息学院，上海 200090；2. 同济大学电信学院，上海 201804)

摘 要：WWW为各行各业提供了大量的信息，但如何准确地从这些信息中提取出相关领域的权威信息是目前研究的热点问题之一。该文提出评判网站信息的多因素综合评估模型，该模型对网站的权威值进行合理计算，给出基于表格数据的语法树模型，完成了表格数据的自动提取。通过实例证明，该方法很好地解决了权威信息的准确和自动提取。

关键词：数据提取；Web数据挖掘；语法树；多因素综合评估；表格

Study and Application of Automation Extraction Technology from Web Authoritative Information

LI Jing, YUAN Xiao-hua, SHEN Xiao-jing
(1. School of Information, Shanghai Fishery University, Shanghai 200090; 2. School of Telecommunication, Tongji University, Shanghai 201804)

【Abstract】Although WWW has provided much information for all fields, how to extract the authoritative information from related fields exactly is becoming a hot topic. This paper provides a process of extracting table data it provides a multiple factors assessment model to judge the Web page. Using the model, the authoritative value of Web page can be gained correctly. It provides a table-based phrase tree method to extract the interesting data automatically. Example proves that this method can extract the authoritative information exactly and automatically.

【Key words】data extraction; Web data mining; phrasing tree; multiple factors assessment; table

1 概述

随着Internet的发展，网上信息正在呈指数级增长[1-2],在大量的数据中自动、准确地提取感兴趣的数据是非常必要的。但是在Web中自动、准确地提取用户感兴趣的数据非常困难，主要原因如下：

(1)难于确定权威网页。网上的信息来源很多，目前对于网页权威性的评价是根据网页URL的重要性来评价的[3-4]。比较常用的算法有PageRank, HITS, Kleinberg, SALSAD等。这些算法由于没有考虑网页的主题信息，因此由以上算法评价的权威网页却并不一定是主题相关的权威网页。

(2)Web 网页的自动提取困难。Web 页面特别是 HTML 页面(包括 ASP 或 PHP 等脚本编写的页面)，其设计无规律可循，用程序手段处理、分析比较困难；页面中的大多数内容描述是与数据驱动的系统无关的格式编排，并且由于要动态添加标题以及编写其他服务器端脚本，因此文档结构可能在每次连接到页面时都需要进行更改。

针对这些困难本文提出通过多因素综合评估法来确定权威网页，对权威页面的感兴趣数据(笔者认为感兴趣数据是以表格的形式展现的)通过基于语法树模型进行提取，给出了一个实例。

2 Web数据的提取

WWW 是个浩瀚的信息源，获取嵌入在 Web 页面，如 HTML, XML 或文本文件中的感兴趣的数据，重新将半结构化的数据组织为结构化数据是面向 Web 内容挖掘的第 1 步，也是一个基于 Web 数据挖掘工作成功与否的关键。为了使其获取的数据具有一定的权威性，提取过程(见图 1)包括：页面预处理，页面评估以及数据的提取 3 个过程。

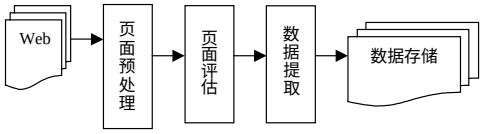

图 1 Web 信息获取的一般过程

(1)页面预处理

通过搜索引擎可以获得一个页面集合，这个集合包括权威信息源和一般信息页面。在评估这些页面之前首先需要将获得一组相关文档按照其所属网站进行分组，然后对每个组的文档进行清除、过滤页面冗余和噪声信息，最后对这些网页集合进行语义分析得到其行业信息。

预处理的步骤如下：

1)页面格式信息的规范、无表格页的删除和无用标签的删减[4]：对超文本数据中的各种脚本Scripts加以删除，去除无表格的网页(主要关心表格中的数据)，去除文件中的一些图片、视频等并对不规范的文本进行规范化。

2)抽取关键信息：在 Web 网页中包含了很多关键信息，例如标题、粗体字等。这些关键信息本身便表明其有特定的含义，需要使用有效的方法将其抽取出来。采用了简单匹配方法完成此任务。即要想抽取其中的标题信息，将标签<title>和</title>之间的内容直接抽取出来作为本网页的标题信息。

3)将网页按照网站分组存放。

基金项目：上海高校优秀青年教师科研专项基金资助项目
作者简介：李 净(1977－)，女，讲师、博士研究生，主研方向：数据仓库，数据挖掘；袁小华，副教授、博士；沈晓晶，讲师、博士
收稿日期：2007-08-20 **E-mail：**j_li@shfu.edu.cn

(2)页面评估

对于权威页面的评估通常采用网页的被访问频率来衡量，如基于 HUB 页面的 HITS 算法，目前的一些搜索引擎如著名的搜索引擎 Google 就是基于 HITS 算法的，这种方法的优点是系统会自动找到某些页面，使得系统本身具有一定的“智能”性，但这种方法得到的权威页面具有不稳定性，另外这种方法可能会产生大量“权威”页面，给系统造成巨大负担。本文采取多因素综合页面评估法，使用这种方法在一定程度上保证了权威页面的正确确定。另外为了保证权威网页的相对稳定性，需要对权威网页的评估频率进行控制。其评估步骤包括：1)评价因素的确定：评价因素是预先确定好的一般为网站 PageRank 值、点击率、主题相关率等。2)评价因素权值的确定。3)评价结果的产生。

(3)数据提取

目前基于 Web 的内容挖掘常常挖掘的是文本，但是实验证明人们关心的是数据，这步操作就是从相对稳定的权威页面中提取出感兴趣的数据，最后将这些提取出来的数据以 XML 的格式存放到数据库中，为以后的数据挖掘做准备。感兴趣数据的自动提取方法包括：1)构造网页语法树：认为感兴趣的数据是以 table 的形式体现的。将网页按<Table></Table>, <Tr></Tr>和<Td></Td>等标记展开，按照下面的算法构造一棵语法树。2)根据表格确定感兴趣的内容的数据，遍历该分枝的所有叶节点，获取数据。

3 关键技术及系统设计

(1)网页内容的获取方法

获取网页内容的方法有：1)使用相关的 OCX 控件，提供相关的入口参数 URL 即可获取页面的超文本内容。2)利用相关开发工具，如微软基础类库 MFC 中有 CinternetSession 类可获取网页；Java 中有 URL 和 URLConnect 类；Powerbuilder 中有 Inet 用户对象都支持获取网页文本的功能接口。

本文系统采用 Java 的 URLConnect 类来获取文档内容。

(2)多因素综合评估模型

多因素综合评价模型如下：

$$Q(P)=f(PR,T,U,\cdots)$$

其中，Q 是综合法加权得到的网页 P 的等级，各个因素的个数和每个因素的权值系统可以动态调整。得出的值越大，其权威性越高。

PR 为 Google 的 PageRank 所得的网页级别：

$$\text{PageRank}(A)= (1-d) + d(\text{PageRank }(T1)/C(T1) + \cdots + \text{PageRank }(Tn)/C(Tn))$$

其中，T 是网站专业领域的评价因子。T 的值通过简化文献[5]所提供的方法得到；U 是网站的超文本的点击率。

T 值得获取步骤：1)输入一组主题特征词($T1, T2,\cdots,Tn$)，以及这些词的权值($W1, W2,\cdots,Wn$)。2)分别检索网站中每个特征词 Ti，得到站内相关网页数 Pi 构成 n 维特征向量($P1, P2,\cdots,Pn$)。3)对($P1,P2,\cdots,Pn$)和($W1,W2,\cdots,Wn$)加权平均就得到 T 值。

(3)数据提取中语法树的构造

本文仅对表格的数据进行信息抽取，所以，设定“title(表头/标题)、table(表)、td(列)、tr(行)”为关键标记。其构造过程如下：1)建立起始头接点，表示一个文件。2)如果搜索到的当前关键标记与上一个标记相同时，则将该标记作为上一个标记的右孩子。3)如果搜索到当前关键标记，当此关键标记与上一个标记不同时，则回朔找到与当前标记相同的标记，如找到了把该标记作为回朔后的标记的右孩子；如果找不到就将该节点作为回朔前标记的左孩子。4)搜索所有关键标记直到结束。

系统中与之相关数据结构类有 2 个：PickupInfo 类(用于语法树的创建和数据提取)和 IBTree 类(存储语法树)。

IBTree 类：

```
{   private object data;          //该节点的数据
    private IBTree Lchild;       //左子树
    private IBTree Rchild;       //右子树
    Public String tag;           //节点的标志
    Public Object GetValue();    //获取当前节点的值
  Public Object SetValue(object data);   //设置当前节点的值
  Public IBTree ReturnLChild();    //获得左子树节点
  Public IBTree ReturnRChild();    //获得右子树节点   }
```

PickupInfo 类

```
{  Private IBTree ibTree;    //表示信息抽取用的二叉树
  Private String ibs 一“ title table td tr”; //关键标记
  public String CreateIBTree();    //语义树的创建
  public String PickupData();      //数据提取   }
```

4 应用实例

从权威网页中抽取水产品的价格：

(1)搜索关键字“农产品价格”，将得到的结果按照所属网站聚类，进行预处理。

(2)权威网站评估：综合因子有 PR,T,G(更新率)，其权值为 0.3,0.5,0.2，为了方便，选择中国水产网和温州农网为例。

输入主体特征词为：(农产品，农业)，权值为(0.6,0.4)。

程序计算得到：中国农网的 PR,U,T,G 为(7,1,0.9)，其权威等级为 2.78。

程序计算得到：温州农网的 PR,U,T,G 为(3,0.98,0.75)，其权威等级为 1.54。

(3)从权威网站：中国农业网进行数据提取，提取前的表格数据如表 1 所示。

表 1　提取前的数据

品种	价格/(元·kg^{-1})	日期
白鲢鱼	4.5	2007-05-10
百花鱼	68.0	2007-05-10
斑节虾	150.0	2007-05-10
草鱼	15.0	2007-05-10

(4)构造的语法树如图 2 所示。

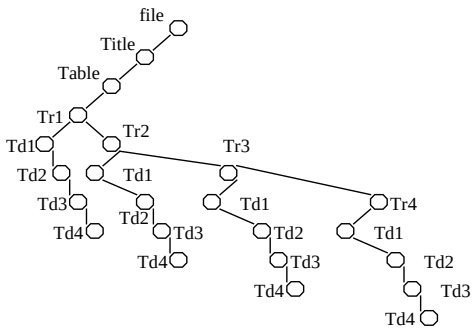

图 2　语法树

(5)抽取的结果为：

白鲢鱼	4.5	2007-05-10
百花鱼	68	2007-05-10
斑节虾	150	2007-05-10
草鱼	15	2007-05-10

（下转第 66 页）