文章编号: 1003-0077(2008)01-0022-08

一种全自动生成网页信息抽取 Wrapper 的方法

梅 雪[1,2],程学旗[1],郭 岩[1],张 刚[1],丁国栋[1]

(1. 中国科学院 计算技术研究所,北京 100080; 2. 中国科学院 研究生院,北京 100049)

摘 要: Web 网页信息抽取是近年来广泛关注的话题。如何最快最准地从大量 Web 网页中获取主要数据成为该领域的一个研究重点。文章中提出了一种全自动化生成网页信息抽取 Wrapper 的方法。该方法充分利用网页设计模版的结构化、层次化特点,运用网页链接分类算法和网页结构分离算法,抽取出网页中各个信息单元,并输出相应 Wrapper。利用 Wrapper 能够对同类网页自动地进行信息抽取。实验结果表明,该方法同时实现了对网页中严格的结构化信息和松散的结构化信息的自动化抽取,抽取结果达到非常高的准确率。

关键词: 计算机应用;中文信息处理;网页信息抽取;网页结构分离;包装器

中图分类号: TP391 **文献标识码**: A

Fully Automatic Wrapper Generation for Web Information Extraction

MEI Xue[1,2], CHENG Xue-qi[1], GUO Yan[1], ZHANG Gang[1], DING Guo-dong[1]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Web information extraction has been a hot topic in recent years. The challenge is how to extract important information from a large number of web pages as quickly and accurately as it can. In this paper a novel method is proposed for fully automatic wrapper generation for Web information extraction. This method makes use of structure of Web templates abundantly. It uses Web Page Link_Sort algorithm and Web Page Structure_Seperator algorithm to extract information from Web pages and output a wrapper accordingly. Experimental results showed that this method performs well in both rigidly and loosely structured records in Web pages.

Key words: computer application; Chinese information processing; Web information extraction; Web structure seperator; wrapper

1 引言

随着互联网的快速发展,任何利用 Web 数据进行生产或者研究的项目必然先遇到 Web 数据抽取的问题。因此,近年来各种与 Web 数据抽取相关的研究工作大量出现在各种学术会议、期刊杂志中。这些工作涵盖了人工智能、数据挖掘、数据库和信息检索等多个领域。我们迫切需要一种高自动化的网络信息抽取技术来整合纷繁的网络信息资源。目前比较流行的自动化信息抽取工具有 MDR[4]、基于 MDR 的改进方法 Depta[5] 等。但这些工具的目标都是从产品列表或表格中抽取信息,也就是说对待抽取网页中信息的结构化程度要求比较严格。然而,网络中还存在着大量的网页,它们承载着松散的结构化信息,例如博客信息等。无论是严格的结构化信息,还是松散的结构化信息,对于信息挖掘者来说都是宝贵的财富,都需要借助高效的抽取工具来获得这些信息。本文基于网页模板的设计准则提出了全自动生成网页信息抽取 Wrapper 的方法 PSNT (extraction based on temPlate Structure aNd Tag

收稿日期: 2007-05-21 **定稿日期**: 2007-12-03

基金项目: 国家高技术研究发展计划(863)资助项目(2005AA142110)

作者简介: 梅雪(1982—),女,硕士生,研究方向为信息抽取与数据挖掘;程学旗(1971—),男,研究员,博导,主要研究方向为网络信息安全、大规模信息检索与信息挖掘、PZP 计算;郭岩(1974—),博士,助研,研究方向为信息抽取与挖掘。

tree) ,该方法同时实现了对网页中严格的结构化信息和松散的结构化信息的自动化抽取。

动态网页是指由网站后台数据库中的数据通过事先定义好的模板动态生成的网页。网页模板的设计必须满足布局清晰,可读性强。当该模板生成的网页呈现在浏览器下时,用户能立即识别出该网页所包含的各个数据区域及区域内部各条数据信息。为了便于用户快速而准确地识别,网页中区域之间以及区域内部各信息之间都需要存在明显的分界。网页信息的规整划分很大程度上是由生成该网页的高度结构化的模板设计所实现的。这种高度结构化的设计特点有效地将区域之间以及区域信息之间的区分转化为他们所在的结构体之间的层次关系。我们可以利用这一特点对网页信息进行抽取。抽取的关键和实质是挖掘出网页背后所隐藏的模板中各个结构体之间的层次关系,从而准确抽取出我们需要的信息。基于这种思想,本文提出了结构体相对等势原理和结构体分离算法,进而提出了全自动生成网页信息抽取 Wrapper 的方法 PSNT。该方法充分利用网页设计模版的结构化,层次化特点,抽取出网页中各个信息单元,并输出相应 Wrapper。利用 Wrapper 能够对同类网页自动地进行信息抽取。实验结果表明,该方法同时实现了对网页中严格的结构化信息和松散的结构化信息的自动化抽取,抽取结果达到非常高的准确率。

2　相关工作

目前,国内外在网络信息抽取领域的研究主要集中在: 如何建立针对各类网站的全自动化信息抽取工具,并将这些信息按照一定的格式进行整合,支持各类计算机应用。目前发表的大部分文献都是针对某一类特定网站的信息抽取。比如 MDR[4] 和 Depta[5] 是针对商品清单或表格信息的抽取,文献[2,3]致力于对搜索引擎返回页面记录的提取。这些特定网站中的网页往往承载着格式化很强的信息,例如记录信息,商品价目信息等。而我们的方法 PSNT 既能够处理严格的结构化信息,又能够处理松散的结构化信息。WDE[1] 提出了一种可自适应的全自动抽取方法,文献中表明该方法对严格的结构化信息和松散的结构化信息都有很好的抽取效果。但该方法和大多数抽取方法一样,都是将树编辑距离(the tree-edit distance)[6] 作为确定信息抽取模式的基本依据。树编辑距离是对标签串(信息包含的所有标签顺次排列所形成的字符串)相似度的量化,通过将树编辑距离与一个预设的临界值相比较达到对两个标签串相似度的判断,从而实现对信息的识别。事实上,即使网页中同一区域内的各信息,它们所包含的标签串也可能存在无法预估的差异。本文所提出的 PSNT 算法未使用树编辑距离,而是运用信息所在的结构体之间存在的某种关系来实现对信息的抽取。

3　全自动生成网页信息抽取 Wrapper 的方法 PSNT

我们提出的全自动生成网页信息抽取 Wrapper 方法的框架结构如图 1 所示。该方法的输入是同一网站的两个同类(即组织结构相似) Web 网页的 HTML 文档,输出是针对该类网页的一个信息抽取 Wrapper。图中虚线框内的三个模块各自的抽取规则共同作用生成一个 Wrapper。下面我们将对图 1 中的各模块进行详细阐述。

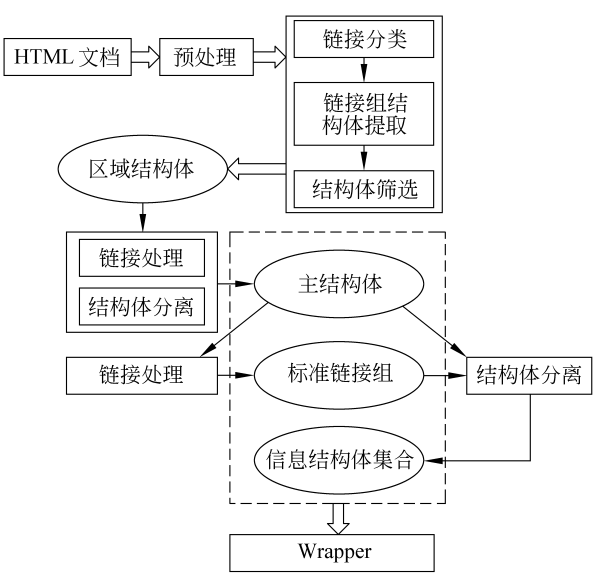

图 1　PSNT 方法的框架图

3.1　预处理过程

我们首先获得同一网站的两个同类 Web 网页,通过比较这些网页中的链接,去除所有公共链接,然后任选其中一个网页,为之建立标签树。

3.2　区域结构体的生成

3.2.1　标签路径及结构体的概念

一个页面基于它的 HTML 源文件中的标签可

以转化成一棵树来表示,这棵树叫做标签树。这棵树的根节点是〈HTML〉标签,并且所有的内容节点(文本,图片等)都是这棵树的叶节点。每一个内部节点代表一对标签(开始的标签和结束的标签),或者仅代表一个标签(该标签没有对应的结束的标签,比如说〈BR〉),根标签和内部的节点统称为标签节点。标签树中的任何一个节点可以通过从根节点到该节点的一条路径来定位,我们称该路径为标签路径。一个标签路径包含一系列的路径节点,根据对标签路径的不同定义,每一个路径节点的构成元素是不同的。本文借用文献[2]中定义的两种标签路径,一种是不严格的标签路径(Relaxed Path),另一种是基于索引的标签路径(Index Path),分别记为 $Rpath$ 和 $Ipath$。本文运用 $Rpath$ 对链接进行分组,运用 $Ipath$ 确定结构体路径。

定义 1:如果 n_1, n_2, …, n_{k-1}, n_k 是标签树 T 中的标签节点,其中 n_1 是 n_2 的双亲节点,n_2 是 n_3 的双亲节点,以此类推,直到 n_{k-1} 是 n_k 的双亲节点,那么节点 n_k 的 $Rpath$ 定义如下:$Rpath(n_k) = n_1.n_2 \cdots n_{k-1}.n_k$。

定义 2:如果 n_1, n_2, n_3 是标签树 T 中的三个标签节点,其中 n_2 是 n_1 的第 i 个直接孩子节点,n_3 是 n_2 的第 j 个直接孩子节点,那么节点 n_3 的 $Ipath$ 定义如下:$Ipath(n_3) = n_1.n_2[i].n_3[j]$。

我们提出以下几个关于结构体的概念:

(1) 结构体:标签树中的一个标签节点及其子树共同构成一个结构体,该结构体在网页中呈现为一片连续的区域。

(2) 结构体的表示:标签树中的一个标签节点及其子树共同构成的结构体由该标签节点的 $Ipath$ 来表示,该结构体称为该标签的结构体。

(3) 结构体的嵌套关系:假设 n_1, n_2, …, n_{k-1}, n_k 是标签树 T 中的标签节点,其中 n_1 是 n_2 的双亲节点,n_2 是 n_3 的双亲节点,以此类推,直到 n_{k-1} 是 n_k 的双亲节点,那么节点 n_k 的结构体的嵌套关系就是指节点 n_1, n_2, …, n_{k-1}, n_k 分别对应的结构体之间的有序列,这种有序列在网页中呈现为若干连续区域之间的有序嵌套。结构体的嵌套关系可以用 $Rpath$ 描述,例如节点 n_k 的结构体的嵌套关系是 $n_1.n_2 \cdots n_{k-1}.n_k$。因此,如果两个节点的结构体的嵌套关系相同,那么这两个节点的 $Rpath$ 必然相同;反之亦然。

3.2.2 链接分类

通过观察,我们有第一个发现:网页中的数据信息可以是相对独立的记录信息(如搜索引擎结果页面的记录),同一表格中的条目信息(如论坛版面页面中的各条帖子记录),也可以是相对离散的信息(如博客)等等。这些信息有一个共同点,即每条信息至少包含一个超链接,该链接指向与本条信息相关联的网页。

另外我们还有第二个发现,一个设计规整的网页在浏览器下呈现时,具有如下两个重要的特点:

(1) 网页的各个区域之间存在明显的分界,这让用户能够立刻识别出各个区域;

(2) 网页的同一个区域内各信息之间同样存在分界,并且同一区域内各信息之间的分界是相似的。

基于以上两个特点,我们可以得出以下两点分析结果:

(1) 网页的各个信息区域(广告区域,主数据区域等)通过程序置于该网页模板的不同结构体中;

(2) 各个信息区域内的各条信息有两种放置方式:

a) 通过程序置于该区域所在结构体中的不同子结构体之中,这些子结构体具有相同的结构体嵌套关系,这就意味着各条信息之间既区别又联系;

b) 通过程序顺次置于该区域所在结构体之内,即包含每条信息的最小结构体就是该条信息所在的整个信息区域的结构体。

基于以上两点分析,我们可以得出:同一信息区域内的各条信息所在的最小结构体拥有相同的结构体嵌套关系。

基于以上第一个发现,我们可以将对信息的研究转化为对信息的标题链接的分析。基于以上第二个发现,我们可以得出,同一信息区域内各条信息的标题链接所在的最小结构体拥有相同的结构体嵌套关系。这里需要说明的是,每一个链接所在的最小结构体的结构体嵌套关系就是该链接所对应的〈a〉标签在标签树中的 $Rpath$。我们的任务是抽取出主数据区域(包含网页主要内容的区域)中的各条信息,于是我们可以通过分析网页中各链接所在的最小结构体的结构体嵌套关系,来间接完成任务。因为结构体的嵌套关系可以用 $Rpath$ 描述,所以我们用 $Rpath$ 对网页中所有链接进行分类,即 $Rpath$ 相同的链接分为一组。

3.2.3 区域结构体的生成

根据经验,我们假设页面中各信息区域所包含的信息数不小于 4,即所有信息区域都包含至少 4 个 $Rpath$ 相同的链接。于是我们根据链接的

Rpath 对链接进行分组,去掉链接数小于 4 的链接组。

根据经验,我们提出结构体的筛选原理: 主数据区域(包含网页主要内容的区域)中的字符数占整个页面字符总数的一半以上。基于该原理,我们提出结构体筛选的启发式规则: 在字符数大于页面字符总数一半的结构体中,挑选所含字符数最小的结构体作为区域结构体。于是我们得出区域结构体的抽取算法,基本步骤如下:

(1) 求出页面中所有链接的 *Rpath*;

(2) 将 *Rpath* 相同的链接分在同一组中;

(3) 删除链接数小于 4 的链接组;

(4) 计算出每组中各个链接的 *Ipath*,并将该组中所有链接 *Ipath* 的最大公共部分作为该组链接确定的最小结构体路径(*Ipath*);

(5) 将相同的结构体(*Ipath* 完全相同)进行合并,并计算出每个结构体所包含的字符总数;

(6) 在字符总数大于页面字符总数 1/2 的所有结构体中,将字符总数最小的结构体作为我们的区域结构体。

这里,对以上第(4)个步骤举例如下: 假设两个链接的 *Ipath* 分别为 $n_1.n_2[i].n_3[j]$ 和 $n_1.n_2[i].n_3[k]$,那么这两个链接的 *Ipath* 的最大公共标签前缀是 $n_1.n_2[i]$,于是,这两个链接所在的最小结构体的路径为 $n_1.n_2[i]$。

3.3　主结构体的生成

区域结构体中一定包含了所有主要数据信息,但它不一定是包含所有主要数据信息的最小结构体。主结构体是对区域结构体进行结构体划分而得到的。主结构体的确定使我们的信息抽取过程更简单。

3.3.1　结构体分离算法

我们先提出有序链接组的抽取算法: 将结构体中的所有链接依照 *Rpath* 进行链接分组(*Rpath* 相同的链接分为一组)。删除链接数小于 4 的组。剩余链接组按如下启发式规则进行链接组的筛选:

(1) 删除重复链接最多的组,重复是指两个链接包含完全相同的字符,如快照;

(2) 计算剩余每组链接的平均链接长度,平均链接长度 = 链接总字符数/链接个数;

(3) 将平均链接长度最大的链接组作为该结构体的有序链接组。

这里,我们提出结构体相对等势原理: 如果结构体 Q 和 P 是两个不存在任何重叠的结构体,将 P 和 Q 所确定的最小结构体记为 $S(P,Q)$,那么 Q 和 P 相对于 $S(P,Q)$ 是等势的。本文中通过引入相对等势原理,反映了网页模板结构体之间的层次关系。需要指出,本文将单独的链接视为一个结构体,该结构体是该链接所对应的〈a〉标签的结构体,因此链接之间同样满足结构体相对等势原理。基于结构体的相对等势原理,我们提出了结构体分离算法。

结构体分离算法是利用待划分结构体的有序链接组对该结构体进行的划分。假设待划分的区域结构体为 M,该结构体的有序链接组 L 为: $L_1, L_2, \cdots, L_n$,R 是结构体 M 的最终划分结果集,R 中的各个元素相对于结构体 M 都是等势的。那么,结构体 S 的分离算法如下:

(1) 取出有序链接组中的前两个元素 L_1 和 L_2,计算 L_1 和 L_2 共同所在的最小结构体,记为 $S(L_1, L_2)$;

(2) 比较 $S(L_1, L_2)$ 和待划分结构体 M,如果结构体 $S(L_1, L_2)$ 就是结构体 M,说明结构体 L_1 和结构体 L_2 相对于 M 是等势的,此时将 L_1 从有序链接组 L 中删除,同时将结构体 L_1 作为一个独立的元素标记到 R 中;

(3) 如果结构体 $S(L_1, L_2)$ 包含于待划分结构体 M,那么,将 L_1 和 L_2 从 L 中删除,同时将 $S(L_1, L_2)$ 作为有序链接组的第一个元素;

(4) 以上步骤直到有序链接组中已不存在任何单链接,此时将 L 中剩余元素(最多一个)的结构体作为一个独立的元素标记到 R 中。

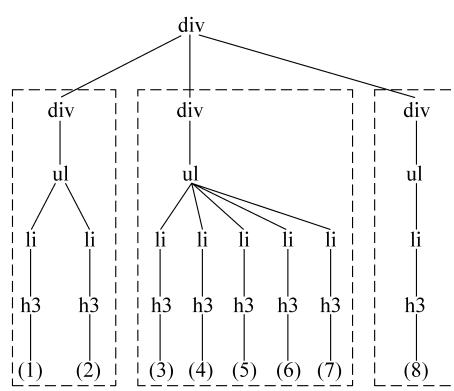

图 2　结构体分离算法举例

下面我们举例说明结构体分离算法。图 2 是 HTML 页面标签树的一部分,该树的根节点 div 的结构体是我们待划分的区域结构体,所有叶节点是该区域结构体的有序链接组。为了便于描述,我们用数字表示有序链接组中的各个链接。我们对该区

域结构体进行结构体分离: 计算链接(1)和(2)所确定的最小结构体为 div. div[1]. ul[1],该结构体包含于根节点 div 的结构体中,进一步计算 div. div[1]. ul[1]和(3)所确定的最小结构体,该结构体即为根节点 div 的结构体,因此 div. div[1]. ul[1]和(3)相对于根节点 div 的结构体是等势的,于是将 div. div[1]. ul[1]进行标注,作为根节点 div 的结构体中一个独立的子结构体。接着计算 $S(3,4)$,该结构体为 div. div[2]. ul[1],该结构体包含于根节点 div 的结构体,进一步计算 div. div[2]. ul[1]和(5)所确定的最小结构体,以此类推,直到 S(div. div[2]. ul[1], 8) = 根节点 div 的结构体,此时将 div. div[2]. ul[1]进行标注,作为根节点 div 的结构体中的另一个独立的子结构体。此时,剩余链接(8)未标注,按照结构分离算法,将(8)作为独立的子结构体进行标注。

3.3.2 主结构体生成

假如 R 是结构体 M 的最终划分结果集,我们按照以下原则确定主结构体:

(1) 计算 R 中各个结构体所包含的字符总数,如果 R 中最大结构体(字符数最多)所包含的字符数大于整个页面字符总数的一半,将该最大结构体作为主结构体;

(2) 如果 R 中最大结构体包含字符数小于页面字符总数的一半,说明没有任何结构体独立包含所有的主要数据信息。如果 R 中的元素数目小于 4,本文认为主结构体为划分结果集合 R,此时的主结构体是一些独立的结构体集合;如果 R 中的元素数目大于或者等于 4,则认为主结构体为区域结构体本身。

基于以上原则,图 2 中根节点所代表的区域结构体 div 经过结构体分离算法后,得到的主结构体为 div. div[2]. ul[1]。

3.4 标准链接组生成

因为主结构体的有序链接组是构成信息抽取 Wrapper 的重要组成部分,所以为了和信息抽取过程中其他结构体所产生的有序链接组进行有效区分,故将主结构体的有序链接组称为标准链接组。

标准链接组的确定有以下三种情况:

(1) 当区域结构体就是主结构体时,区域结构体的有序链接组就是主结构体的有序链接组,此时的标准链接组即为区域结构体的有序链接组;

(2) 当主结构体是结构体集合 $R(|R|<4)$ 时,区域结构体的有序链接组就是主结构体的有序链接组,此时的标准链接组即为区域结构体的有序链接组;

(3) 当主结构体仅仅是区域结构体的一个组成部分时,主结构体按照有序链接抽取算法(参见第 3.3.1 节)找到自己的标准链接组。

3.5 信息结构体集合的生成

主结构体中一定包含了所有主要数据信息,但并不是其中所有的信息都是我们想要的,只有和标准链接组相关的信息才是我们要的。于是我们提出信息结构体集合的概念: 主结构体中每条数据信息(指我们需要的信息)所在的最小结构体集合称为信息结构体集合。

信息结构体集合不同于主结构体基于标准链接组进行结构体分离算法所得到的结果集合。假如主结构体 M 经过结构分离算法后得到的结果集合为 $R\{r_1, r_2, r_3, \cdots, r_m\}$,其中各个元素代表一个独立的相对等势的子结构体,一个子结构体中可能只包含一个标准链接,也可能是包含若干个标准链接的(如某一信息含有至少两个标准链接组中的链接)。R 的作用是将主结构体进行了分割,R 中的元素和主结构体内每条信息之间是一一对应的关系。

$R\{r_1, r_2, r_3, \cdots, r_m\}$ 中元素和信息所在最小结构体之间的映射为: 假设结构体 z_1 包含子结构体 r_1,z_1 和 r_2 相对于主结构体等势,且任何包含 z_1(z_1 除外)的结构体都不与 r_2 相对于主结构体等势,那么,本文将 z_1 称为 r_1 所对应的信息结构体。以此类推,我们分别将 $z_2, z_3, \cdots, z_m$ 作为 $r_2, r_3, \cdots, r_m$ 所对应的信息结构体,则有 $Z\{z_1, z_2, z_3, \cdots, z_m\}$ 称为信息结构体集合。

特别是当主结构体为结构体集合 R 时,我们对该集合中的一个结构体(如 $R1$),基于它所包含的标准链接组(主结构体标准链接组的一部分)进行结构体分离算法,并得到划分结果集合 $R1$。将 $R1$ 中各元素映射到信息最小结构体。最终,信息结构体集合为 R 中各结构体的信息结构体集合的相加。

如图 2 所示,该网页的主结构体为 div. div[2]. ul[1],标准链接组为(3),(4),(5),(6),(7),首先对该主结构体进行结构分离。$S(3,4)$ 为 div. div[2]. ul[1],将链接(3)进行标注,作为划分结果集合中独立的结构体元素,计算 $S(4,5)$,结果仍然是 div. div[2]. ul[1],将链接(4)进行标注,并且加入划分结果集合,以此类推,最终得到主结构体基于链接(3),(4),(5),(6),(7)的划分结果集合 $R\{3,4,5,6,$

7}。进一步找出 R 中每个元素所对应的信息结构体。包含链接(3)并且与链接(4)相对于 div. div[2]. ul[1]等势的最大结构体为 div. div[2]. ul[1]. li[1],因此将 div. div[2]. ul[1]. li[1]作为 R 中链接(3)所对应的信息结构体。同理,包含链接(4)且与链接(5)相对于 div. div[2]. ul[1]等势的最大结构体为 div. div[2]. ul[1]. li[2]。最后,我们将结构体集合{div. div[2]. ul[1]. li[1],div. div[2]. ul[1]. li[2],div. div[2]. ul[1]. li[3],div. div[2]. ul[1]. li[4],div. div[2]. ul[1]. li[5]}称为该网页的信息结构体集合,集合中的每个元素包含了且仅包含了一条数据信息。

3.6　信息抽取 Wrapper 的生成

3.6.1　信息抽取 Wrapper 的描述公式

本文描述 Wrapper 的公式为: 主结构体的 $Rpath(N)$(标准链接组相对主结构体的 $Rpath$)[信息结构体相对主结构体的 $Rpath$]。公式中: 若 $N \geq 1$,则 N 代表主结构体的表现形式;若 $N = 1$,则表示主结构体为一个独立的整块结构体;若 $1 < N < 4$,表示主结构体为一个结构体的集合,它所包含的结构体数目为 N。

图 2 中根节点 div 的 $Rpath$ 为 HTML. body. div. div. div. div. ul(图中省略了从标签树的根节点 html 到区域结构体根节点 div 的标签树部分)。图 2 对应的网页生成的信息抽取 Wrapper 的描述公式为: html. body. div. div. div. div. ul(1)(li. h3. a)[li]。

3.6.2　Wrapper 在网页信息抽取的运用

已知一类网页的 Wrapper,则对于该类的任何网页做如下处理: 首先对网页进行预处理,得到网页的标签树;然后直接运用 Wrapper 公式的第一部分提取出该网页的主结构体,如果主结构体数目与 N 不符合,则选取字符数最大的 N 个结构体作为主结构体;运用公式的第二个部分找到该主结构体中的标准链接组,标准链接组中的各个链接对公式第三部分直接提取出的信息结构体进行识别(即每个信息结构体至少包含一个标准链接);将符合条件的信息结构体中的内容抽取出来,得到一个信息集合,完成对该网页的信息抽取。

这里需要说明的是: 如果公式第三部分(信息结构体相对于主结构体的 $Rpath$)最末端为一个链接标签,此时,该路径下的每个链接的最小结构体(除本身)都为主结构体,则应该将链接之间的部分(包含开始链接,不包含下一链接)作为一条完整的信息。

4　实验和结果分析

为了对本文提出的算法进行评价,我们将 PSNT 抽取算法和比较流行的抽取算法 MDR 进行了比较。实验使用了 31 个实际网站,其中 24 个网站是包含商品条目,表格信息,搜索记录信息等严格结构化信息的网站(参见表 1),另外 7 个网站是包含用户评论,博客等松散的结构化信息的网站(参见表 2)。本文用传统意义上的准确率和召回率来评价我们的算法,其中召回率是指算法抽取出的正确信息条数所占网页中实际的信息总条数的百分率,准确率是指算法抽取出的正确信息条数与所有抽取信息条数的百分比。最后实验所得的召回率和准确率是手工计算所得的。需要说明的是,我们的方法对实验中各网站所生成的 Wrapper 数不超过 2 个,大多数网站只有唯一的 Wrapper。

表 1 和表 2 的各列如下描述: 表中的第一列是实验对象网站,每个算法下对应的第一列为信息总条数,第二列为抽取正确的信息条数,第三列为抽取错误的信息条数。

表 1　严格的结构化信息的抽取结果

	PSNT			MDR		
shopping. yahoo. com	15	15	0	15	11	4
youtube. com	20	20	0	20	20	0
google. com	10	10	0	10	0	0
baidu. com	10	10	0	10	0	0
foodtv. ca	10	10	0	10	10	0
ebay. com	50	50	0	50	50	0

续表

	PSNT			MDR		
taobao.com	40	40	0	40	0	0
exactseek.com	10	10	0	10	0	0
mozdex.com	10	10	0	10	0	0
searchhippo.com	10	10	0	10	10	0
aaas.org	20	20	0	20	20	0
dmoz.com	20	20	0	20	20	9
download.com	10	10	0	10	9	0
msn.com	9	9	0	9	0	25
buzzle.com	10	10	0	10	0	0
alltheweb.com	10	10	0	10	0	0
sogou.com	10	10	0	10	0	0
forum.xinghuanet.com	36	36	0	36	36	0
joeant.com	10	10	0	10	0	0
discussion.forum.nokia	22	0	5	22	0	0
www.kodak.com	10	10	0	10	10	0
netsearch.org	10	10	0	10	0	0
del.icio.us	12	12	0	12	0	0
nextag.com	11	11	0	11	11	0
平均召回率	94.3%			53.8%		
平均准确率	98.6%			84.4%		

表 2 松散的结构化信息的抽取结果

	PSNT			MDR		
blackberry.com	7	4	0	7	0	0
messages.yahoo.com	20	20	0	20	20	1
operawatch.com	5	5	0	5	0	0
engadget.com	15	15	0	15	0	0
thinkprogress.org	30	30	0	30	0	0
gizmodo.com	6	6	0	6	6	19
forums.gentoo.org	11	11	4	11	11	0
平均召回率	96.8%			39.3%		
平均正确率	95.8%			64.9%		

从表 1 和表 2 中我们可以看到,无论是抽取严格的结构化信息,还是抽取松散的结构化信息,无论在准确率方面还是召回率方面,我们的方法 PSNT 都高于 MDR,这种差异在表 2 中尤为明显。从表 1 和表 2 中我们还可以看到,PSNT 对一些网站的抽取出现了失误。经过分析我们认为这是因为实验中 PSNT 仅仅支持 4 个区域以下的抽取,当区域数大于等于 4 时,PSNT 将区域当作记录进行抽取,于是出现抽取错误。

5　结束语

本文提出了一种全自动化信息抽取方法 PSNT。该方法充分利用网页设计模版的结构化、层次化特点,运用网页链接分类算法和网页结构分离算法,抽取出网页中各个信息单元,并输出相应 Wrapper。利用 Wrapper 能够对同类网页进行自动信息抽取。不同于以前的信息抽取方法,该方法无论是对严格的结构化信息还是松散的结构化信息的抽取都表现出很好的准确率和召回率。PSNT 的下一步工作将致力于对多区域(区域数不小于 4)的网页信息的抽取。

参考文献:

[1] Justin Park and Denilson Barbosa. Adaptive Record Extraction From Web Pages [A]. WWW 2007 [C].

[2] Mundluru, D., Katukuri, J. R., and Celebi, S.. Automatically Mining Search Result Records [J]. Data Mining 2005.

[3] Zhao H., Meng W., Wu Z., Raghavan V., and Yu C.. Fully Automatic Wrapper Generation for Search Engines [A]. In: WWW 2005 [C]. 66-75.

[4] Liu B., Grossman R., Zhai Y. Mining Data Records in Web Pages [A]. KDD 2003 [C]. 601-606.

[5] Zhai Y., and Liu B.. Web Data Extraction Based on Partial Tree Alignment [A]. WWW 2005[C]. 76-85.

[6] Zhang K., and Shasha D.. Tree Pattern Matching. In: Pattern Matching Algorithms [M]. Oxford University Press, 1997.

国家"十五"重点图书出版规划项目之一
《中国的语言》由商务印书馆出版

经过 6 年左右近百名语言学界专家学者和商务印书馆编辑们的共同努力,第一部全面介绍中国境内 129 种语言基本情况的专著——《中国的语言》,于 2007 年由商务印书馆出版。该书是国家"十五"重点图书出版规划项目之一,由中国社会科学院民族学与人类学研究所孙宏开、胡增益、黄行主编,国内少数民族语言学界的 90 多位一流的专家学者编写。

该书的基本框架和所收语言在国内外学术界基本上已经形成共识,根据《中国大百科全书》早已经确定的语系、语族、语支的框架编排,分为七编:第一编为导论,简要介绍中国语言调查的成就(包括汉语方言调查和少数民族语言调查),讨论中国语言的调查方法、谱系分类、语言和方言的区分、类型学特征、研究方法论,书面语的体系和类型等问题。从第二编起,按照语系(及其语族)的顺序,分别介绍了汉藏语系、阿尔泰语系、南亚语系、南岛语系、印欧语系、混合语等的 129 种具体语言(其中包括分布在台湾省高山族使用的 15 种少数民族语言)的特点,主要内容包括:语言的分布、使用人口、使用状况、语言系属;语言的语音、语法、词汇特点;语言的方言情况,包括分布、使用人口、差异、特点;语言的书面形式及特点;语言的简要研究情况以及 500 字左右的英文提要。

在该书所介绍的语言中,有许多是近年来新发现的语言。可以说,这部书反映了中国语言调查的最新研究成果,是了解中国国情的一个基本内容,也是国内外语言学界掌握中国语言的资料最全面、观点最新的专著。书中对当前学术界一些有争议的问题,如语言和方言的界限、语言的谱系分类等,提出了作者的观点。

在该书编辑过程中,还得到了全国人民代表大会常务委员会副委员长许嘉璐教授、国家语言文字工作委员会前副主任王均研究员、中国社会科学院副院长江蓝生研究员的指导和支持,他们分别为该书撰写了序言。

《中国的语言》16 开 1 卷本,2 600 余页,约 360 万字,定价 338 元。