网际网 数据抽取 知识发现 半结构化数据 (12)

计算机科学1999Vol.26№.10

49-52

网际网上半结构化数据抽取与知识发现方法及其实现*)

On Semi-structured Data Extraction from WWW and Its Schema Knowledge Discovery Method

陈恩红 范 焱 王行甫 蔡庆生

(中国科学技术大学计算机系 合肥230027)

TP393 TP18

Abstract It is well known that World Wide Web has become a huge information resource. However, the information on WWW can not be queried and manipulated in a general way. Large amount of information is stored in a static HTML format and can only be viewed through browser. Therefore, it is very important for us to utilize this kind of information effectively. This paper proposes a semi-structured data extraction method to get the useful information embedded in a group of relevant web pages, and store it with OEM(Object Exchange Model). Then, we adopt data mining method to discover schema knowledge implicit in the semi-structured data.

Keywords Semi-structured data, Knowledge discovery

1. 引言

在信息化程度日益提高的今天,半结构化信息已遍及社会的各个领域。例如,网际网(World Wide Web,又称WWW)已成为一个巨大的信息源,然而WWW上的信息并不能以一种通用的方式进行查询及操纵,大量的信息是以静态的HTML文本形式存储并只能通过浏览器来浏览,因此如何有效利用这类信息显得尤为重要[1]。虽然某些站点或许也提供了一些搜索引擎,但一般是通过关键词匹配来进行其查询,且查询结果仍是HTML文本。有关的信息还须通过浏览器到相应的Web站点去浏览,用户难以获得关于整个网站的信息结构。

基于上述问题,本文首先实现一种数据抽取方法,将相关的一组网页中的有用信息提取出来。这类信息的特点是无任何预定义的结构,通常又被称为半结构化数据[2]。这种类型数据广泛存在于电子图书馆,在线文档,电子商务的众多应用领域[3]。由于它难以用传统的关系模型表示出来,因此本文采用OEM(Object Exchange Model)[4]模型对其进行组织存储。考虑到网上信息量往往较大,我们的半结构化数据抽取的实现是在线处理方式,即系统根据要访问的服务器地址及HTML文件的存取路径,将文件取回抽取信息,然后根据信息抽取过程中得到的链接,再去取文件进行信息抽取。最后再用数据挖掘方法[5-8]进行模式知识发现。

2. OEM模型

半结构化数据的特点是缺少任何固定或严格的模式,因此对它的研究显得更吸引人。下面将介绍本文采用的OEM模型。在OEM中,每个对象由标识符和值组成,标识符唯一标识了对象,对象 id 的值表示为 $f(id)$。$f(id)$可以是原子值(如整数,字符串),也可以是一个对象包(object bag),如$\{l_1:id_1, \ldots, l_k:id_k\}$,标记 l_i 是描述对象 id 与其子对象 id_i 间关系的字符串,子对象的类型可以不同,具有原子值的对象称为原子对象,由对象包构成的对象称为复杂对象。以OEM表示的对象可以用一个带标记的有向图来表示,图中的节点表示对象,有向边上的标记表示对象间的关系(即语义信息),出度为0的节点表示原子对象,并与某个值关联。其它节点为复杂对象。

图1是从 http://us.imdb.com/Title?Star+Wars+(1977)上抽取的部分信息的有向图表示。具体的抽取方法及算法将在后面给出。

*)本文研究部分得到国家自然科学基金资助。陈恩红 博士,主要研究领域为知识发现,机器学习及约束满足问题。范焱 博士生,主要研究领域为知识发现。王行甫 工程师,主要研究领域为知识发现。蔡庆生 博士导师,主要研究领域为人工智能,机器学习与知识发现。

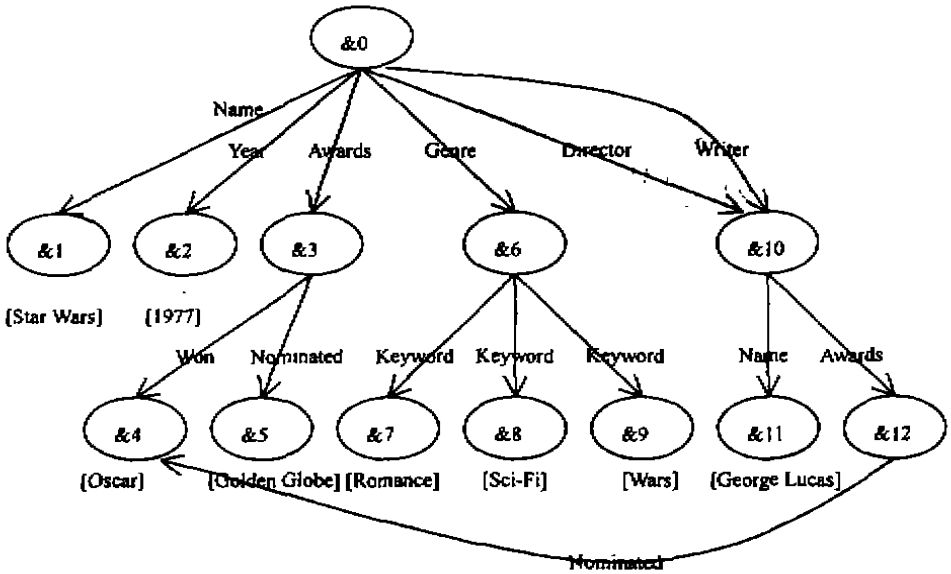

图1 OEM 模型表示的半结构化数据

图1表示的信息来自多个相关的 HTML 页面。例如关于"Star Wars"的主页上有"Awards"的图标，它链接到一个关于其获奖信息的页面，在该页面上还有到其它页的链接。因此，在半结构化数据抽取中，我们采用宽度优先策略，将所需要访问的链接及该链接所对应的边的出发节点标识保存在队列中，以保证所有要抽取的页面都能被取回处理。

3. 半结构化数据抽取算法

半结构化数据的特点是无固定模式，且同一属性在不同页面上取不同数量的值，甚至在某些页面上根本无值，这对信息抽取造成一定的困难。由于页面是以 HTML 形式存储的，因此，我们为每一类具有相似结构的页面提供一种辅助文件(见图2)，用于指导程序自动抽取所需属性的值，并且自动加上标记。辅助文件共有四种信息，其中第一列对应要抽取信息前带的 tag 标志，第二列对应需要加的标记，若为空(NULL)，则仅表示对应的第一列信息作结束标志用，因为有的属性可取多个值，所以我们需要一些信息来判断在 HTML 文件上何处为该属性取值结束处。第三列表示要取值的个数，其中"1"表示仅有一个值，"N"表示所有值都取，"-1"表示此信息只作结束标志，等。如果抽取的信息为另一页面的 http 地址时，第四列提供该页面的类型信息，它与地址一同放入队列中，以便抽取算法(见算法1)找到相应于该类页面的辅助文件。

〈TITLE〉	Name	1	0
HREF="/Sections/Countries/	Country	1	0
HREF="/Sections/Years/	Year	1	0
HREF="/More?awards+	Award	1	1
HREF="/List?production-companies=	Production	1	0
CLASS="smallkey">Genre/keyword:	Genre	-1	0
HREF="/List?genres=	Keyword	N	0
〈/TD〉〈/TR〉	NULL	N	0
HREF="/List?distributors=	Distributor	1	0
HREF="/Glossary/D#director"	NULL	1	0
HREF="/Name?	Director	1	2

图2 用于影片类主页信息抽取的辅助文件

算法1

```
Procedure extract_info(Q)
Input:Q:Queue to store the http address
Output:Directional Graph describing Semistructured Data
{
   Match←True;
   While(Q<>empty)do{
      addr←first entry in Q;
      get an HTML document Doc(addr)from remote web
      server;
   read the corresponding tag file Tag(Doc(addr));
   repeat{
      if (Match==True or Cur_tag==NULL) then
        S←Doc(addr)中的下一个串的起始位置;
      Cur_tag←Tag(Doc(addr))中指针所指的当前
           tag 信息;
      if (Cur_tag 为 S 所指的子串)then
        Match=True;
        P(S,Tag(Doc(addr)));
```

```
            else
                前移 Tag(Doc(addr))中的指针;
                Match=False;
            endif
        }Until EOF(Doc(addr))or EOF(Tag(Doc(addr)))
    }end while
}
```

在该算法中,首先从某 Web 站点获取 HTML 文本 Doc 后,借助与其对应的 tag 文件抽取所需的信息,其中过程 P(S,Tag(f))执行具体的数据抽取任务。该算法还对某些链接进行检测,若其所链接的下一页面有用(即存在所需抽取的信息),还需将相应的链接地址保存在队列中,以采用宽度优先策略来抽取所有所需信息。

由于同一个值可能出现在不同的页面,因此需要建立 Hash 表以保证值与标识唯一对应,即在有向图中均指向同一节点。由于我们是在 VC++5.0环境下实现的,因此 Hash 表的建立采用提供的模板类 CMap,构造指针变量 XCMapPtr。

```
typedef CMap(CString, LPCSTR, unsigned int, unsigned int)XCMap;
typedef XCMap * XCMapPtr;
```

算法2

```
insert_hash(XCMapPtr map,CString key)
Input:hash 表 map,及字符串型值 key,已分配的最大标识 seq;
Output:字符串型值 key 对应的标识;
{
    unsigned int seq1;
    //查找 key 是否已在 hash 表 map 中,若已在,
    //则将相应的标识赋给 seq1,
    if(!map->Lookup(key,seq1))
    {
        seq++;
        //当 key 不在 map 中,将其与相应的标识插入 hash 表
        map->SetAt(key,seq);
    return seq;
    }
    return seq1;
}
```

4. 半结构化数据中模式知识的发现

上述半结构化数据的抽取任务完成后,可用于各种目的的知识发现任务,就如同关系数据库中的知识发现一样。本文接下来介绍基于上述有向图形式描述的数据中的模式知识发现的过程及结果。

4.1 有关定义

算法建立在关联规则中高频数据集(frequent itemset)的发现方法基础上,但由于传统的关联规则发现方法考虑的是平面式的结构化数据,因此在实现上有较大区别。首先表现在对交易数据的定义上:

定义1 交易数据:无入边的节点对应的复杂对象称为交易数据。

以图1所示为例,其对应的交易数据为复杂对象 &0,具体值为:

(Name:Star Wars,Year:1977,Genre:Keyset,…,Director:George Lucas)

前已叙及,因为在半结构化数据中,标记信息表明语义关系,因此应在交易数据的表示中带有标记,如在 Name:Star Wars 中,Name 为标记,Star Wars 为原子对象 &o1的值。

此外,关于数据集也有不同的定义:

定义2 数据集:设$1\leqslant i_1<\cdots<i_n\leqslant P$ 且 $k>0$,则有:(i)每个标识符是自身的一个数据集;(ii)若 $f(id)=\{l_1:id_1,\cdots,l_p:id_p\}$ 且 x_i 是标识符 id_i 的数据集,则 $\{l_i:id_1,\cdots,l_p:id_p\}$ 是 id 的一个数据集。

定义3 k-高频数据集:有 k 个"标记:值"对的数据集称 k-数据集;支持度超过预定阈值的 k-数据集称为 k-高频数据集。若以有向图的形式来描述 k-高频数据集,则有 k 个节点的出度为0。

4.2 知识发现算法

由上述定义可看出,在半结构化数据的知识发现中,以第一层数据为交易数据,但在知识挖掘过程中,通过有向图利用深层的数据来发现高频数据集。实际上,这种高频数据集正是隐含在半结构化数据中的模式知识。算法3是模式知识发现的简要描述:

算法3

```
Procedure MineFreqScema()
Input:交易数据库 T;
Output:模式知识;
{
    F1←计算高频1-数据集;
    For (k=2,1≠Φ;k++){
        Ck←由 Fk-1生成选高频 k-数据集;
        计算 Ck 在 T 中的支持度;
        生成 Fk;
    }endfor
    精炼所生成的 F1∪F2∪…∪Fk-1;
}
```

虽然上述算法与传统的关联规则发现算法极为相似,但前已叙及,由于数据的形式由平面的关系型数据转向层次型的半结构化数据,所以在生成 k-高频数据集的具体实现上是不同的,并且根据所得的数据集是否需要对象值,可以得到两类不同的模式信息:一类是结构模式,它只含有标记信息,而忽略对象(或对象值),其目的是对全部交易数据的结构信息的获取,其中"1?"为通配符,表示任何标记;另一类是模式对象值关联模式,它不仅考虑标记,而且还包含对象(或对象值)的信息,它类似于传统的关联规则发现中的关联数据项。实际上,k-数据集可以视为一种树形结构,称之为嵌套树,可表示成序列 $P_1\cdots P_k$,其中 P_i 是数据集中的第 i 个(标记:值)对的嵌套路径,表示为$[\perp,l_1,id_1^{j_1},\cdots,l_n,id_n^{j_n}]$,即从初始节点⊥开始的标记、值序列,其中 $id_i^{j_i}$ 表示标记值对$(l_i,id_i^{j_i})$在 $f(id_{i-1})$中为第 j_i 次出现。K-数据集 $P_1\cdots P_k$ 对应的嵌套树通过如下途径构

造：①嵌套树的初始节点为⊥；②对$1 \leqslant i \leqslant k$，将$P_i$沿着最长匹配路径插入到嵌套树中。

由上可知，F_1的发现可通过寻找支持度(support)超过阈值的嵌套路径来完成。此外，在数据集中并不出现中间节点的标识，使得所有具有标记序列$l_1, l_2, \cdots, l_n$并中止在同一节点的嵌套路径都表示相同的1-数据集，因此一个交易数据有可能对某些1-数据集的支持度超过1，这显然不同于传统的关系型数据的关联规则发现中，一个交易数据对某个数据集的支持度不超过1的情况。在完成对F_1的计算后，就可对所有$F_k(k>=2)$进行计算，即由$P_1 \cdots P_{k-1}$和$P_1 \cdots P_{k-2}P_k$构造$P_1 \cdots P_{k-2}P_{k-1}P_k$。

在发现所有高频数据集后，将消除所有非最大的高频模式，因为这些非最大的高频模式所表达的信息同样也包含在最大的高频模式中。需要注意的是，无论$i<, =, >j$，一个i-数据集都可能使j-数据集成为非最大的高频数据集。例如，3-高频数据集{11:&o1,12:&o1,12:&o2}使2-高频数据集{12:&o1,12:&o2}成为非最大的高频数据集。若f(&I1)={10:{11:&o1,12:&o1},12:&o2}，2-高频数据集{10:&I1,12:&I2}使3-高频数据集{10:{10{11:&o1,12:&o1}},11:&I2}。由此可知，消除非最大高频数据集的任务只能在所有高频数据集都被发现后才能进行。

结束语 本文针对WWW上的半结构化数据的抽取及相应的模式知识发现进行了研究。随着信息化程度的日益提高，这类形式的数据将越来越丰富，因此对它的研究也日益受到人们的重视。今后我们将在上述工作基础上，作进一步的研究，如将机器学习方法引入到用于数据抽取的tag信息自动识别中，以减少对人的依赖性。此外，我们还将研究基于聚类方法的半结构化数据中的知识发现。

参考文献

1 Berners-Lee T, et al. The World Wide Web. Comm. Of ACM, 1994, 37(8): 76～82

2 Quass D, et al. Querying Semistructured Heterogeneous Information. In: Deductive and Object-Oriented Databases (DOOD), 1995(12月): 319～334

3 Abiteboul S. Querying Semi-structured Data. In: Proc. of Interl. Conf. on Database Theory, 1997

4 Goldman R, et al. A Standard Textual Interchange Format for the Object Exchange Model (OEM): [Technical Report]. Stanford University, 1996

5 Agrawal R, et al. Mining Association Rules between Sets of Items in Large Databases. SIGMOD, 1993: 207～216

6 Wang K, Liu H Q. Discovering Typical Structure of Documents: A Rad Map Approach. In: ACM SIGIR Conf. On Research and Development in Information Retrieval. Aug. 1998

7 王清毅，陈恩红，蔡庆生. 知识发现的若干问题及应用研究. 计算机科学，1997，24(5)：73～77

8 叶绚，陈恩红，蔡庆生. 关联规则的发现算法研究. 小型微型计算机系统，已录用

(上接第40页)

业系统与信息系统的接口代表。

·描述员工之间的合作关系：由于流程往往是一个协同过程，我们需要为相互合作的工作组提供支持，该过程的模型主要描述员工的合作关系。

相信随着BPR理论的完善，过程建模在BPR领域会得到更好的应用。

结束语 在竞争日益激烈的今天，如何提高企业的市场竞争力，使中国企业逐步走向世界是当代中国企业所面临的难题。西方管理者提出的业务流程重组实际上是要使企业各个环节的组合能更好地适应市场日趋向买方倾斜的竞争要求，增强企业整体的运行效率。在这个过程中，信息技术和现代计算机技术将发挥巨大的作用。所以我们应将信息技术和计算机技术合理应用于BPR领域，努力提高企业的效率和效益，为中国民族工业的腾飞作出贡献。

参考文献

1 屠晓光. BPR的概念. 经理世界，1998(9)

2 Poh H L, Chen W W. Business Process Reengineering: Definitions and Model Revisited

3 Enterprise Integration Laboratory Department of Industrial Engineering, University of Toronto Designing Tools to Support Business Process Reengineering, Ontario M5S 1A4

4 Warboys B. Reflections on the Relationship between BPR and Software Process Modeling. IPG Department of CS University of Manchester M13 9PL