ISSN1000-9825 Journal of Software 软件学报 2000,11(1):73~78

从WEB文档中构造半结构化信息的抽取器*

黄豫清 戚广志 张福炎

(南京大学多媒体计算机研究所 南京 210093)

摘要 为了对WEB上不规则的、动态的信息按照数据库的方式集成和查询,采用对象交换模型(object exchange model,简称OEM)建立了WEB信息模型.为了将页面中各个部分表示为对应的OEM对象,设计了半结构化信息的抽取算法,并给出测试结果.该方法可以抽取结构化和半结构化的信息,比现有的抽取方法通用性更强.

关键词 启发式规则,数据抽取格式,对象交换模型.

中图法分类号 TP311

WWW是快速增长的巨大信息库,含有大量有用的信息,其信息存储为静态HTML页面,主要通过浏览器来查看.虽然我们可以利用现有的众多搜索引擎进行有效信息的查询,但是查询结果往往是互不相关的HTML页面,要直接查询页面上精确的信息十分困难.

从数据库的观点看,WWW上的大量资源包含半结构化的信息,按照某种格式显示.考虑AAAI会议资源,它提供了分段的会议论文,包括作者、论文名称等信息.AAAI-1997的页面如图1所示,AAAI-97 Technical Papers,Invited Talks等等是各段中可以识别的标题,对于基于页面结构的查询非常有用.例如,查询"Find Kentaro Toyama's Papers",应该返回页面中Kentaro Toyama的论文名称.我们可以通过抽取页面上的结构信息,经过包装器来提供这样的查询能力.但是,由于信息的数目巨大,网上新资源频繁加入,现存资源的格式经常变动,对WEB文档结构信息手工抽取是不实际的.文献[1]采用启发式方法,按照各个部分字体的大小和缩进距离推导出页面上的层次结构.该方法对于没有标出字体的大小和缩进距离的部分无法抽取,不能处理列表和表格,而且启发式方法也会产生错误,需要用户手工校正系统输出的结构.文献[2]采用用户输入页面描述文件对层次结构进行抽取,该描述文件需要用户描述抽取过程的具体变量和编写抽取方法,只能适用于某种特殊的页面.文献[3]讨论了WWW资源的抽取器,但主要考虑出售商品的页面,需要对搜索的信息类型进行很强的假设,只能抽取关系类型的数据,该方法不能用于更一般类型的页面.

AAAI-97 Technical Papers

Agents

Agent Architecture

- Kentaro Toyama, Gregory D. Hager: If at First You Don't Succeed...
- Juan D. Velasquez: Modeling Emotions and Other Motivations in Synt

Fig.1 Fragments of AAAI conference homepage

图1 AAAI会议主页的片断

当前各种抽取方法存在的问题是:(1)不能将所有数据都抽取出来,(2)不同的页面需要编写不同的抽取器.为了将用户需要的所有数据都抽取出来,并且构造出适用于各种页面格式的抽取算法,我们设计了HTML文本中各个数据片断的抽取算法,该算法采用用户指定的数据抽取格式作为输入,并结合启发式规则进行抽取处理.本文提出的抽取算法立足于不规则的半结构化数据,对各种页面中的结构化数据和半结构化数据都统一

* 作者黄豫清,女,1970年生,博士,讲师,主要研究领域为数据库系统,计算机辅助教学.戚广志,1970年生,博士生,主要研究领域为信息集成系统,数据库系统.张福炎,1939年生,教授,博士生导师,主要研究领域为计算机图形学,多媒体技术.

本文通讯联系人:黄豫清,南京 210093,南京大学多媒体计算机研究所

本文1998-11-17收到原稿,1999-02-12收到修改稿

处理,通用性强.

1 结构信息的抽取

结构信息抽取过程就是将一组超链连接的HTML页面转换为嵌套的数据对象,抽取器将所需数据的网页地址和该数据抽取格式的描述作为输入.如果HTML页面的格式改变,数据抽取格式的描述必须被更新.因为数据抽取格式描述是一个简单的文本文件,它可以用任何编辑器直接修改.由于在WWW上的主要信息是半结构化的[4],不像传统数据库中的数据那样具有规则的和静态的结构,而斯坦福大学提出的OEM模型非常适于描述半结构化数据,因此抽取器的输出是OEM格式的数据.OEM是一个无模式的模型,特别适合表示网上的半结构化数据.OEM中表示的数据组成一个图,在顶部具有唯一的根对象,具有零个或多个嵌套子对象.各个OEM对象包含1个标记、1个类型和1个值.该标记描述了存储在组成部分中的值.存储在OEM对象中的值可以是原子性的(例如类型string),也可以是一组OEM对象.

1.1 识别段标题

用文本表示的段开始的标记常常在HTML中表示为粗体字或者大尺寸的字符、列表段以及表格段.在表1中列出的段表达式用于识别一个页的各段标题,与段表达式形式一致的单词或短语作为段标题被系统保存起来.因为各个段标题代表页面中一个段的开始,在上述段标题识别结束时,所有不同的段就被标识出来.

Table 1 A list of section expression which identify section titles in pages

表1 识别页面上段标题的段表达式列表

Section expression①	Description②	Examples③
⟨h1-6⟩#⟨/h1-6⟩	Title font section④	⟨h2⟩Invited Papers⟨/h2⟩
⟨b⟩#⟨/b⟩	Physical bold-faced section⑤	⟨b⟩Innovation in Database Management: Computer Science ... ⟨/b⟩
⟨strong⟩#⟨strong⟩	Logical bold-faced section⑥	⟨strong⟩Area⟨strong⟩
⟨fontsize=1-7⟩#⟨/font⟩	Section with font size⑦	⟨font size=7⟩good weather⟨/font⟩
⟨ul⟩⟨li⟩#⟨/ul⟩	List section without order⑧	⟨ul⟩⟨li⟩⟨aname="Jacobs97" href="../../indices/a trec/j/...")⟨/a⟩;..⟨/ul⟩
⟨ol⟩⟨li⟩#⟨/ol⟩	List section with order⑨	⟨ol⟩⟨li⟩Today⟨li⟩Tomorrow⟨/ol⟩
⟨table#⟩⟨tr⟩#⟨th⟩#⟨td⟩#⟨/table⟩	Table section⑩	⟨table⟩⟨tr⟩⟨th⟩Food⟨/th⟩⟨tr⟩⟨td⟩A⟨/td⟩⟨td⟩B⟨/td⟩⟨/table⟩

①段表达式,②描述,③例子,④标题字体段,⑤物理粗体字的段,⑥逻辑粗体字的段,⑦有字体大小的段,⑧无序列表段,⑨有序列表段,⑩表格段.

1.2 确定页面初步层次结构

下一步是获取页面中各个段和其他段之间的初步层次关系.为了确定上面各标题段之间的层次,需要以下启发式规则:(1) 子段标题字体一般比所属段的标题字体更小;(2) 向右缩排的段常常是另一个段的子段;(3) 无序列表段、有序列表段和表格段中各元素互为兄弟.由此,根据各个段表达式捕获与之类型相同的段标题,为各个段标题构造新节点,根据标题的字体大小、缩排距离使新节点成为另一个段的亲子节点.同一个列表段各个节点互为兄弟节点.各个节点的亲子是排序的,出现的次序和页面上对应段的次序相同.

但是当完全依靠上面的启发式规则识别一个新页结构时,可能选择出和用户意图不一致的段标题,因为段表达式和启发式规则不能覆盖所有的情况.系统允许用户通过修改数据抽取格式文件来添加系统丢失的段或者删除系统错误选择的段,将启发式规则和数据抽取格式结合起来,可以在自动化和准确性之间得到最好的平衡.

1.3 数据抽取格式的描述

数据抽取格式的描述,就是对感兴趣的页面中的标记的描述.特定的标记表示其中嵌入的单词或短语是一个标题,一个标题的出现就意味着一个新段的开始,因此,识别标题就是识别一页中的段.一页可以分解为不同的段,一段中又包含子段,我们需要识别段内的嵌套结构.例如,AAAI页面上的AAAI-97 Technical Papers,Invited Talks等组成.AAAI-97 Technical Papers又由子段Agents,Automated Reasoning等组成.

一个数据抽取格式是由 N 行形式为{段文本,段变量,资源}的字符串组成.其中,段文本表示从资源中反复抽取满足要求的文本作为新创建的 OEM 对象的值;段变量作为被创建的 OEM 对象的标记;资源取值为一个页面地址或者是前面抽取格式行中出现的段变量,表示用于抽取文本的资源是该页面的 HTML 文本或者是当前父对象的值.第 1 行创建一个根对象,后面的行创建父对象标记为资源的、本行对象标记为段变量的、取值由段文本限定的所有 OEM 亲子对象.用户对于一个页面中感兴趣的内容通常限制在一个范围内,因此,一个页面的数据抽取格式的第 1 行通常说明了该页面中有用部分的开始到结束的字符标记、存放内容的 OEM 对象的标记和该页面的网上地址.下例说明从 AAAI-97 页面的内容中创建 OEM 对象的数据抽取格式.

例:1{〈h2〉#DBLP:,root,get("http://SunSite.Informatik.RWTH-Aachen.DE/.../AAAI-97.html")}

2{〈/hx〉#〈hx〉,〈hx〉#〈/hx〉,root}

3{〈li〉#〈li〉,Section,〈hx〉#〈/hx〉}

行 1 从给定 URL 的资源文件的内容中获取出现在标记〈h2〉和 DBLP 之间的文本,放入变量 *root_var* 中.它用来限定要抽取的范围,并且生成根节点.

行 2 将给定 URL 资源文件中第 1 次出现〈h1-6〉之后的文本抽取出来,将〈hx〉到〈/hx〉之间的文本作为新的段变量名称,将〈/hx〉到〈hx〉之间的文本,即从当前位置到下一个兄弟段之间的文本作为该段变量的值,x 表示 1～6 之间的取值.这里,运用启发式规则,按照字体大小生成该页面的初步层次结构,对于 AAAI 的页面来说,存在〈h2〉、〈h3〉和〈h4〉这 3 种类型的字体,所以在根节点下生成 3 层亲子节点.

行 3 从行 2 创建的段变量值中抽取〈li〉到〈/li〉之间的文本作为变量 *Section_var* 的值,构造复杂对象 Section,作为已经产生的亲子节点的亲子.

图 1 所示页面的 HTML 文本片断如图 2 所示.根据数据格式描述行进行段变量的创建和赋值,结合启发式规则,确定各个段之间的层次结构,该页面最多有 6 层结构,即 root→〈h2〉#〈/h2〉→〈h3〉#〈/h3〉→〈h4〉#〈/h4〉→Section→{author,Title}.该数据被包装到一个 OEM 对象中,如图 3 所示.

```
〈html〉AAAI 97, IAAI 97, July 27－31, 1997...
〈h2〉AAAI-97 Technical Papers〈/h2〉
〈h3〉Agents〈/h3〉
〈h4〉Agent Architecture〈/h4〉
〈ul〉
〈li〉〈a name="ToyamaH97"href="../indices/a-tree
/t/Toyama:Kentaro.html"〉Kentaro Toyama〈/a〉,
〈a bref="../../indices/a-tree/h/Hager:
Gregory_D=.html"〉Gregory D. Hager〈/a〉:
If at First You Don't Succeed...
DBLP:...〈/body〉〈/html〉
```

Fig. 2 HTML fragments of AAAI conference homepage

图 2 AAAI 会议主页的 HTML 文本片断

```
Root complex{
AAAI-97 Technical Papers complex{
Agents complex{
Agent Architecture complex{
Section complex{
Author set{Kentaro T., Gregory D. H.}
Title string{If at First You Don't Succeed..}
                }
    ...         }
    ...     }
    ...                 }
    ...     }
```

Fig. 3 OEM objects extracted from AAAI homepage

图 3 AAAI 页面抽取出的 OEM 对象

1.4 结合数据抽取格式描述和启发式规则的算法

输入:数据抽取格式 $SECTXT_1$,$SECVAR_1$,$SOURCE_1$,...,$SECTXT_M$,$SECVAR_M$,$SOURCE_M$

$SECTXT_i$,$SECVAR_i$,$SOURCE_i$ 分别代表第 i 行的段文本,段变量,资源

输出:页面的 OEM 数据 OEMdata

OEMdata EXTRACTOR($SECTXT_1$,$SECVAR_1$,$SOURCE_1$,...,$SECTXT_M$,$SECVAR_M$,$SOURCE_M$)

按照 $SOURCE_1$ 的网络地址获取 HTML 文本

抽取 $SECTXT_1$ 中描述的左右分割符之间的文本段作为根对象值

创建标记为 $SECVAR_1$ 的根对象 R_0

将根对象 $SECVAR_1$ 名称和值 Value 记录在已处理表中

While(数据抽取格式描述行未处理完)

```
{将已处理表中第 1 个记录作为当前项
While (在当前项中存在和 SOURCEm 一致的 SECVARm 对应的值 Value)
    {将 Value 作为资源
    如果SECTXTm 不出现字体大小标记、缩进标记组合
        {按照 SECTXTm 从资源中获取各文本段作为标记为 SECVARm 的各个对象值
          如果 SECVARm 为常数字符串,所有子对象标记都是 SECVARm
          否则按照 SECVARm 从 Value 值中获取各个对象标记
              生成对象作为标记为 SECVARm 对象的亲子
              将这些亲子的 SECVARm 和值 Value 记录在已处理表中
        }
    如果SECTXTm 出现字体大小标记的组合,用启发式确定层次结构
        {在资源中获取所有尺寸的字体标记,从大到小记录在字体列表中
        将字体列表中首记录作为当前字体记录
        While (当前字体记录不为空)
            {按照当前字体标记从资源中获取各文本段作为各个对象值
            按照 SECVARm 从资源中获取各个对象标记
            生成对象作为标记为 SECVARm 对象的亲子
              将这些亲子的 SECVARm 和值 Value 记录在已处理表中
              将所有 Valuc 作为资源列表
              字体列表记录下移一个记录作为当前字体记录
            }
        }
    如果SECTXTm 出现缩进标记的组合,用启发式确定层次结构
        {在资源中获取所有尺寸的缩进标记,从小到大记录在缩进列表中
        将缩进列表中首记录作为当前缩进记录
        While (当前缩进记录不为空)
            {按照当前缩进标记从资源中获取各文本段作为各个对象值
            按照 SECVARm 从资源中获取各个对象标记
            生成对象作为标记为 SECVARm 对象的亲子
              将这些亲子的 SECVARm 和值 Value 记录在已处理表中
              将所有 Value 作为资源列表
              缩进列表记录下移一个记录作为当前缩进记录
            }
        }
    已处理表下移一个记录作为当前项
    }
    读取下一个数据格式描述行
}
Return OEMdata
```

该抽取算法将启发式规则纳入数据抽取格式中,即当抽取格式段文本 $SECTXT_m$ 中出现字体大小标记的组合、缩进标记的组合时,用启发式确定层次结构.对于出现字体大小标记的情况,找出资源中所有字体尺寸,并且从大到小排列,对于各个尺寸字体依次生成各层对象,对于相邻的两种字体,较大字体生成的对象值作为较小字体生成对象时的资源.对于出现缩进标记的情况,找出资源中所有缩进距离,并且从小到大排列,对于各个缩进

距离依次生成各层对象,对于相邻的两种缩进距离,较小缩进距离生成的对象值作为较大缩进距离生成对象时的资源.

2 与相关工作的比较

我们使用抽取器将几个网上 HTML 页面表示为 OEM 对象,并且评价遵从约束的抽取格式描述的抽取器效果(见表 2).结构相似的资源抽取格式基本相同,例如,VLDB 会议、SIGMOD 会议、PODS 会议和 AAAI 会议.但是,对于结构不相似的资源,例如,上海交通大学研究项目、《中国青年报》和《人民日报》抽取格式的差异很大.我们考虑将资源按照结构的相似性分类,同一类的资源使用一种标准抽取格式,而《中国青年报》和《人民日报》使用另一种标准的抽取格式,只要对一类标准抽取格式稍作修改,就可以用于该类其他资源的抽取格式.另外,抽取时间不仅和文件本身的大小有关,而且和页面的层次结构有关,它随着页面层次的增加而增加.文献[1]的启发式方法,对没有标出字体大小和缩进距离的部分无法抽取,不能处理列表和表格,对于表 2 中各个页面内的对象大部分无法抽取出来.本文和文献[2]的区别在于,本文将抽取的方法和 HTML 页面格式的描述分开,用户只需要描述 HTML 的格式,具体的抽取方法由统一的抽取器完成;而文献[2]需要用户将页面格式和抽取方法都具体给出,这不仅加重了用户的负担,而且其抽取器不具有通用性.文献[3]只能抽取表格类型的数据,无法对具有复杂嵌套层次的对象进行抽取,对于表 2 中的各个对象则无法完整地抽取出来.因此,本文提出的抽取算法可以将用户需要的数据完全抽取出来,并且适用于各种页面格式,通用性强.

Table 2 Time and cost for extracting various WEB sources

表 2 抽取不同 WEB 文档资源需要的时间和代价

WEB sources①	Extracting time(ms)②	Extracting format lines③	Deepest level of OEM④	File size (K)⑤
VLDB Conferences⑥	765	7	5	28.6
SIGMOD Conferences⑦	590	7	4	35.7
PODS Conferences⑧	172	7	4	13.5
AAAI Conferences⑨	1982	6	6	60.4
Research of SHJD⑩	250	6	3	16.7
Info. Center of CAS⑪	125	4	5	12.4
China Youth Daily⑫	32	5	3	13.0
People's Daily⑬	62	6	3	20.0

①WEB 资源,②抽取时间,③抽取格式行数,④OEM 最大层次,⑤文件大小,⑥VLDB 会议,⑦SIGMOD 会议,⑧PODS 会议,⑨AAAI 会议,⑩上海交通大学的研究项目,⑪中国科学院文献情报中心,⑫中国青年报,⑬人民日报.

3 将来的工作

在识别一个新页的层次结构时,系统可能不能识别页面上的段标记,需要在系统中加入从用户例子中学习的新类型标记的结构.另外,对于抽取格式的描述依赖于外部输入,当资源文件改变时,需要对描述文件进行更新,抽取程序中需要插入机器学习技术,自动猜测网点上下层 HTML 结构.

参考文献

1 Ashish A, Knoblock C. Wrapper generation for semi-structured Internet sources. SIGMOD Record, 1997,26(4):8~15

2 Hammar J, Garcia-Molina H, Cho J *et al*. Extracting semi-structured information from the Web. SIGMOD Record, 1997, 26(2):18~25

3 Kushmerick N, Weld D S, Doorenbos R. Wrapper induction for information extraction. In: American Association for Artificial Intelligence ed. International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1997. 729~735

4 Chawathe S, Garcia-Molina H, Hammer J *et al*. The TSIMMIS project: integration of heterogeneous information sources. In: Ashish G *et al* eds. Proceedings of the 10th Anniversary Meeting of the Information Processing Society of Japan. San

Francisco, CA: Morgan Kaufmann Publishers, Inc., 1994. 7～18

Extracting Semi-Structured Information from the WEB

HUANG Yu-qing QI Guang-zhi ZHANG Fu-yan

(*Multimedia Computer Institute Nanjing University Nanjing* 210093)

Abstract In order to integrate and query irregular and dynamic information on WEB in a database-like fashion, the authors use object exchange model (OEM) to construct information model of WEB in this paper. To express each component of pages as an OEM object, the authors design an algorithm which extracts semi-structured data from HTML pages, and the testing results are given. This method can extract structured and semi-structured data. It has better applicability than other existing methods.

Key words Heuristics rule, data extracting format, object exchange model.