表格信息抽取引擎的设计与实现

王治和

（西北师范大学数学与信息科学学院　兰州 730070）

摘　要　讨论针对 Web 表格的信息抽取，分析并给出了表格信息抽取引擎的系统结构，以及实现该系统所涉及的关键技术和数据模型，为用户提供一种以 Web 表格为信息抽取对象的、支持抽取方式选择的 Web 表格信息抽取工具。

关键词　Web 表格，数据挖掘，信息抽取，二叉树模型

The Design and Implementation of Information Extraction Engine on Web Tables

WANG Zhi-He　TIAN Hong

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070)

Abstract　This paper discusses a key technique aiming at information extraction on Web tables. It also analyzes and gives out the system structure of information extraction engine on Web tables, the key technique and data model implementing the system. The paper will provide a kind of information extraction tool which aims at Web tables and supports the method selection of extracting.

Keywords　Web table, Data mining, Information extraction, Binary tree model

1　引言

在 Internet 上进行信息检索时，经常会出现“信息过载”，即网上的信息是海量和无组织的，这就是所谓的“Rich Data Poor Information”问题。为解决这类问题，现在已经出现了很多的搜索引擎来帮助用户查找有用信息。尽管如此，查找的结果也是很庞大的。如何从这些文档中直接抽取所需信息，而不必一一浏览，是用户迫切需要的。

笔者研究了针对某一主题信息抽取的一些关键技术和实现方法，即“以表格为例的信息抽取引擎”的实现，其目的是提供一种以 Web 表格为信息抽取对象的、支持抽取方式选择的 Web 表格信息抽取工具。

2　系统构成

“表格信息抽取引擎”是针对信息抽取对象是 Web 表格的数据而开发的信息抽取工具，具有较高的通用性能，其系统构成如图 1 所示。需构建以下两个工具：

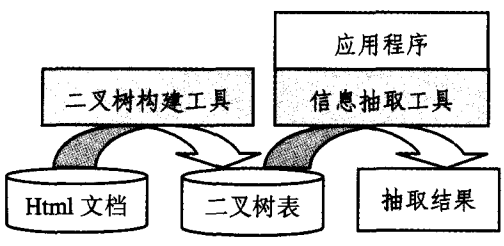

图 1　信息抽取引擎系统结构

① Html 文档分析工具

实现对全文信息的重新构建，包括 Html 分析和创建全文二叉树，即将一篇 Html 文档转化为一棵含有文本信息的二叉树，供信息抽取用。

② 信息抽取工具

利用全文二叉树进行查找、信息抽取，并具有选项抽取等功能，即从一篇 Web 文档的表格中提取出与用户感兴趣的关键词相关的表格信息。

二叉树构建工具以 Html 文档作为输入数据，将文档内的标记(tag)与文本(text)分开，将用户感兴趣的标记及其中的内容构造成一棵含有文本信息的二叉树。在此，仅对表格进行信息抽取，所以设定“title(表头/标题)、table(表)、td(列)、tr(行)”为感兴趣标记。例如下面的表可转化成图 2 所示的二叉树。

Data	Tem
Mon	31℃
Thu	28℃
Wed	33℃

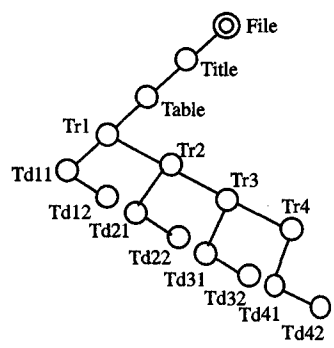

图 2　二叉树

二叉树的根结点为 file，file 的子结点为 title，title 的子结点为 table。二叉树的左结点为父结点的儿子，右结点实际为父结点的兄弟，即左结点嵌套于父结点，而右结点和父结点并列。

当 Html 文档转化成一棵二叉树后，还需提供用于二叉树的遍历和信息抽取的方法，之后就可使用信息抽取工具。

根据用户的要求，通过遍历这棵二叉树查找用户感兴趣的关键词，然后将该关键词结点所在的行、列或所在的子表格中的所有内容作为信息抽取结果进行输出。

3 关键技术

选用 Java 作为开发语言。关键技术为构建上述两个工具时相关的类、方法及函数的构造及编程。具体实现时，二叉树构建工具可包括 Htmlparser(Html 分析)类、TagNode(标记结点)类和 TagTree 类。

3.1 tmlparser 类

Htmlparser 类从一个 URL 地址或文件路径读出一篇 Html 文档，将文档内的标记(tag)与文本(text)分开，并把感兴趣的标记(tag)中的信息存储在二叉树的结点中。

3.2 TagNode 类

TagNode 类主要用来将 Html 标记的细节信息存储在结点对象中，为构建二叉树做准备。

3.3 TagTree 类

这个类的主要任务是把 TagNode 对象存储在二叉树结构中。在此类中有两类方法：一类是建树用的私有方法，不能被应用程序调用；另一类是协助应用程序从二叉树中抽取信息的公有方法。

信息抽取工具将根据用户的要求调用二叉树的不同方法提取符合要求的文档信息，该模块由 ExtractData(抽取数据)类组成。

构造以上类时需定义相关方法，并说明各方法的功能及方法中所使用到的参数及变量。

4 构建二叉树模型

在信息抽取中之所以没有把 Html 标记删除，而是记录在二叉树中，是因为信息抽取不仅需要找出关键词，同时需要找出与这个关键词位置相关的所有位置的文本内容。因此在抽取相关内容时，还需要利用 Html 的结构，不能将其单纯转化为字符流。在目标表示的时候也需要用到 Html 标记和文本文档。

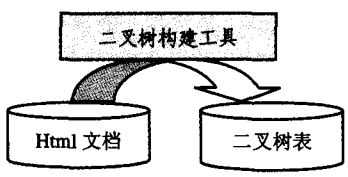

图 3 二叉树构建工具功能图

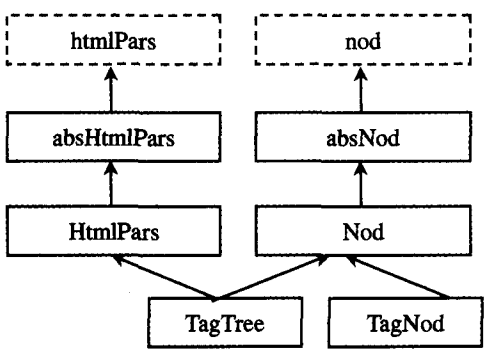

图 4 类的继承关系图

二叉树构建工具的功能是将一篇 Html 文档转化为一棵含有文本信息的二叉树，供信息抽取用，如图 3 所示。

具体实现时，将二叉树构建工具分为 Htmlparser(Html 分析)类，TagNode(标记结点)类和 TagTree 类。类的继承关系如图 4 所示。

5 信息抽取工具

完成二叉树的建立，用于二叉树的遍历和信息抽取，该工具是由 extractData 类组成的。

5.1 extractData 类

5.1.1 类变量

① Private TagTree tagTree：表示信息抽取用的二叉树。

② Private PrintWriter errStream：表示出错信息流。

③ Private Vector finalTextVec：储存找到的信息。

④ Private final String TAGS = " title table td tr "：感兴趣的 tag 标记。

⑤ rivate final int Dispsize：数据的最大长度。

5.1.2 类方法

Private String parseWeather(String tblStr, String keyStr)

(1)功能

ParseWeather 方法用来抽取表格中与关键字相关的信息。

(2)变量及参数

① 参数

a. String tblStr：抽取模式，即抽取一行、一列或是一个表格。

b. String keyStr：信息抽取的关键字。

② 变量略。

5.2 信息抽取示例

下面就一个有关天气预报的 Web 文档进行信息抽取，来说明整个信息抽取的过程。首先来看这个天气预报的 Web 文档(见图 5)，此文档中包含了一个极其复杂的嵌套表格，从中抽取需要的天气预报信息。

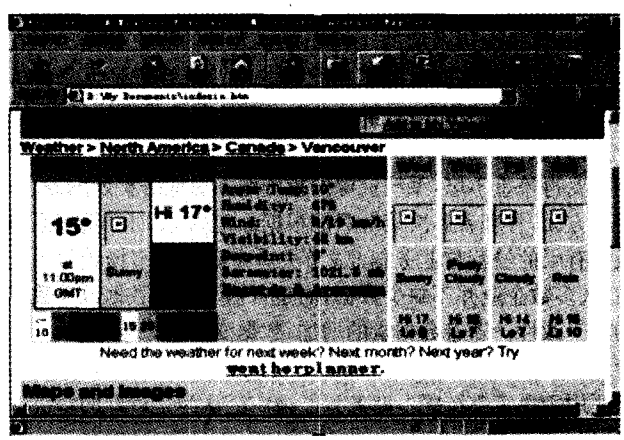

图 5 有关天气预报的表格网页

就图 5 所示的天气预报网页，给出信息抽取文档地址、关键词和模式，得出抽取结果。

当所给参数为：

文件路径：D:\my documents\indexie. htm
抽取方式：List
关键词 ：Wed

(下转第 175 页)

有 Article 的概念上下文：

$$Ctx(Article)=\left\{\begin{array}{l}(Title,property)\\(Author,property)\\(Pages,property)\\(Publication,overlays)\\(Document,\ includes)\\(Paper,similar\text{-}to)\end{array}\right\}$$

Book 的概念上下文：

$$Ctx(Book)=\left\{\begin{array}{l}(Title,property),(Publisher,property),\\(Library,includes),(Chapter,part\text{-}of)\end{array}\right\}$$

Article 与 Book 概念上下文的名称相似度和语义关系相似度如表 2。

表 2 Article 与 Book 概念上下文的名称相似度和语义关系相似度

LA()	Title	Author	Pages	Publication	Document	Paper
Title	1.0	0.25	0.25	0.4	0.4	0.4
Publisher	0.25	0.0	0.25	0.5	0.5	0.5
Library	0.25	0.25	0.0	0.5	0.5	0.5
Chapter	0.25	0.25	0.25	0.64	0.8	0.5

RA()	property	property	property	overlays	includes	similar-to
property	1.0	1.0	1.0	0.3	0.5	0.8
property	1.0	1.0	1.0	0.3	0.5	0.8
includes	0.5	0.5	0.5	0.8	1.0	0.7
part-of	0.7	0.7	0.7	0.6	0.8	0.9

那么，有计算结果：

$LA(Article,Book)=0.5$；

$CA(Article,Book)=0.553$。

$SA(Article,Book)=0.5\cdot 0.5+0.5530.5=0.5265$。

$\widetilde{A}_1(0.5265)=0$，$\widetilde{A}_2(0.5265)=0.9649$，$\widetilde{A}_3(0.5265)=0.0351$。

可知，Article 与 Book 为语义次关联。

用户在资源请求时可以指定资源匹配的类型{弱关联，次关联，强关联}。假设要求返回与资源请求目标 Article 强关联的概念，Book 就不满足条件。用户可以根据反馈的结果调整资源匹配的类型，扩大或缩小查询结果范围。本例中如果将匹配类型改成次关联，查询将返回 Book 相应的数据。

结束语 资源匹配是资源发现中的重要步骤。为了能够根据组织单元提供的信息更加准确地描述并执行资源发现，考虑更加丰富的语义和上下文信息，本文结合概念名称和概念上下文对语义相似度的影响，提出了一种语义模糊匹配算法。通过在资源的描述和资源的匹配中使用语义信息，消除了语义异构，从而保障用户获取语义上相关联的、更多的数据。同传统的资源匹配比较，本方法可以更加准确地定位资源。同时，引入模糊集的理论，实现了一定程度的资源模糊匹配，可以有效地辅助用户进行选择和决策，为资源发现提供了新的解决方法。未来的工作将研究更加通用的匹配方法。

参 考 文 献

1 Mowshowitz A. Virtual Organization. Communication of the ACM, 1997, 40(9): 30～37
2 赵纯均，陈剑，冯蔚东. 虚拟企业及其构建研究. 系统工程理论与实践，2002，22(10)：49～55
3 董方鹏，龚奕利，李伟，等. 网格环境中资源发现机制的研究. 计算机研究与发展，2003，40(12)：1749～1755
4 Raman R, Livny M, Solomon M. Matchmaking Distributed Resource Management for High throughput Computing. In: Proc. of the Seventh IEEE Intl. Symposium on High Performance Distributed Computing, Chicago, IL, July 1998
5 The portable batch system. http://pbs.mrj.com
6 Liu C, Foster I T. A Constraint Language Approach to Matchmaking. RIDE, 2004. 7～14
7 Tangmunarunkit H, Decker S, Kesselman C. Ontology-Based Resource Matching in the Grid-The Grid Meets the Semantic Web. In: International Semantic Web Conference, 2003. 706～721
8 Chakraborty D, Perich F, Avancha S, et al. Dreggie. Semantic Service Discovery for M－Commerce Applications. In: Proc. of the 20th Symposiom on Reliable Distributed Systems, Workshop on Reliable and Secure Applications in Mobile Environment, 2001
9 Chakraborty D, Perich F, Avancha S, et al. An Algorithm for Matching Contextualized Schemas via SAT: [Technical report]. DIT－03－003. DIT University of Trento, Italy, January 2003. Available at: heep://prints.biblio.unitn.it/archive/00000348/
10 Serafini L, Bouquet P, Magnini B, et al. Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions. ITCC (1), 2004. 360～366
11 Zhang X, Freschl J, Schopf J. A Performance Study of Monitoring and Information Services for Distributed Systems. In: Proceedings of IEEE HPDC-12, 2003
12 Keung H, Dyson J, Jarvis S. Predicting the Performance of Globus Monitoring and Discovery Service. In: Proceedings of 4th IEEE/ACM International Workshop on Grid Computing, 2003
13 Miller A G. WordNet: A Lexical Database for English. Communications of the ACM, 1995, 38(11): 39～41

(上接第 127 页)

时，得到的抽取结果为：

```
Wed Sunny Hi 17 Lo 8
```

当所给参数为：

```
文件路径:http://202.117.80.2
抽取方式:Table
关键词 :Wed
```

时(注：同一网页)，得到的抽取结果为：

```
Wed Sunny Hi 17 Lo 8
Thu Partly Cloudy Hi 16 Lo 7
Fri Cloudy Hi 14 Lo 7
Sat Rain Hi 16 Lo 10
```

结束语 本文利用二叉树模型实现了针对表格的信息抽取引擎的开发。该工具具有以下的特征：检索对象是 Web 表格；适用于结构化 Web 文档；高效、通用的信息抽取工具；提供多种信息抽取方式；支持一定的网页容错功能，具有可扩充性。

参 考 文 献

1 田志良，王世普，张皓东，等. 国际网企业网和智能建筑. 昆明：云南大学出版社，1997. 177～215
2 李大友，邱建霞. 计算机网络. 北京：清华大学出版社，1998. 151～166
3 杨明福. 计算机网络. 北京：电子工业出版社，1999. 123～127
4 蔡皖东，陈亚滨. 计算机网络培训教程. 西安：西安电子科技大学出版社，1999. 62～72