Web 信息抽取及知识表示系统的研究与实现①

谭守标 徐 超 江 元 (安徽大学 电子科学与技术学院 安徽 合肥 230039)

宁仁霞 (黄山学院 电子信息工程系 安徽 黄山 245021)

摘 要： 研究了从数据密集型 Web 页面中自动提取结构化数据并形成知识表示系统的问题。基于知识数据库实现动态页面获取，进行预处理后转换为 XML 文档，采用基于 PAT-array 的模式发现算法自动发现重复模式，结合基于本体的关键词库自动识别页面数据显示结构模型，利用 XML 的对象-关系映射技术将数据存入知识数据库，由此实现 Web 数据自动抽取。同时，利用知识数据库已有知识从互联网抽取新知识，达到知识数据库的自扩展。以交通信息自动抽取及混合交通出行方案生成与表示系统进行的实验表明该系统具有高抽取准确率和良好的适应性。

关键词： Web 信息提取；知识表示；数据密集型 Web 页面；基于本体的关键词库

Research and Realization of a Web Information Extraction and Knowledge Presentation System

TAN Shou-Biao, XU Chao, JIANG Yuan (School of Electronic Science and Technology, Anhui University, Hefei 230039, China)

NING Ren-Xia(Electronic Information Engineering, Huangshan University, Huangshan 245021, China)

Abstract: The Web Information Extraction and Knowledge Presentation System is proposed to extract information from data intensive web pages. It downloads dynamic web pages, based on a knowledge database, changes them to XML documents after preprocessing, finds repeated patterns from them, by using a PAT-array based Pattern Discovery Algorithm, recognizes their data display structure models, automatically based on the repeated patterns and an ontology-based keyword library, and then extracts the data and stores them in the knowledge database with the object-relational mapping technology of XML. Through these steps, web data is extracted automatically, and the knowledge database is also expanded automatically. Experiments on the traffic information auto-extraction and mixed traffic travel schemes auto-creation system showed that the system has high precision and is adaptive to web pages in different domains with different structures.

Keywords: web information extraction; knowledge presentation; data intensive web pages; ontology-based keyword library

随着 Internet 的迅猛发展，Web 已经成为全球传播与共享科研、教育、商业和社会信息等最重要和最具潜力的巨大信息源。Web 信息抽取是指从 Web 页面所包含的无结构或半结构的信息中识别用户感兴趣的数据，并将其转化为结构和语义更为清晰的格式，以统一的形式集成在一起，使 Web 信息的再利用成为可能，成为当前研究的一个热点[1]。目前关于 Web 信息抽取的工作可以大致分为以下几个类别：基于特征模式匹配的信息抽取、基于归纳学习的信息抽取、基于网页结构特征分析的信息抽取、基于本体的 Web

① 基金项目:安徽省教育厅自然科学基金(2005KJ004ZD)
收稿时间:2010-01-06;收到修改稿时间:2010-02-26

信息抽取等。由于 Web 页面的种类繁多且信息抽取目的也不尽相同，不存在一种 Web 信息抽取系统，能够适应这种千变万化的应用环境。现有各种抽取方法针对不同领域、不同结构页面的通用性上也都存在一些问题[2-9]。

由于目前很多 Web 页面是动态生成的，以列表或表格的方式集中显示后台数据库中的数据，这种类型的页面对于数据集成等现实应用具有重要意义，抽取准确度也相对较高。本文针对于数据密集型的 Web 页面，开发出一种新的 Web 信息抽取和知识表示系统，通过基于 PAT-array 的模式发现算法[10]和基于本体的关键词库的结合大大提高了信息抽取算法的准确性和通用性，基于 Web 信息抽取的混合交通出行方案生成与表示系统的成功实验也证明了本文提出的 Web 信息抽取算法的实用性。

1　系统概述

本系统总体分成三部分：相关 Web 页面获取模块、Web 信息抽取模块、知识表示模块。系统总体框图如图 1 所示。

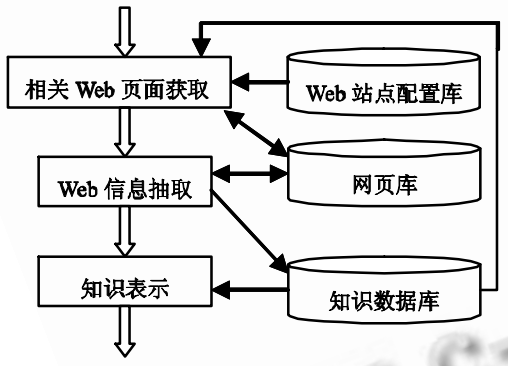

图 1　系统总体框图

相关 Web 页面获取模块：以知识数据库中现有知识为基础，根据 Web 站点配置信息生成动态 URL 从互联网上获取与所需知识相关的 Web 页面。

Web 信息抽取模块：采用基于 PAT-array 的模式发现算法发现数据密集型 Web 页面中的重复模式，结合基于本体的关键词库自动识别页面数据显示结构模型，利用 XML 的对象-关系映射技术，将数据存入知识数据库。

知识表示模块：以 B/S 架构提供知识表示服务，根据用户的输入从知识数据库中智能化搜索并生成用户需要的解决方案。

2　各模块的算法设计与实现

2.1 相关 Web 页面获取

数据密集型页面往往由 Web 站点根据用户的查询请求动态生成，从同一站点能得到大量同类型的动态页面。据此，系统以知识数据库为基础，采用 Web 站点配置方式，根据 Web 站点响应查询请求方式，人工配置含特定知识的 Web 站点信息及其动态页面 URL 生成规则。用知识数据库中现有知识作为查询参数，生成相关 Web 站点的动态 URL，通过 HTTP 协议自动获取相关 Web 页面。

如网站 www.cngoto.com 提供根据地名查询经过该地的所有列车车次信息，其响应查询请求的方式为 http://www.cngoto.com/tr/category91.asp?categoryid=$，此处$代表要查询的站名。系统从知识数据库的地点信息中检索得到各个地名，替换$即生成该网站动态 URL。

2.2 Web 信息抽取

本模块算法流程如图 2 所示：

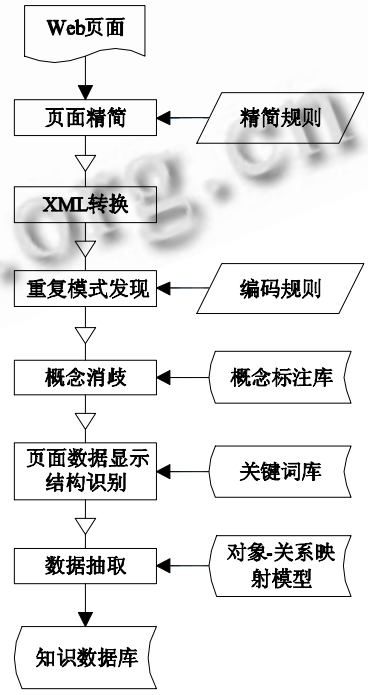

图 2　Web 信息抽取流程

2.2.1 页面精简

普通网页常常包含很多 Header 部页面属性信

息、脚本、样式、注释、图片、隐含数据、空格、标签属性设置及一些无用标签等，这些信息中不含有集中式数据，对造成后续处理速度缓慢，甚至使后续处理无法进行，需要首先进行页面精简，去掉这些冗余信息。本系统采取采用正则表达式技术进行如下页面精简操作：

① 清除 body 以外的部分；

② 清除文档中的脚本(<script 脚本内容</script>)、样式(<style 样式内容</style>)、注释(<!--注释内容-->)、隐含内容(<input type="hidden" 隐含内容>)、图片内容(<img 图片内容>)；

③ 清除文档中没有实际内容的标签对（只含空格、换行符等）（递归清除）；

④ 将连续多个“ ”和“ ”替换成一个空格“ ”；

⑤ 清除标签的属性信息。

2.2.2 XML 转换

由于 HTML 语法的随意性，即使经过页面精简，仍无法保证 HTML 文档的结构特性。而 XML 是一种结构化的自解释语言，更方便于进行重复模式发现，且在数据抽取过程中采用了 XML 的对象-关系映射技术，需要将 HTML 文档转换成 XML 文档。

本系统采用开源的 Jtidy 工具，实现 HTML 文档到 XML 文档的转换[11]。

2.2.3 重复模式发现

数据密集型 Web 页面的一个显著特点是数据显示区域（绝大部分情况是列表或表格形式）具有很强的重复模式，针对这一特点，可以通过重复模式的发现，很方便的确定页面数据显示区域的结构。

本系统采用基于PAT-array的算法实现快速的文档内重复模式的发现。具体步骤如下：

① 令牌翻译：对 HTML 中与数据显示相关的标签进行编码，将转换得到的 XML 文档翻译成二进制字符串；

② PAT 数组构造：罗列二进制字符串的所有半串（从每个编码到结束位置构成一个半串），按序排列后得到每个半串起始位置序号构成 PAT 数组；

③ 候选重复模式发现：使用栈操作，搜索得到所有半串的共同前缀即为候选重复模式；

④ 最佳重复模式确定：根据最优化标准从候选重复模式中确定出最佳重复模式。

2.2.4 概念消歧

单纯的重复模式发现算法只能得到笼统的数据显示结构，无法区分真正的数据及其语义（标题）。本系统采用基于本体的关键词库从重复模式中区分出标题项和数据项，最终确定准确的数据显示结构。

对于自然语言表示的 Web 文档，其中存在大量同义的词汇，在进行标题识别前需要进行概念消歧处理，利用概念标注库，将特定领域的同义词汇转换为关键词库中的本体词。

2.2.5 页面数据显示结构识别

本系统采用 XML 的对象-关系映射技术实现数据抽取，页面数据显示结构的识别即为 XML 文档对象模型(DOM)的确定。步骤如下：

① 标题定位：使用关键词库中特定领域的本体词集合，对页面中符合重复模式的数据进行搜索和定位，确定出其中的标题项；

② 标题-数据映射关系识别：根据确定出来的标题项集合的相对关系及与重复模式中其他数据项的相对关系，确定出各个标题项与数据项的映射关系；

③ DOM 树生成：根据重复模式及确定出的各个标题项与数据项的映射关系，生成对应的 DOM 树。

对于如下的 xml 文档：

```
<?xml version="1.0" encoding="GB2312"?>
<table>
  <tr>
    <td>车次</td>
    <td>1019</td>
    <td>运行时间</td>
    <td>19 小时 34 分</td>
  </tr>
  <tr>
    <td>始发站</td>
    <td>合肥</td>
    <td>终点站</td>
    <td>东莞东</td>
  </tr>
  ……
</table>
```

得到的 DOM 树结构如图 3 所示：

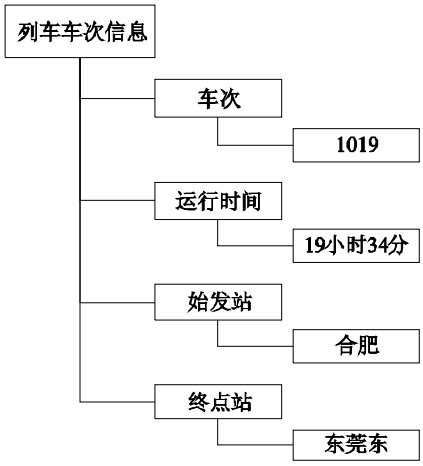

图 3　示例文档对应的 DOM 树结构

2.2.6 数据抽取

根据得到的 DOM 树进行数据模型映射，步骤如下：

① 利用 DOM 树生成对应的 XML 文档和 DTD 模式定义文档；

② 利用 DTD 进行对象-关系映射，将 XML 映射到知识数据库中，生成映射规则，进行数据抽取及存入数据库处理。

2.3 知识表示

采用 B/S 架构，利用数据库检索技术结合智能化方案生成算法，直接为用户提供解决方案，而不仅仅是知识的罗列。对解决方案给出一些评价标准，根据用户的选择按照评价标准对方案进行排序，使用户方便快捷的找到符合自己需求的方案。

3　实例及分析

本文以交通信息抽取及混合交通出行方案查询作为实例，通过相关 Web 页面获取及 Web 信息抽取模块从互联网上逐步抽取得到地点信息、站点信息、列车时刻信息、航班时刻信息、长途汽车客运时刻信息、各种交通票价信息等，实验中从配置的 15 个站点的约 30 万个动态页面中抽取相关数据，抽取准确率接近 100%。图 4 即是通过 Web 信息抽取得到的火车时刻信息。

开发了混合交通的出行方案生成系统，前台提供出行方案查询页面，可以指定多种交通工具和转车次数进行查询，按时间、金额、转车次数等进行排序显示。结果页面按序显示符合条件的各种出行方案，每条方案中全面给出从起点到终点的详细信息。

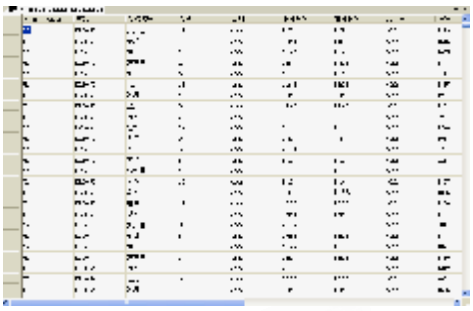

图 4　通过 Web 信息抽取得到的火车时刻信息

实验结果表明，本系统具有如下一些优点：

1) 以知识数据库作为支撑，通过配置网站库，能实现各种特定领域相关知识动态 Web 页面的自动下载；

2) 使用了基于本体的关键词库及概念标注库，使信息抽取能适应不同知识领域，适应无统一语义的 Web 环境。

4　结论

针对现有 Web 信息抽取方法对不同领域、不同结构 Web 页面的信息抽取缺乏通用性，本文提出了一种新的 Web 信息抽取和知识表示系统，实现不同知识领域下各种数据密集型动态 Web 页面的自动信息抽取，系统具有如下创新点：

1) 传统 PAT-array 算法无法区分重复模式区域的标题项和数据项，本系统采用基于本体的关键词库从重复模式中区分出标题和数据，自动识别数据显示结构模型和语义；

2) 将 Web 信息抽取和知识数据库有机结合起来，把知识数据库已有知识作为 Web 信息抽取的基础，从互联网上抽取新知识再存入知识数据库。从而达到知识数据库的不断自扩展。

实验表明该系统具有高抽取准确率和良好的适应性。下一步在页面数据显示结构模型自动识别中将利用基于本体的页面结构识别方法，提高具有复杂标题结构的重复模式结构识别能力。

参考文献

1 张岭.智能信息检索中的 Web 挖掘研究[博士学位论文].上海:上海交通大学, 2003.

(下转第 9 页)

(上接第 4 页)

2 Ana-Maria P. Information extraction from unstructured Web text [Ph.D Dissertation]. University of Washington, 2007.

3 李海健,王晓丰.Web 信息抽取的现状及未来展望.廊坊师范学院学报(自然科学版), 2009,9(3):39－40.

4 Wong TL, Wai L. An unsupervised method for joint information extraction and feature mining across different Web sites. Data and Knowledge Engineering, 2009,68(1):107－125.

5 韩存鸽,燕敏.Web 信息抽取方法研究.计算机系统应用, 2009,18(7):172－174,189.

6 Chang CH, Kayed M, Girgis MR, et al. A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10):1411－1428.

7 Gatterbauer W, Bohunsky P, Herzog P, Krupl B, Pollak B.Towards domain-independent information extraction from web tables. Proc. of the 16th international conference on World Wide Web, May.2007.71－80.

8 Crescenzi V, Mecca G. Automatic information extraction from large websites. Journal of the ACM, 2004,51(5):731－779.

9 邓尚民,孙玉伟.信息抽取系统的研究现状.现代图书情报技术, 2006,(3):54－58,81.

10 林建敏,谢康林.基于 PAT-array 和模糊聚类的文本聚类方法.计算机工程, 2004,30(12):126－127.

11 Jtidy 说明. [2008-11-21]. http://jtidy.sourceforge.net/.