

文章编号:1001-9081(2010)03-0818-03

## 基于局部标签树匹配的改进网页聚类算法

李 睿,曾俊瑀,周四望

(湖南大学 软件学院,长沙 410082)

(rj\_rli@hnu.edu.cn)

**摘要:**Web信息抽取中需要对目标网站的网页进行聚类分析,以检测并生成信息抽取所需的模板。传统的基于DOM树编辑距离的网页聚类算法不适合文档对象模型(DOM)树结构复杂的动态模板网页,提出了一种基于局部标签树匹配的改进网页聚类算法,利用标签树中模板节点和非模板节点的层次差异性,根据节点对布局影响的大小赋予节点不同的匹配权值,使用局部树匹配完成对网页结构相似性的有效计算。实验结果表明,改进的算法较传统的基于DOM树编辑距离的网页聚类算法,在对采用模板生成的动态网页进行聚类分析时具有更高的准确率,且时间复杂度低。

**关键词:**Web信息抽取;网页聚类;树编辑距离;局部标签树匹配

中图分类号: TP301 文献标志码:A

## Improved Web page clustering algorithm based on partial tag tree matching

LI Rui, ZENG Jun-yu, ZHOU Si-wang

(Software School, Hunan University, Changsha Hunan 410082, China)

**Abstract:** In the process of Web information extraction, Web pages on the target websites should be clustered in order to detect and generate templates that are used to extract required information. Traditional page clustering algorithm based on DOM tree edit distance is not suitable for the complex Document Object Model (DOM) tree structure pages created from dynamic templates. In this paper, an improved Web page clustering algorithm was proposed based on partial tag tree matching. In the proposed algorithm, the appropriate weights were assigned to the nodes according to their effects on the layout of Web pages and the level difference between template nodes and non-template nodes. After that, the structure similarity between Web pages was computed efficiently based on partial tree matching approach. Compared with the traditional algorithms, the experimental results show that the proposed algorithm is of higher accuracy in clustering dynamic Web pages and lower computing complexity.

**Key words:** Web information extraction; Web page clustering; tree edit distance; partial tag tree matching

### 0 引言

随着Internet的快速发展,如何利用Web上的海量数据资源提供更好的信息服务开始成为研究的重点。Web信息抽取通过将互联网上半结构化的数据抽取出并转化为结构化数据存入数据库中,从而能够利用数据库领域的技术高效地管理Web上的数据<sup>[1]</sup>。然而,由于Web网页资源的异构性以及没有一个严格的规则来规定如何构造HTML和声明Web页面的结构,使得设计一个通用的抽取规则在网页中准确定识别所需的信息成为一项极为复杂的任务。文献[2]提出一种基于直推式支持向量机的Web信息抽取方法,直接从分类的角度抽取Web信息;文献[3]对Web特征的适应性进行分析,引入了相互关联的三层规则,并用XML查询语言XQuery来进行表达。动态网页与静态网页在生成方式上存在很大不同,动态网页由模板生成,检测出网页模板,可实现高效精确的信息抽取。

目前,动态网页的模板检测主要通过将相同结构的网页自动聚类,从聚类后的网页簇中自动泛化生成高效准确的抽取模板。因此,网页自动聚类作为实现数据精确抽取的前提,

是Web信息抽取流程中极为重要的环节。文献[4]通过采用观察网页URL规则的方法对网页进行聚类,但随着动态URL的流行,这种方法的准确性非常有限。文献[5]研究了基于网页DOM树编辑距离的聚类方法,在对结构简单的新闻网页进行实验时取得了不错的效果,但该方法的计算复杂度较高,而且对于网页DOM树较为复杂的动态网页进行聚类时准确率不高。简单树匹配(Simple Tree Matching, STM)算法<sup>[6]</sup>在现有的网页聚类过程应用较为广泛,但该算法不允许节点进行交换操作和跨层操作<sup>[7]</sup>。

本文在分析DOM树的网页层次结构特征的基础上,利用HTML标签树中模板节点和非模板节点的层次差异性,提出了一种基于局部标签树匹配的网页聚类算法,所提出算法根据节点对布局影响的大小赋予节点不同的匹配权值,使用局部树匹配完成对网页结构相似性的计算。实验结果表明所提出的算法相对于STM具有如下良好性质:1)计算时间复杂度低;2)对于结构复杂的动态模板网页的聚类精确度高。

### 1 基于局部树匹配的网页聚类算法

网页的DOM树结构在一定程度上反映了页面的内容安

收稿日期:2009-09-24;修回日期:2009-11-12。 基金项目:湖南省自然科学基金资助项目(09JJ3123)。

作者简介:李睿(1975-),男,湖南汨罗人,讲师,博士研究生,主要研究方向:Web信息抽取、Web信息检索; 曾俊瑀(1987-),男,湖南怀化人,硕士研究生,主要研究方向:Web信息检索; 周四望(1971-),男,湖南岳阳人,副教授,博士,主要研究方向:信息处理。

排以及内容之间的关系,而且基于模板生成的动态网页的 DOM 树表示往往具有相对固定的结构,如图 1 所示。当采用 DOM 树结构来表示网页时,网页之间的相似性可以很自然地采用树编辑距离(Tree Edit Distance)进行衡量<sup>[6]</sup>。树编辑距离计算的原理与字符串编辑距离的计算类似。

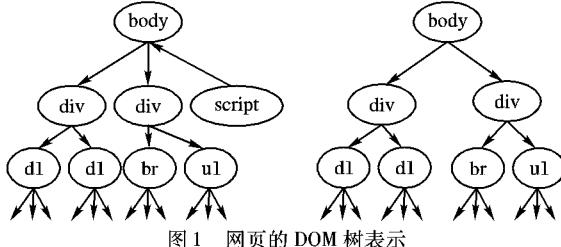
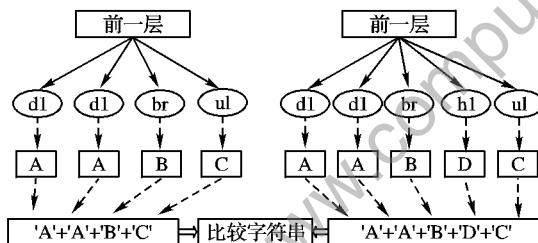


图 1 网页的 DOM 树表示

### 1.1 网页标签树相似性计算

基于 HTML 标签的字符串编辑距离算法不适合直接用于比较两个 Web 文档的相似性,因为它没有考虑文档的结构,而 DOM 树恰恰是最能表达 HTML 标签的分布情况进而表达网页的布局。通过观察 DOM 结构可以发现:同一模板构造的网页自顶向下差异度逐渐增大,换言之,从根节点到叶子节点的路径中,网页 DOM 树节点对布局的影响逐渐减小。结合 DOM 树表示网页树型结构的优点和字符串编辑距离计算的高效性,本文将 DOM 树的每一层节点的 HTML 标签连接成串,计算两棵 DOM 树对应层所连接的串的距离,并将每一层的字符串编辑距离加权相加所得的结果作为最终两个网页的编辑距离,并以此来衡量两个 Web 文档的相似性。为进一步提高算法效率,在将标签连接成串的过程中,先将标签压缩为字符,以此减小连接串的长度。比较两棵 DOM 树第  $k$  层节点标签构成的字符串如图 2 所示,通过对前  $N$  层节点的标签字符串比较,最终可以求得网页标签树的相似度。

图 2 比较两棵树的第  $k$  层节点标签构成的字符串

### 1.2 基于局部标签树匹配的网页聚类算法

由于 DOM 节点的层次越大,对树结构的影响越小,本文为每层所得的字符串编辑距离附上一个权重  $W_i$ ,以体现层次对结构的影响,权重为层次  $i$  的函数。

**定义 1** DOM 树中层次节点字符串编辑距离的权重为:

$$W_i = \frac{1}{(i+1)^x} \quad (1)$$

其中: $i$  为 DOM 树的层次编号,根节点为第 0 层; $x$  为权重调节参数,在实验中,对多个  $x$  值进行了比较,以确定最合适的  $x$  值。

另外,DOM 树的节点总数也会对字符串编辑距离产生影响,当某一层的节点数目较多时,得出的编辑距离也较大,因而包含较多节点数目的网页相似性会比包含较少节点树的网页相似性小,这会在一定程度上影响网页聚类的准确率。为了抵消节点数目所带来的负面影响,为第  $i$  层的字符串编辑距

离加上一个影响因子  $\delta_i$ 。

**定义 2** DOM 树中层次节点的影响因子  $\delta_i$  如下:

$$\delta_i = \frac{1}{length(i)} \quad (2)$$

其中: $i$  为节点所在层次; $length(i)$  为第  $i$  层节点字符串的长度。

由于层次较低的节点对 DOM 树布局影响较小,且在同一模板的网页集合中,低层次的节点会因网页内容而差异较大,因此,本文对 DOM 树的前  $N$  层( $N < 5$ ,可根据实验结果来确定)进行分析计算。算法实验表明,只对前  $N$  层进行分析计算不但提高了时间效率,还避免了许多低层节点带来的噪声干扰,提高了聚类的精度。

**定义 3** 基于 HTML 标签树字符串编辑距离的网页相似距离:

$$Distance = \sum_{i=1}^N SED(layer1[i], layer2[i]) \times W_i \times \delta_i \quad (3)$$

其中: $layer1[i]$  和  $layer2[i]$  分别表示由两棵不同 DOM 树第  $i$  层节点标签构造的字符串; $SED$  为求字符串编辑距离的函数; $N$  为计算的 DOM 最大层次。

本文利用标签树中模板节点和非模板节点的层次差异性,根据节点对布局影响的大小赋予节点不同的匹配权值,提出了一种基于 HTML 标签树局部匹配的改进网页聚类算法。算法伪代码如下:

```
Algorithm: DSED(T1: Tree, T2: Tree, N)
begin
for (i = 1; i < N; i++)
  S1 := T1 中第 i 层节点构造的字符串
  S2 := T2 中第 i 层节点构造的字符串
  layer1 := T2S(S1)
  layer2 := T2S(S2)
  lengthi := (|layer1| + |layer2|) / 2
  sigmai := 1 / lengthi
  Wi := 1 / (i + 1)^x
  singleDistance = SED(layer1, layer2)
  distance += SED(layer1, layer2) * Wi * sigmai
end for
return distance
end
```

算法最终返回两个网页的距离(即差异度),其中  $N$  为 DOM 树的层次, $T2S$ (Tags to String) 将 DOM 节点的标签字符串连接成字符串,以提高字符串编辑算法的效率。易知该算法的时间复杂度为  $O(|S1| \times |S2|)$ 。

## 2 实验结果及分析

对于精确度的评估标准,本文采用聚类算法通用的 *F-Measure* 来评估。*F-Measure* 的值在 0 到 1 之间,值越大表示聚类准确度越高。

在实验数据方面,对五个在线商务网站:www.ebay.com、www.dangdang.com、www.amazon.cn、youa.baidu.com、www.paipai.com,和五个博客网站:www.blogbus.com、blog.sina.com.cn、www.williamlong.info、www.cppblog.com、hi.baidu.com 分别进行了实验。

考虑到实验的数据覆盖范围,商务网站涵盖了国内外知名的商务网站,而博客网站包括了较为主流的门户博客与其

他类型的个人博客。对于每个网站分别抓选取了 200 个网页进行了实验,同一个网站的 200 个网页都是爬取的具有同一个模板的网页。由于商务网站和博客网站的构造有着非常明显的区别,对两类网站分别进行实验,可以更好地证明本文所提出的算法的适用范围,避免可能存在的过适应问题。

## 2.1 实验 1

根据定义 1,本文对多个权重参数  $x$  值进行实验,以确定最合适的是  $W_i$  值,实验结果如图 3 所示。实验结果表明,权重调节参数取 1 附近的值时, $F\text{-Measure}$  值较高。为了计算简便,本文取权重调节参数等于 1。

根据定义 3,本文对多个 DOM 层次数  $N$  值进行实验,对实验结果取平均值,以确定最合适的是 DOM 树层次,在本实验中,取定义 1 中的权重调节参数为 1。实验结果如图 4 所示。实验表明, $N$  为 4 时, $F\text{-Measure}$  值较高。

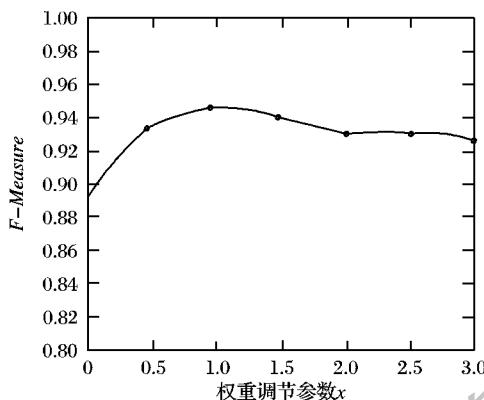


图 3 不同权重调节参数  $x$  的  $F\text{-Measure}$  值

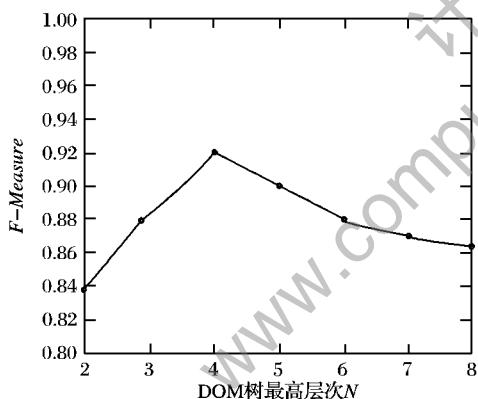


图 4 不同 DOM 树层次  $N$  值的  $F\text{-Measure}$  值

## 2.2 实验 2

在实验 1 的最优条件下,应用新的 DSED 算法和传统的 STM 聚类算法进行比较。为了规范结果,将网页相似距离的值映射到  $[0, 1]$ ,并取了多个阈值来分别计算网页聚类结果,通过用阈值和网页间相似距离进行比较来判断网页是否相似。在图 5、6 中,取 X 轴阈值的值域为  $[0.05, 0.8]$  间隔为 0.05,对多组实验数据取平均值。

由聚类效果比较图可知,对于电子商务模板网页,STM 算法和 DSED 算法在不同阈值下的  $F$  值基本相近;但对于博客模板网页,DSED 算法在不同阈值下的  $F\text{-Measure}$  值上明显高于 STM 算法。因此,总的来说,基于 DSED 算法的网页聚类具有更好的适应性,对于允许自定义模块的动态模板网页如博客网页,基于局部树匹配的网页聚类算法优于一般基于简单树匹配的网页聚类算法。

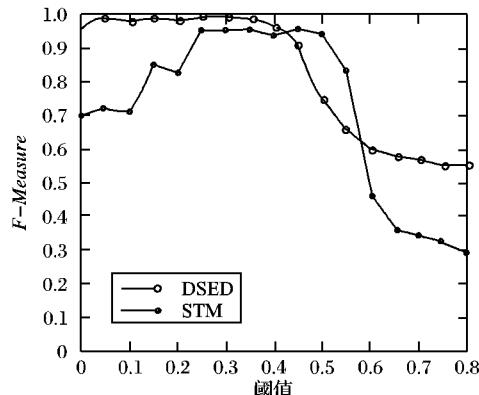


图 5 STM 和 DSED 算法比较(博客网页)

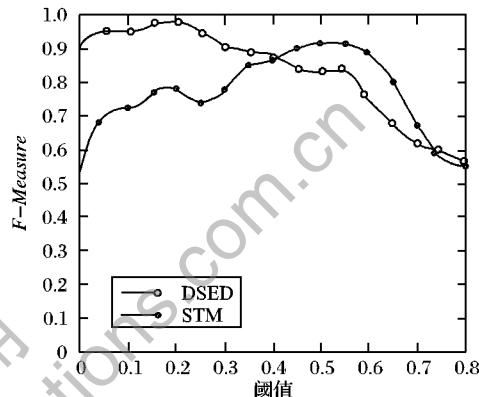


图 6 STM 和 DSED 算法比较(电子商务网页)

## 3 结语

本文提出了一种基于局部标签树匹配的改进网页聚类算法,该算法利用标签树中模板节点和非模板节点的层次差异性,根据节点对布局影响的大小赋予节点不同的匹配权值,使用局部树匹配完成对网页结构相似性的有效计算。实验结果表明,新的算法具有更好的聚类效果和适应性,从而使得算法更符合 Web 信息抽取的实际应用需求。

## 参考文献:

- [1] FLORESCU D, LEVY A, MENDELZON A. Database techniques for the world-wide Web: Survey [J]. SIGMOD Record, 1998, 27(3): 59 - 74.
- [2] 肖建鹏, 张来顺, 任星. 直推式支持向量机在 Web 信息抽取中的应用研究[J]. 计算机工程与应用, 2009, 45(2): 147 - 149.
- [3] 支宗良, 陈少飞. 一种基于 XQuery 的优化 Web 信息抽取方法 [J]. 计算机应用, 2008, 28(1): 152 - 154.
- [4] CRESCENZI V, MECCA G, MERALDO P. Wrapping-oriented classification of Web pages [C]// Proceedings of the 2002 ACM Symposium on Applied Computing. New York: ACM Press, 2002: 1108 - 1112.
- [5] REIS D C, GOLGHER P B, SILVA A S, et al. Automatic Web news extraction using tree edit distance [C]// Proceedings of the 13th International Conference on World Wide Web. New York: ACM Press, 2004: 502 - 511.
- [6] ZHAI Y, LIU B. Structured data extraction from the Web based on partial tree alignment [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(12): 1614 - 1628.
- [7] YANG W. Identifying syntactic differences between two programs [J]. Software-Practice and Experience, 1991, 21(7): 739 - 755.