第 36 卷 第 3 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.3 Computer Engineering February 2010
·博士论文· 文章编号：1000—3428(2010)03—0001—03 文献标识码：A 中图分类号：TP391

基于 Web 页面结构和主色调的聚类算法

赵涓涓，陈俊杰，李元俊
(太原理工大学计算机与软件学院，太原 030024)

摘 要：针对目前 Web 聚类准确率不高的问题，提出一种基于 Web 页面链接结构和页面中图片主色调特征的聚类算法。通过分析 Web 页面中的链接结构和 Web 页面中所显示图片的主色调来比较页面之间的相似度，对 Web 站点中的 Web 页面进行聚类。聚类过程兼顾 Web 页面结构和页面的主要色彩特征。系统实验结果表明，该算法能有效提高聚类的准确性。

关键词：聚类；Web 挖掘；链接结构；主色调

Clustering Algorithm Based on Web Pages Structure and Dominant Color

ZHAO Juan-juan, CHEN Jun-jie, LI Yuan-jun
(College of Computer and Software, Taiyuan University of Technology, Taiyuan 030024)

【Abstract】Aiming at the problem that the efficiency is low in Web clustering, this paper proposes a clustering algorithm based on linkage structure and the character of the dominant color on Web pages. It compares the similarity between Web pages by analyzing the linkage and the dominant color on them. It can cluster the Web pages on Web sites. In this procedure, the clustering has both the structure and the main character of tone. Experimental results of the system prove that it has made the clustering become more efficient and it has improved a lot than before.

【Key words】clustering; Web mining; linkage structure; dominant color

1 概述

Web 数据是异构的、非结构化的、动态变化的，这就需要将 Web 页面分类(聚类)，对不同的分类设计不同的分装器(Wrapper)，并进行信息抽取，最终得到的结构化数据进行数据挖掘。

目前该领域的研究主要有基于文本内容的Web页面主题聚类[1]和基于 Web 页面结构的聚类[1]；前者仅考虑 Web 页面的内容信息，聚类时间效率低，而且它不关心 Web 页面的结构，不利于分装器的设计；后者虽然聚类效率较高，但是它没有利用不同页面在页面内容方面的个性化特征(如文本内容、色彩、样式等)致使聚类结果不是很准确。

本文设计一种兼顾 Web 页面结构特征和 Web 页面内容的聚类方法(CABSW)，该方法巧妙地利用了 Web 页面的超链接结构[2]的规律性和 Web 页面的主色调特征的相似性，通过分析 Web 站点中一部分(而不是全部页面)但是具有代表性的页面得到站点分类模型，从而为设计分装器提供训练样本，保证整个数据挖掘过程顺利进行。实验结果表明，该方法能在聚类的效率和准确率方面达到较好的效果。

2 聚类方法分析

2.1 Web页面特征剖析

根据对 Web 页面的深入分析，发现其具有更适合于聚类算法的结构。

2.1.1 反映Web页面的链接结构

利用 Web 页面解析技术可以发现，“网页链接标签路径[3](link-path)集合”可作为 Web 页面的特征项和判别 Web 页面相似度和聚类的主要依据；图 1 是一个 Web 页面生成的 DOM 树型结构，该页面的 link-path 集合为{HTML-TABLE-UL-LI, HTML-TABLE-TD-TR-TD}。

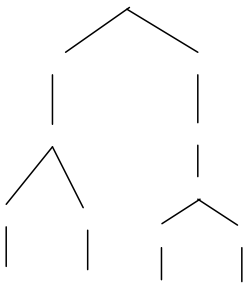

图 1 DOM 树形结构

2.1.2 相同link-path下的相似性链接

如图 1 中的 DOM 树所示，该 Web 页面中的 link-path {HTML-TABLE-TR-TD}下面包含有 2 个链接{url1, url2}，认为这 2 个链接极有可能指向同一类别的 Web 页面，这是因为格式良好的 Web 页面的样式和布局都是很有规律的，超链接布局如图 2 所示。

基金项目：国家自然科学基金资助项目(60773004)；山西省自然科学基金资助项目(2006011030, 2007011050)

作者简介：赵涓涓(1975－)，女，讲师、博士研究生，主研方向：智能信息处理，情感计算，数据挖掘；陈俊杰，教授、博士、博士生导师；李元俊，硕士研究生

收稿日期：2009-07-03 **E-mail：**zh_juanjuan@126.com

涨幅排行 更多>>

排名	板块	涨跌幅	领涨股
1	券商	2.87%	国金证券
2	工程建筑	2.23%	重庆路桥
3	机械	2.09%	林海股份
4	钢铁	1.85%	大连金牛
5	运输物流	1.82%	申通地铁

图 2　超链接布局

图 2 界面来自 http://finance.sina. com.cn/stock/，其中超链接的 link-path 相同，而且指向同类页面，甚至指向同一个页面，只是传递的参数不同。因此，进行聚类时可以先比较同一个区域中的链接指向的 Web 页面，因为它们相似度很高，所以可以提高聚类的效率。

2.1.3　可提供页面内容信息的Web页面部分标签

多数 Web 站点虽然页面结构基本相似，但是不同主题的 Web 页面主色调都不相同。如网易网站的新闻板块的主色调为蓝色、体育板块的主色调为红色而音乐板块的主色调为橙色；所以在做 Web 聚类时，充分考虑页面主色调提供的分类信息，有利于提高 Web 聚类的准确性。

2.2　Web页面主色调相似度

Web 页面中包含有很多的图片，这些图片的色调风格决定了整个 Web 页面的色调风格[4]，而整个 Web 页面的色调决定了该页面所属的信息板块。因此，可以提取每个 Web 页面中最大的 10 个的图片，在图片的 10 个固定坐标点上各抽取出 5×3 个像素点的区域，如图 3 所示。计算这些区域的颜色直方图，如图 4 所示。

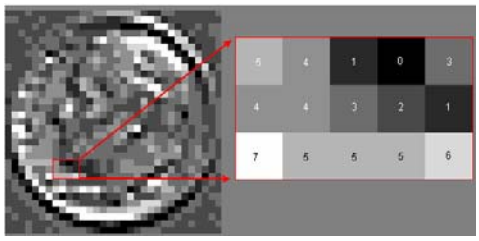

图 3　像素抽取

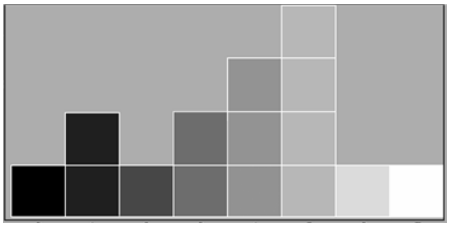

图 4　颜色直方图

将这些区域的颜色直方图合并，得出整个图片的颜色直方图，这样就可以得出图片的主色调；不同的 Web 页面通过比较颜色直方图的相似度(即图片主色调的相似度)来判断 Web 页面之间的相似度，从而为聚类提供依据。

3　聚类算法实现

本系统中实现的聚类方法是基于Web页面链接结构的主色调的基础上，使用 Web 结构信息和内容信息相结合来进行聚类的方法。采用 DOM 和 Xml 等技术协助完成聚类系统，其重要意义在于对 Web 站点的页面进行聚类，利用聚类的结果可以训练 Wrapper，然后对 Web 页面进行数据抽取和数据挖掘。

3.1　算法描述

根据上述 Web 页面结构和网页中主色调的特性，基于 Web 页面结构和主色调的聚类算法描述如下：

Seed：链接种子，如：www.qq.com；

Qcol：存放链接集合的队列；

Maxw：规定从链接集合中返回的页面数上限；

Tol：算法要分析的 Web 页面总数，输出 *WebM*。

聚类后的页面类别集合，也即 Web 站点类别模型

Step1　算法开始；设定初始相关数值与参数。

Step2　设定 *Qcol* 初值,初始链接种子，存放队列初始置空。将 *Seed* 页面中的所有链接集合返回给 *Qcol* 队列，相同 link-path 的链接被放入同一个链接集合；Set *Qcol*={*Seed*}。

Step3　弹出 *Qcol* 中最高优先权的链接集合 *listm*,返回被 *listm* 指向的最多 *Maxw* 个页面，放入集合 *L* 中，同时保存 Web 页面的图片直方图信息，为 *Clr* 函数的计算做准备。do Until((*Qcol*!=Null)&&*Tol*>0)。

Step4　修改页面最大值，其中 *count*() 为集合 *L* 的势：*Tol*=*Tol*-*L.count*()。

Step5　根据页面 link-path 集合的不同将 L 中的 Web 页面分组为 *G*1*G*2…*Gk*，并根据 *G* 中页面数的不同将 *G* 从大到小排序：Set *F*=(*G*1, *G*2,…,*Gk*)*L*。.

Step6　进行相似度计算 *DISTC*(*Gi*, *Gj*)，如果在距离范围之内则添加到当前组，否则产生一个新的分组。循环依次比较，直到与全部组比较结束。

Step7　保存新的 *Clr* 的值。

Step8　从 W 中添加新的链接到队列 *Qcol* 中，如果队列不为空，转到 Step5。

Step9　保存分组数据，算法结束

程序实现的 Java 代码如下：

```
for (i=1, i<k+1, i++)
{for (j=k, j>i, j--)
    {if (DISTC(Gi,Gj)<0.18)
//DISTC 为 DISTC 函数，0.18 为实//验过程得出的合理阈值
        {add Gj to Gi;
            set Gj=Null;}}}
  Set T=(C1,…,Cg)//C 为 G 的合并结果
  for each C ∈ T
  {
   if (Clr(C,WebM))>0.4) //Clr 函数在算法解释中说明
            WebM={C}∪WebM;
   else
     WebM=WebMm;
  //WebMm 为将 C 归入 WebM 后的页面类别集合
  }
   add to Qcol the new links collections from L; }
```

3.2　算法解释

文献[3]对上述算法做过阐述；本文在单纯的聚类算法的基础上提出了新的网页优先搜索策略以提高算法的运行效率，提出了新的 Web 站点类别模型生成方法，并且在聚类过程考虑到了 Web 页面中页面主色调特性,以提高聚类准确率。

3.2.1　链接集合队列 Qcol 的优先权

队列 *Qcol* 中链接集合的优先权通过以下启发式来计算：

(1)势最大的链接集合优先,因为同一链接集合 *listm* 内的链接所指向的页面很有可能属于同一类别 *C*，所以集合的势越大对于 *C* 的支持度就越高。

(2)势非常小的链接集合优先，比如：势为 1 的链接集合极有可能指向新的未曾访问过的区域。

因为将要访问的 Web 页面数即 *Tol* 是一定的，所以上述 2 种策略会得到截然不同的效果：策略(1)会得到包含页面数很大的类别，策略(2)则会得到页面数很少的类别，但是类别数量会很多。

本文基于上述综合考虑，提出的方法如下：首先是按照从大到小的顺序分析势大于 15 的链接集合；分析完势大于 15 的链接集合之后再优先分析势为 1 的集合；最后分析其他集合，直到 *To*≤0，分析过程结束。

3.2.2 *DISTC* 函数

DISTC 函数是计算 2 个 Web 页面相似度距离，公式如下：

$$DISTC(G_i,G_j)=\frac{|(G_i-G_j)\cup(G_j-G_i)|}{|G_i\cup G_j|}$$

将 link-path 集合相同的 Web 页面分组为 *G*1*G*2…*Gk*，此处的 *Gi*, *Gj* 表示 Web 页面组的 link-path 集合，根据集合的并运算和差运算来计算 Web 页面的相似度距离。

3.2.3 *Clr* 函数：

Clr 函数是通过计算 Web 页面间颜色直方图的相似度对聚类过程进行约束[5]，如果不使用 *Clr* 函数约束聚类过程的话，Web 页面就会按照页面链接结构来聚类，这样的结果没有考虑到页面提供的内容特性，聚类结果不够准确；*Clr* 函数的基本思路是：

(1)将网页集合 *C* 做为一个新的网页类别计算与其他已有的网页类别的相似度，得到 *Clr* 函数的值。

(2)网页集合 *C* 与其他已有网页类别相似度计算公式为

Clr(*C*,*C*0)=*DIST*(*C*,*C*0)+*CDIST*(*C*,*C*0)；

其中，*CDIST* 为页面颜色直方图相似度计公式(计算思路与 *DIST* 函数同)；*C*0 为某个已有的网页类别；

(3)如果有多个 *C*0 使 Clr(*C*,*C*0)<0.4 则选取 *Clr* 值最小的 *C*0，将新类别 *C* 归入 *C*0；

(4)如果所有的 *C*0 都使 *Clr*(*C*,*C*0)>0.4 则将 *C* 做为一个新的、独立的类别加入 *WebM*。

这样在按照 Web 页面结构聚类的同时考虑到了 Web 页面内容特性，聚类结果更有说服力。

4 系统实现及结果分析

4.1 系统实现环境

本系统是在 Windows2000 操作系统下，主要开发语言为 Java 语言，开发平台为 Eclipse＋Tomcat＋JDK＋SQL server。系统界面如图 5 所示。

图 5 聚类系统输入界面

4.2 用例及结果

本文系统数据源选取了 4 个较大型网站的子站点，它们结构规整、链接丰富、具有代表性，具体实验用例见表 1。

表 1 实验用例

站名	URL	Pages
SINA	http://spots.sina.com.cn/	213
BAIDU	http://mp3.baidu.com/	230
QQ	http://www.qq.com/	160
HAO123	http://www.hao123.com/	170

本文使用 F-measure[3]对实验结果进行评测：

$$F_i=\max_{j=1,2,\cdots,m}\left(\frac{2P(C_i,\hat{C}_j)R(C_i,\hat{C}_j)}{P(C_i,\hat{C}_j)+R(C_i,\hat{C}_j)}\right)$$

其中，*P* 表示准确率(Precision)；*R* 表示召回率(Recall)：

$$P(C_i,\hat{C}_j)=\frac{|C_i\cap\hat{C}_j|}{|\hat{C}_j|},\quad R(C_i,\hat{C}_j)=\frac{|C_i\cap\hat{C}_j|}{|C_i|}$$

其中，$C_1,C_2,C_i,\cdots,C_n$ 表示应该得到的正确类别；$\hat{C}_1,\hat{C}_2,\hat{C}_j,\cdots,\hat{C}_m$ 表示实验所得的类别；$|C_i\cap\hat{C}_j|$ 表示出现在 $\hat{C}_j$ 中的属于 C_i 的页面数。

综合所有类别，使用以下公式对系统进行评测：

$$F=\sum_{i=1}^{n}F_i\cdot\frac{|C_i|}{\sum_{p=1}^{n}|C_p|}$$

通过 F-measure 评测后得到本系统和仅基于 Web 结构的聚类(常使用 K-means 算法)的比较结果见表 2。

表 2 评测结果比较

站名	本文方法		传统方法	
	F 值	准确率/(%)	*F* 值	准确率/(%)
SINA	0.85	68	0.81	56
BAIDU	0.77	61	0.74	54
QQ	0.88	71	0.77	63
HAO123	0.67	56	0.66	51

可见，本文提出的聚类方法对提高页面聚类的性能有很好的作用，尤其在查准率方面，同时比较站点中页面的结构情况，页面结构相对简单的站点，*F* 值较高，准确率也较高，聚类结果更理想，比如：BAIDU 的站点页面结构比 HAO123(一个提供网址大全的页面，站点中提供了很多进入其他网站的链接)简单得多，对应的 *F* 值前者也要高得多。

5 结束语

本文介绍了一种基于 Web 页面链接结构和 Web 页面主色调的聚类方法，采用 DOM 和 Xml 等技术完成实验，并改进了原有的 Web 页面聚类方法，使用 Web 结构信息和主色调相结合来进行聚类，结果更有说服力，提高了聚类的准确率。本文对 Web 站点的页面进行聚类，利用聚类的结果可以训练 Wrapper，然后利用 Wrapper 对 Web 页面进行数据抽取和数据挖掘。本文的聚类方法对通过动态提交而生成的页面处理效果还不理想，这是下一步的研究方向。

参考文献

[1] 刘远超，王晓龙. 文档聚类综述[J]. 中文信息学报，2006，20(3)：55-62.

[2] Menczer F. Lexical and Semantic Clustering by Web Links[J]. Journal of the American Society for Information Science and Technology, 2004, 55(14): 1261-1269.

[3] Crescenzi V, Merialdo P, Missier P. Clustering Web Pages Based on Their Structure[J]. Data & Knowledge Engineering, 2005, 54(3): 279-299.

[4] 冯 霞，黄亚楼. 基于色彩主特征的快速图象检索[J]. 数据采集与处理，2005，20(2)：198-201.

[5] 石 林. 基于对象的 Web 图像检索研究[D]. 济南：山东师范大学，2007.

编辑 金胡考