

一种基于页面 Block 的 Web 信息提取方法

蒙 韧,邵延振,袁鼎荣

(广西师范大学,广西 桂林 541004)

摘 要:基于页面结构的信息提取是 Web 数据挖掘中三大研究领域之一。该研究的关键技术是如何识别 Web 页面的组织形式,从中挖掘所需要的页面信息。文中基于页面的语义分块(Block)给出一个新的块主题提取算法,与传统的以页面为单位的 Web 信息提取相比,更符合实际情况,粒度优势明显。该算法针对页面中不同分块的重要性给予不同的权值,依据权值大小取舍页面信息提供给用户。针对该算法进行了模拟实验,从实验结果可以看出该算法具有一定的实用性和有效性。

关键词:语义 Block;Block 权值;Block 主题提取;Web 信息挖掘

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2010)01-0197-04

A Web Information Extraction Algorithm Based on Web Page

MENG Ren, SHAO Yan-zhen, YUAN Ding-rong

(Guangxi Normal University, Guilin 541004, China)

Abstract: Information extraction based web page structure is one of three web data mining's research fields. Key technology of the research is how to recognize web page's organization form and mine the needed information. Introduces a new block topic-extracted algorithm based on semantic block. Compared with traditional information extraction based on web page, it is more accordant to the fact and the advantage of granularity is evident. This algorithm gives different block weight values according to the importance of different blocks in a web page. Extract useful information for users according to magnitude of block weight. Simulation experiment was preformed for this algorithm. This algorithm has high practicability and effectiveness.

Key words: semantic block; block weight; block topic extraction; web data mining

0 引 言

互联网技术不断成熟和发展,网络信息资源的急剧增长,要想找到符合要求的信息资源越发困难,不仅需要浏览的网页数量剧增,而且网页的结构和内容越发复杂和多样,从网页中快速提取主要信息变得非常必要。

以整个的 Web 页面作为最小的信息提取单元的方式已逐渐不能适应 Web 页面信息提取的快速发展,因此,把页面按照一定的算法划分为若干个区域(Block),把这些区域作为基本的信息处理和提取单元,提出该方法的理由如下:

(1)当今的 Web 页面大部分是由多个不相关的主题页面 Block 构成,以页面 Block 为最小单位来提取 Web 信息更符合实际情况,粒度上优势明显。

(2)基于页面 Block 进行提取信息,可以适当忽略稳定的且内容无关的区域,节省处理和存储代价。

1 页面 Block 的特征提取

1.1 页面 Block 的基础:

Block 是页面中在内容和显示上独立的、闭合的矩形区域。Web 页面可以分割为若干互不相交的 Block,把这个过程称为页面分区。递归定义一个 Block 可以由多个相互不重叠的 Block 组成,则图 1 所示的页面可以分解为图 2 所示的树形结构,称为 Block 树。分区级别表示该分区算法产生的 Block 树的最大深度,用 Level 表示。选择合适的分区级别有助于得到最理想的分区结果。可以把 Block 定义成一个由主题、内容组成的二元组:Block = < topic, content >, 其中, topic 为 Block 主题,即能够表示 Block 内容概要的一段文字;content 为 Block 内容,即 Block 中存储的 HTML 文档。用 Layout 表示网页的布局,即剔除了 Block 的网

收稿日期:2009-04-17;修回日期:2009-07-27

基金项目:广西自然科学基金(桂科自 0640069)

作者简介:蒙 韧(1973-),男,工程师,研究方向为数据挖掘。

页框架。

1.2 页面分块的基本技术

1) DOM 树。

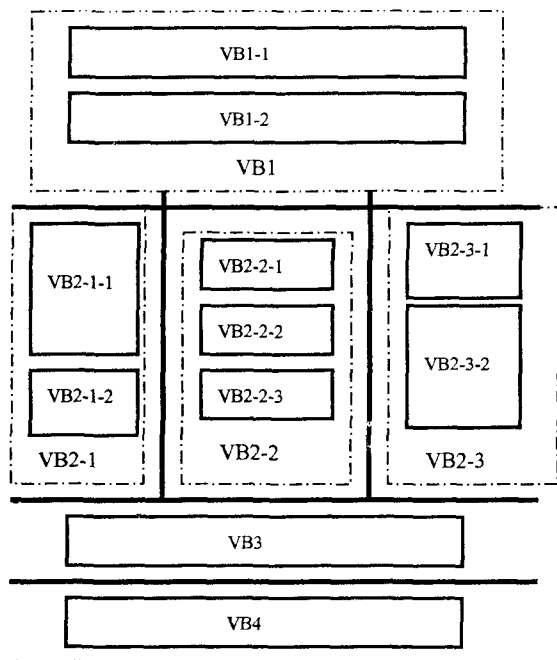


图 1 VIPS 页面分区的示例

DOM 树把文档中的每个组成部分定义为一个节点,在 DOM 树中,每个 HTML 元素都对应一个节点,节点之间的父子关系表示 HTML 标签在页面之前的嵌套关系,兄弟关系表示了 HTML 标签在页面中的并列关系,这样在使用 DOM 树进行解析的时候,

它在内存中构建起一棵完整的解析树,借此实现对整个 HTML 文档的全面的、动态的有层次的访问,将所有的 HTML 页面中的元素都解析成树中的节点,最后通过 HTML 解码器将解析的结果以语法树的形式输出,就生成 HTML DOM 标记树,每一个网页对应一个 DOM 标记树。

2) VIPS (Vision-based Page Segmentation)。

DOM 提供了树形结构的页面模型,因此,可以基于 DOM 来建立 Block 树。可以借助大量 HTML 标签来获取布局 and 位置信息,如 <P>, <TABLE>, <TR>, , <H1> <H6> 等。由于 HTML 语法的灵活性,很多 Web 页面并不完全遵守 W3C 的 HTML 规范。此外,依据 DOM 的分区只能代表布局结构独立的区域而不能完全代表语义上独立的区域,比如,同

一个父节点下面的两个子节点并不一定代表相同的主题。因此,VIPS 算法不能完全依赖 DOM,还需要考虑其他一些因素。通常,显示上独立的区域一般表示相同的主题,Web 中提供了大量可见的元素来划分页面,例如字体、颜色、图像、空白,这些都是页面分区算法需要考虑的元素。

VIPS 算法将一个 Web 页面 Ω 可以被描述成一个三元组 $\Omega = (O, \Phi, \delta)$ 。其中 O 是块的有限集合,所有的这些块都不是重叠的。 Φ 是分隔符的有限集合,包括水平线和垂直线以及 HTML 的各种标签,并根据其可视性赋予了一定的权值。 δ 表示两个分割块之间的邻接关系。

算法抽取出第 n 级 Block (n 初始化为 1),组成深度为 n 的 Block 树,同时保存 Layout。然后,判断 n 是否满足给定的分区级别,如果小于该级别,则依次把 Block 树中的每个叶子节点 Block 的内容 (content) 作为新的 DOM,重新抽取 Block 并组装 $n+1$ 级 Block 树,如果某个 Block 已经无法再分割,则忽略这个 Block。如此反复,直到 Block 树达到给定分区级别为止。对于页面分块的详细规则见文献[1~3]。图 1 例子页面通过 3 次分区算法迭代过程中产生了图 2 的 Block 树。

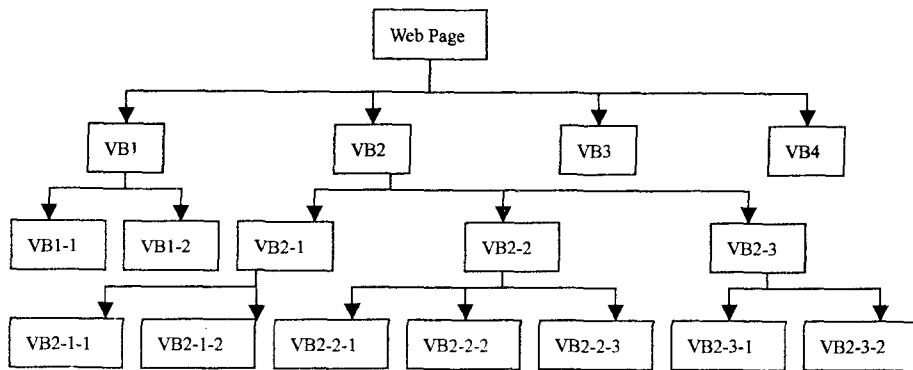


图 2 页面 Block 树示例图

1.3 Block 主题生成算法

为了便于对 Block 进行检索,文献[4~8]给出了为每一个 Block 生成一句主题文字的简单的方法。主题生成算法要求具有唯一性,即相同的内容抽取的主题相同,但主题相同对应的内容则不一定相同。通常可以采用两种算法生成主题:一种是通过字体、颜色和位置等信息查找主题;另一种是利用数据挖掘的方法抽取主题。

1.4 Block 赋权

是什么特征使人们可以用来从不重要的页面 Block 中区别出重要的页面 Block 见文献[9],一般来说,网页的设计者喜欢把最重要的信息放在页面的中心,把导航条放在页面的上方或者左边,并且版权信息放在页面的底端。语义 Block 的重要性可以通过位

置、大小等空间特征来反映;另一方面,块的内容对判断块的重要性也起着很重要的作用。

1) 空间特征。

每一个语义块都被一个矩形框定位在页面上,每个语义块的空间特征可由以下四方面构成: {BlockCenterX, BlockCenterY, BlockRectWidth, BlockRectHeight}。其中,BlockCenterX 和 BlockCenterY 是指语义块的中心点的坐标,BlockRectWidth 和 BlockRectHeight 是指该语义块的宽度和高度。这样的空间特征被称作绝对空间特征。使用页面的绝对空间特征很难在不同的网页之间进行比较,因此用页面的长度与宽度来规范化绝对空间特征,得到相对空间特征的定义如下:

$$\{ \text{BlockCenterX}/\text{PageWidth}, \text{BlockCenterY}/\text{PageHeight}, \text{BlockRectWidth}/\text{PageWidth}, \text{BlockRectHeight}/\text{PageHeight} \}$$

对页面大小规范化以后给人们带来另一个问题,就是有的页面的长度是显示器长度的好几倍,这样的话,很多很重要的语义块在规范化之后,就会挤压在屏幕的上方,同那些不太重要的块如导航条、广告条一样。对于比较长的网页来说,第一屏的信息是非常重要的,应当避免这个问题;宽度规范化则没有相应的问题,因为大多数的网页宽度是不会比屏的宽度大的。

为了解决问题,用固定大小的窗口大小来代替页面高度进行规范化:

$$\text{BlockRectHeight} = \text{BlockRectHeight}/\text{WindowHeight};$$

同样的 BlockCenterY 也需要进行调整:

$$\text{BlockCenterY} =$$

$$\begin{cases} \text{BlockCenterY}/(2 * \text{HeaderHeight}) & \text{if } \text{BlockCenterY} < \text{HeaderHeight} \\ 0.5 & \\ \text{if } \text{HeaderHeight} < \text{BlockCenterY} \\ 0.5 + (\text{PageHeight} - \text{BlockCenterY})/(2 * \text{FootHeight}) & \\ \text{otherwise} \end{cases}$$

其中 HeaderHeight 和 FootHeight 为一个页面头部与底部的高度的预定义常数。

2) 内容特征。

以下 9 个特征用来表示块的内容特征:

$$\{ \text{ImgNum}, \text{ImgSize}, \text{LinkNum}, \text{LinkTextLength}, \text{InnerTextLength}, \text{InteractionNum}, \text{InteractionSize}, \text{FormNum}, \text{FormSize} \}$$

其中,ImgNum 和 ImgSize 表示包含在块中的图像的数量和大小;LinkNum 和 LinkTextLength 表示块中超链接的数目和块中链接文本的长度;InnerTextLength 是指块中文本的单词数目;InteractionNum

和 InteractionSize 是指块中的 < INPUT > 和 < SELECT > 标记的数目和大小;FormNum 和 FormSize 是指块中的 < FORM > 的数目和大小。这些特征都是和重要性相联系的,例如,一个广告通常只包含图片而没有文字,而导航条一般要包含一定数目的超链接。

一般情况下,得到块的重要性通常有两种方法:一种是以经验规则为基础的,根据空间特征和内容特征来得出块的重要性,但是这种方法存在很多问题;现在考虑第二种方法,从例子中学习的方法,通过一定数量的人对一些模块进行预标记,每一个标记的块被表示成 $\langle x, y \rangle$ 的形式,其中 x 为块的特征表示, y 为标注好的重要性。标注好的块的集合被用作训练集 T , 则问题就转化成了发现一个函数 f , 求 $\sum_{(x,y) \in T} |f(x) - y|^2$ 的最小值的问题,如果 y 是离散的就是一个分类的问题,如果是连续的就是一个回归问题。参考文献 [2] 用 RBF 神经网络的方法最小化的常函数 $f^* = \arg \min_f \sum_{i=1}^m \|f(x_i) - y_i\|^2$ 的问题,这样便对 Web 页面中的每一个语义块赋予一个相应常数权值。

2 加权 Block 的主题信息提取算法

算法 1: 基于 Block 的信息提取算法。

Input: pNode is DOM, depth is the depth of pNode in DOM tree, i 为抽取主题数目的参数, ω 为设定的块的权值阈值。

Output: 抽取符合粒度要求的语义 Block 的权值和主题。

ExtractBlockValues(pNode, depth)

1 If depth > MaxDepth

2 Return;

3 End;

4 If VIPS.IsBlock(pNode = True)

5 Bw = BlockWeightGeneration(pNode);

6 Save pNode as Block into BlockPool;

7 If Bw $\geq \omega$

8 For 1 to i

9 ExtracBlockTopic(Block.text _{i});

10 End

11 Else

12 VIPS.GenerateBlock(pNode, depth);

13 ExtractBlockValues(child, depth + 1);

14 End。

算法 2: 页面信息提取算法。

Input: 一个相对规范的网页(或者某个网页的链接)。

Output:按照网页中的块的权重的顺序,从网页中块提取出特定量的主题信息。

ExtractWebValues(Web_i)

1 将一个网页表示成 DOM 树的形式;

2 调用 VIPS 算法,将 Web_i 分块;

3 ExtractBlockValues(pNode,depth);

4 按块中权值的大小,从 BlockPool 中将每块中输出数目为 *l* 的主题;

5 Return 网页中的提取信息。

3 实验结果

抽取国内一些知名的门户网站的部分网页作为模拟测试对象,测试结果如表 1 所示。

表 1 页面 Block 信息的提取结果

来源网站	网页数目	正确率	平均提取时间
新浪	150	97%	0.62s
搜狐	268	96%	0.68s
网易	288	98%	0.59s
腾讯	122	95%	0.62s
新华网	234	96%	0.61s
人民网	168	97%	0.68s

实验测试对象包括了当前的主流的门户网站,经过大量的实验,发现本方法对新闻和 BBS 领域的信息有很好的效果,提取块信息的正确率能够达到 97% 左右,提取一个网页的时间大约在 0.2s 到 0.6s 之间。从实验结果可以看出,对于页面机构和页面风格相对稳定的网站提取效率较高,提取错误产生的原因主要是某些网站的个别网页的页面分区和页面风格与网站的整体不相同。

4 结束语

页面信息提取的方法很多,页面分区的理论早已提出,页面 Block 赋权的工作也有人开展,但是从未有

人将二者联系起来从大量的网页中挖掘主要的页面信息。这是文中的创新之处。在以后的研究工作中,一是提高算法的效率;二是提高页面的预处理能力,规范化 HTML 页面,进而提高算法的准确率;三是对于一些页面结构相对不变的大网站,可以考虑是否将其页面分区和页面 Block 的权值存储起来,那将大大提高网页的信息提取速度。

参考文献:

- [1] Cai D, Yu S, Wen J R, et al. VIPS: a version - based page segmentation algorithm [R]. Microsoft Technical Report, 2003.
- [2] Cai D, Yu S, Wen J R, et al. Block - based Web Search[C]//in 27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, South Yorkshire, UK: [s. n.], 2004.
- [3] Cai D, Yu S, Wen J R, et al. Block - based Link Analysis [C]//in 27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, South Yorkshire, UK: [s. n.], 2004.
- [4] 宋 杰,王大玲,鲍玉斌,等.基于页面 Block 的 Web 档案采集和存储[J].软件学报,2008,19(2):275 - 290.
- [5] 王晓宇,熊 方,凌 波,等.一种基于相似度的主题提取和发现算法[J].软件学报,2003,14(9):1578 - 1585.
- [6] 李晓明,朱家稷,阎宏飞.互联网上主题信息的一种收集与处理模型及其应用[J].计算机研究与发展,2003,40(12):1667 - 1671.
- [7] 张 敏,高剑锋,马少平.基于链接描述文本及其上下文的 Web 信息检索[J].计算机研究与发展,2004,41(1):221 - 226.
- [8] 宋聚平,王永成,尹中航,等.面向主题的网页搜索系统[J].上海交通大学学报,2003,37(3):401 - 403.
- [9] Song Ruihua, Liu Haifeng, Wen Ji - Rong, et al. Learning Block Improtance Models for Web Pages[C]//the 13th international conference on World Wide Web. [s. l.]:ACM,2004.

(上接第 196 页)

- [4] 卢雅琴,邹凌超.基于数学形态学的车牌定位方法[J].计算机工程,2005,31(3):224 - 226.
- [5] 韩丽萍,尹王保,李月娥.一种有效的滤波尺度自适应调整边缘检测方法[J].计算机工程与应用,2005(11):70 - 72.
- [6] Zheng D, Zhao Y Z, Wang H X. An efficient method of license plate location[J]. Pattern Recognition Letters, 2005, 26(15): 2431 - 2438.
- [7] 张 玲,刘 勇,何 伟.自适应遗传算法在车牌定位中的应用[J].计算机应用,2008,28(1):184 - 186.
- [8] 李庆庆,张燕平.基于模糊边缘检测算法的车牌定位[J].计算机技术与发展,2006,16(12):7 - 8.
- [9] 徐 慧. Visual C++ + 数字图像实用工程案例精选[M].北京:人民邮电出版社,2004.
- [10] 郭 亚,王水波.基于灰度图像的车牌定位算法研究与实现[J].现代电子技术,2008(2):137 - 139.
- [11] 章毓晋.图像分割[M].北京:科学出版社,2001.