# 一种通用论坛信息提取方法

刘 锐[1,2],谭文韬[1,2],付园斌[1,2],王 红[1,2,3]

[1](山东师范大学 信息科学与工程学院,济南 250014)
[2](山东省分布式计算机软件新技术重点实验室,济南 250014)
[3](山东师范大学 生命科学研究院,济南 250014)
E-mail: wanghong106@163.com

**摘 要:** 网络论坛的分类和正文提取是网络数据挖掘的一项重要技术. 传统的网页分类方法没有考虑到论坛网址的结构特性,以内容特征为根据,易受噪声影响,效率较低,难以满足通用性的需求. 传统的正文提取方法以文本密度和布局结构为依据,忽视了论坛内容的语义信息,难以从多样化的论坛中有效提取正文. 本文提出基于网址结构的聚类方法(Universal Resource Locators´ Structure Clustering,USC)以及基于词汇关键程度的关键词打分筛选方法(Keyword Scoring Filter,KSF). 两种方法仅需要对数据集中的少量样本进行解析,提取出通用规则,便可满足大规模提取的需要. 实验验证,在相同测试集下,USC 方法的 F 值较传统分类方法高 18.99%,KSF 方法的准确率较传统正文提取方法高 18.46%,适合大规模论坛提取作业.

**关 键 词:** 信息提取; 网址结构; 内容关键度; 聚类分析
**中图分类号:** TP301 **文献标识码:** A **文 章 编 号:** 1000-1220(2018)07-1398-07

## General Method of Forum Information Extraction

LIU Rui[1,2],TAN Wen-tao[1,2],FU Yuan-bin[1,2],WANG Hong[1,2,3]

[1](School of Information Science and Engineering,Shandong Normal University,Jinan 250014,China)
[2](Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology,Jinan 250014,China)
[3](Institute of Life Sciences,Shandong Normal University,Jinan 250014,China)

**Abstract:** The classification and information extraction of online forums are two important technologies of online data mining. The traditional web page classification methods do not take the structure features of their URLs into full account. They are often based on the characteristics of the content,therefore,they are susceptible to noise,of low efficiency and they can't meet the needs of versatility. The traditional information extraction methods are based on text density and layout structure,ignoring the semantic information of the content. They are difficult to extract the content from a variety of forums effectively. This paper proposes a clustering method based on URLs´structure(USC) and a filter method based on keyword scoring(KSF). Both methods only need to analyze a small number of samples in the data set and extract general rules to meet the demand of large-scale extraction. In the same data set,the F value of the USC method is 18.99% higher than that of the traditional classification method,and the accuracy of the KSF method is 18.46% higher than that of the traditional information extraction method.

**Key words:** information extraction; URL structure; content significance; cluster analysis

## 1 引 言

随着互联网的不断发展,人们在互联网中进行交流的方式也在不断增多,互联网论坛从网络开始推广普及之时起便已成为人们在网络中交流和分享经验的主要平台. 据文献[1]的中国互联网络信息中心发布的第 38 次《中国互联网络发展状况统计报告》显示,截至 2016 年 6 月,中国互联网论坛和 BBS 用户达 1.08 亿人,占网民总量的 15.2%. 然而在当下的网络论坛中,以广告为主的各类无关信息充斥在论坛中,这些网页噪声对信息检索和用户体验都会带来极大的不便. 因此,如何有效消除网页噪声,提取出论坛主题帖正文内容,依然是当下研究的重要课题之一. 传统的基于文本密度的正文提取方法[2-4]没有充分考虑到网页中噪声的影响,将网页源码中文本的长度作为判别正文的依据,使得其算法难以被有效应用到正文内容长度跨度大,网页噪声与网页正文混杂交错的网络论坛中.

为了系统地解决网络论坛主题帖正文提取的问题,本文从主题帖页面识别和主题帖正文提取两个方面入手,分别进行解决. 两种方法均是对数据集中的样本数据进行处理,利用解析所得的规则完成后续分类和提取步骤,保障准确度和执行效率. 实验表明: 本文定义的网址相异度函数适合描述网址

收稿日期: 2017-06-12 收修改稿日期: 2017-08-18 基金项目: 国家自然科学基金项目(61672329, 61373149,61472233,61572300, 81273704)资助; 山东省科技计划项目(2014GGX101026)资助; 山东省教育科学规划项目(ZK1437B010)资助; 山东省泰山学者基金项目(TSHW201502038,20110819)资助; 山东省精品课程项目(2012BK294, 2013BK399, 2013BK402)资助; 大学生创新创业项目资助. 作者简介: 刘 锐,男,1997 年生,研究方向为数据挖掘、机器学习; 谭文韬,男,1997 年生,研究方向为机器学习; 付园斌,男,1996 年生,研究方向为数据挖据; 王 红,女,1966 年生,博士,教授,博士生导师,研究方向为数据挖掘、复杂网络.

间的差异程度; USC 方法与 KSF 方法均具有极强的通用性，适合大规模提取; 两种方法在准确度上均明显优于传统方法，且易于扩展和自定义.

本文其余章节安排如下: 第二节介绍信息提取领域的相关工作进展，第三节对两种方法进行理论介绍和阐释，第四节验证实验的结果和分析，第五节是本文的总结和对未来研究方向的展望.

## 2　相关工作

在学术领域，对网页进行分类已经有诸多方法. 文献 [5] 根据语义结构对 XML 网址进行分类，在实验中可以达到较高的准确率; 文献 [6] 使用遗传算法，以网页标签和属性为分类特征，对网页进行分类; 文献 [7] 基于网址结构对网页进行分类，为本文 USC 方法的提出提供了灵感. 文献 [8] 利用上下文特征，使用支持向量机对网页进行分类，是一种经典的网页分类方法. 文献 [9] 使用蚁群算法优选网页特征，并用朴素贝叶斯、KNN 等算法根据优选的特征进行分类，以提高分类的准确度和执行速度. 文献 [10] 将向量机和无监督聚类优势互补，旨在解决向量机效率低和无监督聚类准确度低的问题.

在正文提取方面，前人也已有诸多突出的成果: 文献 [11] 提出一种针对微博的正文提取方法，并以推特为例进行了实验. 文献 [3] 使用 DOM 节点的文本密度为标准进行正文提取，该方法便捷且高效. 文献 [2] 则在此基础上，提出针对短正文网页的正文提取方法，其提取正文的方法仍然为文本密度. 除了文本密度方法外，文献 [12] 提出根据网页结构和文本特征进行正文提取的方法，文献 [13] 则使用布局相似性作为依据进行提取，然而这些方法在处理短正文时也存在一定缺陷.

总体而言，传统的网页分类更多倾向于内容特征方法，而正文提取则以文本密度为基本思想. 内容特征的优点是易于分离出不同网页的个性与共性，为分类提供依据，且这些特征不易受主观因素影响，而缺点是内容特征自身的选择易受主观影响，且分类结果易受选择的结果影响，缺乏通用性和指导依据. 而文本密度方法优点是简单高效，具有一定通用性，但缺点是完全忽略了网页中正文的语义因素，容易受到长度较高的噪声影响，且难以应用到正文较短，结构复杂的论坛之中.

## 3　通用论坛正文提取

### 3.1　网页结构化聚类方法

在主题帖页面识别部分，本文提出了一种基于网址结构的聚类方法 USC. 在该方法中，首先根据分隔符将网址划分为若干部分，随后使用聚类算法将网址划分到不同簇中，筛选出主题帖所在的簇，以此提取出所有主题帖页面的网址. USC 方法使用基于网址结构运作，而不对网页内容直接进行处理，相比传统方法更具有针对性，且在性能上也略胜一筹.

#### 3.1.1　网址结构特征

统一资源定位符( Universal Resource Locator，URL) ，俗称网址，是用于唯一定位互联网上网络资源的一种表示方法，其主要由传输协议、服务器、端口号、路径、查询、片段六部分组成. 隶属于同一论坛的网页，除了协议、服务器名和端口号完全相同外，在路径和查询部分也有一定的相似之处，如共用的文件夹，共用的附加参数等等. USC 从网址中提取若干结构特征和内容特征，以描述该网址，进而反映出网址指向的网页所属的类别.

论坛中的各类网页根据其生成方式的差异可以被划分为动态网页，静态网页，伪静态网页三种类别. 三类网页在网址结构上有着各自的特征:

动态网页: 动态网页的网址中会包含 "?"，其后的内容为向服务器提交的参数列表，该部分由若干以 "&" 符号连接的键值对组成.

静态网页: 静态网页不需要查询部分. 但为便于分类和查找，网页文件往往会被放置在特定的文件夹中，故网址中会包含较长的路径部分.

伪静态网页: 伪静态网页的网址结构既不含查询部分，且路径部分也相对较短，但由于实质是将动态网页的网址进行重写，所以路径部分往往会包含各种分隔符号.

由上面的说明可以得出，同一论坛下的各类网页，在网址结构上也会有诸多相似. 为此本文引入网址的结构向量，该向量对网址进行定量表示，对网址中不同位置的结构块按其内容类别( 文本或数字) 和内容进行编码.

**定义 1.** 网址的结构向量. 一个结构向量由若干结构块编号元组组成，二者的定义分别为:

$$p(u,i) = (t(u,i), v(u,i)) \tag{1}$$

$$S(u) = \{p(u,i) \mid i = 1,2,\cdots,N\} \tag{2}$$

其中 $u$ 为网址，$i$ 指网址中第 $i$ 个结构块，$t(u,i)$ 为类别编号，$v(u,i)$ 为值编号，$p(u,i)$ 为结构块编号元组; $N$ 为总结构块数，$S$ 为结构向量.

一个结构块即为被分隔符号包围的一段字符串，在为网址 $u$ 的第 $i$ 个结构块进行编号时，首先为其类别进行编号，当该结构块的类型( 包括空值) 为首次出现时，为其赋予新的编号，否则沿用已经为该类型分配的编号; 对值分配编号时类似，若该结构块的值( 包括空值) 为首次出现时，赋予新的编号，否则沿用既有编号. $N$ 的值取数据集中拥有最多结构块的网址 $u$ 的结构块数. 若某网址的结构块数不足 $N$，则不足的部分以空值的编号补齐. 此外，当对同一论坛提取的网页构造结构向量时，可以忽略其传输协议和域名部分.

在上述定义下，若有如下 5 条 网址在数据集中:

1. http://example.com/query.php?id=001&grade=100
2. http://example.com/query.php?id=001&grade=99
3. http://example.com/query.php?id=002&grade=100
4. http://example.com/query.php?id=002&grade=99
5. http://example.com/query.php?id=003

分别简记为 $u_1$ 到 $u_5$，类型编号如表 1 所示.

表 1　类型编号

Table 1　Type numbers

| 编号 | 类型 |
|---|---|
| 0 | ( 空值) |
| 1 | 纯字母 |
| 2 | 纯数字 |

对值编号如表 2 所示.

表 2 值编号

Table 2 Value numbers

| 编号 | 值 | 编号 | 值 | 编号 | 值 |
|---|---|---|---|---|---|
| 0 | (空值) | 4 | 001 | 8 | 002 |
| 1 | Query | 5 | grade | 9 | 003 |
| 2 | php | 6 | 100 | | |
| 3 | id | 7 | 99 | | |

则对上述网址分别构造结构向量为:

$S(u_1) = [(1,1),(1,2),(1,3),(2,4),(1,5),(2,6)]$

$S(u_2) = [(1,1),(1,2),(1,3),(2,4),(1,5),(2,7)]$

$S(u_3) = [(1,1),(1,2),(1,3),(2,8),(1,5),(2,6)]$

$S(u_4) = [(1,1),(1,2),(1,3),(2,8),(1,5),(2,7)]$

$S(u_5) = [(1,1),(1,2),(1,3),(2,9),(0,0),(0,0)]$

在对上述网址构造结构向量时,由于其均隶属于同一域名,故统一忽略域名部分,从路径部分开始构造. 每次读取一条网址,分别查找对应的结构块的类型和值的编号,若找到则采用现有编号,否则在编号表中添加新的编号.

定义 2. 网址之间的相异度. 如下所示:

$$N_s(u_m,u_n) = \min\{i \mid p(u_m,i) \neq p(u_n,i)\} \tag{3}$$

$$D(u_m,u_n) = \left\lfloor \frac{N!}{N_s(u_m,u_n)!} \right\rfloor \tag{4}$$

其中 $u_m$,$u_n$ 代表不同的网址,$p(u,i)$ 与 $N$ 的定义为定义 1 中的定义,$D(u_m,u_n)$ 为网址 $u_m$ 和 $u_n$ 的相异度.

$N_s(u_m,u_n)$ 是要找到最小的 $i$,使得两个网址在位置 $i$ 的结构块不同. 如此定义可防止当两个网址完全不同和仅第一个结构块相同时相异度相等的情况.

目录结构是典型的树形结构,上级目录的差异会在后续层级中不断被扩大,因而对高层目录级赋予较高权值,他们体现为阶乘中较大的乘数. 当其产生差异时,后续子目录会进一步扩大这种差异,故采用乘法连接不同层级.

在上述定义下,若有如下三个结构向量:

$S(u_1) = [(1,1),(1,2),(1,3),(1,4),(1,5)]$

$S(u_2) = [(1,1),(1,2),(1,3),(1,1),(1,2)]$

$S(u_3) = [(1,1),(2,2),(1,3),(1,4),(1,5)]$

则对于 $u_1$ 和 $u_2$,其从第 4 个结构块开始不同,故 $N_s(u_1,u_2) = 4$,$D(u_1,u_2) = 5!/4! = 5$; 而对于 $u_1$ 和 $u_3$,虽然仅有第 2 维不同,但由于其为第一组不同的,所以 $N_s(u_1,u_3) = 2$,$D(u_1,u_3) = 5!/2! = 60$.

3.1.2 KNN-PDC 聚类

本文以 KNN-PDC 聚类[14] 为例,介绍聚类算法在 USC 方法中的应用. KNN-PDC 聚类是 PDC 聚类[15] 的一种改进. PDC 聚类利用簇中心局部密度较高和不同簇中心距离较远两条假设,分别求取每个点的局部密度和到另一个密度更高的点的距离,以此确定簇中心. KNN-PDC 聚类在此基础上利用样本点的 K 近邻信息,统一了 DPC 聚类的局部密度定义,摆脱了 DPC 算法受截断距离影响较大的缺点[14].

在 KNN-PDC 聚类中,簇中心的确定依赖于样本 $i$ 局部密度 $\rho_i$ 和局部密度大于 $i$ 且离 $i$ 最近的样本点距离 $\delta_i$,下面,我们给出二者的定义.

定义 3. 局部密度 $\rho_i$ 和局部密度大于 $i$ 且离 $i$ 最近的样本点距离 $\delta_i$. 公式如下所示:

$$\rho_i = \sum_{j \in KNN(i)} \exp(-d_{ij}) \tag{5}$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{6}$$

其中,$KNN(i)$ 代表点 $i$ 的 K 近邻集合,$d_{ij}$ 为点 $i$ 与点 $j$ 之间的距离. K 近邻集合中的点到 $i$ 的距离越近,则 $\rho_i$ 值越高,局部密度越大. $\delta_i$ 则当比 $i$ 局部密度大的最近的点离 $i$ 较远时 $\delta_i$ 较大.

如下是 KNN-PDC 聚类的主要流程.

**Algorithm 1** KNN – PDC Clustering

```
Input points: distance function is available
Input K: number of neighbors
Output clusters: result clusters
procedure Clustering(points, K)
    for all i in points do
        //KNN(i) 指点 i 的 K 近邻点集
        ρ_i ← sum of distance to points in KNN(i)
        δ_i ← min distance to point j where ρ_j > ρ_i
    end for
    DecisionGraph ← scatterplot of each δ_i and ρ_i
    //更高的 ρ_i 和 δ_i 意味着更可能是族中心
    centers ← points in DecisionGraph w/high ρ_i and δ_i
    clusters ← BFS(points, centers, K)
    return clusters
end procedure
```

在上述流程中,用到了寻找各簇中心的 K 近邻点的函数 BFS,该函数的流程如算法 2 所示.

**Algorithm 2** Breadth First Search

```
Input points: distance function is available
Input centers: centers with high ρ_i and δ_i
Input K: number of neighbors
Output clusters: result clusters
procedure BFS(points, centers, K)
    clusters ← Array < Cluster >
    for all c in centers do
        allocate c and KNN(c) to an empty cluster
        Queue.push(KNN(c))
        While Queue.is Empty() = False do
            q ← Queue.pop()
            for all r in KNN(q) do
                //KNN-DPC 判断隶属关系的条件
                if r has not been allocated and d_qr ≤ mean({d_rj | j ∈
                KNN(r)}) then
                    allocate r to cluster with center c
                    Qucue.push(r)
                end if
            end for
        end while
    end for
    return clusters
end procedure
```

通过如上算法,可以将输入的所有网页划分为若干簇,以便进行后续处理. 此外,由于主题帖网址大多具有极高的结构

相似性,因而本文中没有考虑对离群点的处理,如有相关需求,可进一步参考文献[14].

3.1.3　网址解析模块

为了能够快速对大量的网址进行类别筛选,同时为了适应分布式计算的要求,我们需要将分类后的结果进行解析. 本文给出了较为通用的解析器定义,通过对给定网页各结构块的排列方式,内容加以限定,进而得到网址的解析器. 下面,我们将给出其公式定义.

**定义 4.** 解析其模块 $r(i)$ 和解析器 $R$. 其公式如下:

$$r(i) = (t(i), v(i)) \tag{7}$$

$$R = \{r(i) \mid i = 1, 2, \cdots, N\} \tag{8}$$

其中 $t$ 为类别集合,描述所有待解析网址定位置的结构块类型,$v$ 为与 $t$ 中元素对应的值集合,描述所有待解析网址子特定位置的结构块值,$r$ 为对应于特定位置的结构块的规则元组,$R$ 为对所有网址的解析器.

在使用时,解析器和网址结构向量类似,不同处在于解析器中每个位置可能存放多个编号值,使用时采用令网址相异度最小的组合方式,与普通网址计算相异度.

3.1.4　结构化聚类流程

上文中分别介绍了结构向量,网址相异度,KNN-DPC 聚类和解析器的有关内容,该小节将结合前述内容,介绍 USC 方法的完整流程. 在网页类型识别的过程中,往往需要对从同一论坛获取的大量网页同时进行处理,这并不要求对所有网页都进行聚类等操作,而是先对少量网页进行解析,将解析所得规则应用于其他页面,以此快速完成对所有页面的分类,也便于利用集群进行分布式处理. 上述方法的伪代码表示如算法 3 所示.

**Algorithm 3** URLs' Structure Clustering

**Input** *urls*: urls from single forum
**Input** $K$: number of neighbors
**Output** *topics*: urls of topic pages
1: **procedure** USC(*urls*, $K$)
2:　　*samples*←*random urls in urls*
3:　　*a*←*specified topic page url in samples*
4:　　//类别有如动态网页、伪静态网页等
5:　　*samples*←*samples with same type of a*
6:　　*vectors*←*structure vectors of all urls*
7:　　*clusters*←Clustering(*vectors*, $K$)
8:　　*topics*←*cluster to which a belongs*
9:　　//根据主题帖网址构造解析器
10:　　*Rule*← *resolve rule from topics*
11:　　*topics*←*topic urls identified by Rule*
12:　　**return** *topics*
13: **end procedure**

### 3.2　关键字打分筛选方法

在主题帖正文提取部分,本文提出关键词打分筛选方法(Keyword Scoring Filter,KSF). 在该方法中,需要使用特定方法确定词条关键程度,我们以词频-逆向文件频率(TF-IDF)方法为例,来介绍 KSF 方法的主要思想和算法流程.

首先,根据停用词库排除无关文本,其后根据 TF-IDF 值找出文本关键词,接着对关键词出现的区域进行打分,解析得到得分最高的区域,对其他网页采用解析得到的规则直接进行正文提取. 采用这种方法,不仅提升了主题帖正文提取的准确率和运行效率,同时可以获得适用于其他页面的通用解析规则,以简化后续的对于同一论坛的提取工作.

3.2.1　TF-IDF 统计方法

TF-IDF 统计方法是一种用于信息检索和文本加权的技术,常用于评估某一词条在一个语料库中的一份文件中的重要程度. 词条的重要性随着其在文件中出现的次数上升而上升,随着其在语料库中出现的次数上升而下降.

词频(Term Frequency,TF)的计算是该方法的第一个步骤,表示一词条在当前文件中出现的频率,其计算公式如下:

$$TF(f, w) = \frac{N(f, w)}{\sum_{i \in F} N(i, w)} \tag{9}$$

其中,$F$ 为语料库中所有文件的集合,$f \in F$ 为文件,$w$ 为词条,$N(f, w)$ 为词条 $w$ 在文件 $f$ 中出现的次数.

当一个词条在文件中多次出现,如在计算机类文章中频繁出现"分布式",则我们有理由认为该词条在该文章中重要性较高. 体现到 TF 值中,则会被赋予较高的值.

逆向文件频率(Inverse Document Frequency,IDF)是该方法的第二步,该指标体现一个词条在所有文件中出现的普遍程度,其公式如下:

$$IDF(f, w) = \log \frac{\sum_{i \in F} 1}{1 + \sum_{i \in F} P(i, w)} \tag{10}$$

其中,$P(f, w)$ 为词条出现函数,当词条 $w$ 出现在文件 $f$ 中时,值为 1,否则为 0,分母部分 +1 是为了防止词条 $w$ 在语料库中没有出现过,导致分母为 0 的情况发生.

当一个词条在多个文件中频繁出现,如在各类文件中都大量出现的"的"字,则我们有理由认为该词条在整个语料库中重要性较低. 体现到 IDF 值中,也就会导致分母值上升,使得 IDF 值下降.

结合上述二者,最终可以根据以下公式得到词条 $w$ 在文件 $f$ 中相对于整个语料库 $F$ 的重要度.

$$W(f, w) = TF(f, w) \times IDF(f, w) \tag{11}$$

其中 $W$ 为词条 $w$ 的重要度,该重要度体现了词条在文章中的突出程度和语料库中普遍程度二者的调和,倾向于选择在文件中真正突出的词条,而排除掉因行文需要而大量出现的常用字词.

3.2.2　打分评价方法

在利用上文中的 TF-IDF 方法对文本重要度进行加权的过程中,难免会遇到异常页面,如无人回复的冷门帖,含有大量图片而缺乏文字的图片帖等等,这些噪声的存在可能会对主题帖关键词提取造成误导,使得无法得到正确的提取规则,进而使得对所有页面的正文提取全部出错. 为预防这类问题,本文基于正文出现频率大于噪声出现频率的假设,采用打分评价方法对每个网页反馈的结果进行评估,依照得分高低决定采用哪一个结果作为最终的输出. 其伪代码流程如算法 4 所示.

除了可以将算法 4 应用到对每个网页返回结果的评估中外,还可以将其应用到对单一网页内正文位置的判别中,其基

本流程与上述流程类似,故不再赘述.

**Algorithm 4** Scoring

```
Input objs: candidates group
Output result: results group
1: procedure Scoring(objs)
2:     scores←Map<Object,int>
3:     for all obj in objs do
4:         scores[obj].value←scores[obj].value+1
5:     end for
6:     result←objs in scores with highest score
7:     return result
8: end procedure
```

### 3.2.3 关键字打分筛选流程

上文中分别介绍了 TF-IDF 统计方法和打分评价方法,下面将两种算法结合,加以部分优化步骤,构成从主题帖页面中提取出主题帖正文的完整算法.

首先,从所有网页中抽取部分网页作为样本; 然后,初步清理样本网页中的噪声并剔除重复样本; 接着,对样本网页中的文本进行分词并计算权重; 再后,对词条隶属网页元素打分; 最后,借助打分结果构造解析规则,利用打分结果对其他网页提取正文. 其伪代码流程描述如算法 5 所示.

**Algorithm 5** Keyword Scoring Filter

```
Input htmls: source code of web pages
Output content: content extracted from web pages
1: procedure KSF(htmls)
2:     samples←random web pages in htmls
3:     results←Array<Location>
4:     for all html in samples do
5:         lince←html w/o tags or irrelevant contents
6:         //规则包括停用规则、相似规则等
7:         remove lines satisfying Rules
8:         words←split lines by word
9:         locs←Array<Location>
10:        for all word in words do
11:            word.weight←TF-IDF value of word
12:            locs.add(locations where word appears)
13:        end for
14:        //需要的位置而非关键词本身
15:        results.add(S_CORING(locs))
16:    end for
17:    //根据打分结果构造提取规则
18:    ExtractRule←obtained from Scoring(results)
19:    return content extracted by ExtractRule
20: end procedure
```

在算法 5 中,涉及到了停用规则,内容相似规则,提取规则三种规则,这三类规则的具体实现可以根据用户自身需求自行定义. 停用规则是当单一行内的文本满足特定条件时,该行文本会被视噪声而被去除,常用的停用规则有根据停用词库匹配,设定文本长度阈值,是否存在特定文本结构等. 内容相似规则用于判断两部分内容是否由网页模板生成,而非用户所写,若时是则将其去除,常见的生成内容如回帖发表日期,用户个人信息栏,"只看楼主"等论坛操作按钮,等等,为了辨别这些内容,可以选择每行文本的开头数个字符作为键,若键相同则视为模板生成. 提取规则用于精确表示需要进行提取的位置,选用的方法必须能够恰好涵盖所有正确的位置,通常选用代码中标签的 class 属性值.

## 4 实验与分析

### 4.1 结构化聚类

相比传统分类方法,USC 并不直接对所有网页进行分类,而是从所有网页中抽取部分样本作为训练数据,根据分类结果构造解析器,再利用解析器对剩余网页进行解析,因而最终的分类结果在很大程度上取决于解析器,构造解析器的规则中,除直接使用结构向量外,也可使用正则表达式以提升程序的通用性. 本文实验中为便于描述,解析器将仍然使用结构向量.

#### 4.1.1 实验数据

我们利用网络爬虫在互联网上随机爬取了若干网页,经过人工去除非论坛网页、广告页等无关网页,剩余有效网页数 13346,其中主题帖网页数 5888,在所有有效网页中,Discuz! 论坛页面 11822,独立论坛页面数 1524. 这些页面来自不同领域不同话题的论坛,具有一定的代表性.

#### 4.1.2 评价指标

在信息提取领域,通用的评价指标是召回率 $R$,准确率 $P$ 和 $F$ 值. 三者的计算公式分为:

$$R=\frac{N_{et}}{N_t} \tag{12}$$

$$P=\frac{N_{et}}{N_e} \tag{13}$$

$$F=\frac{2PR}{P+R} \tag{14}$$

其中,$N_{et}$ 为提取主题帖数,$N_t$ 为主题帖网址数,$N_e$ 为提取数. 一般而言,提取结果的优劣由 $F$ 值进行评估,其值越高,效果越好.

#### 4.1.3 结果分析

本部分实验主要是与文献[16]所采用的基于 DOM 树的网页聚类算法相比较,在同样的数据集下,二者的运行结果分别如表 3 和表 4 所示.

在此次实验中,我们根据预实验的结果对 K 值进行了微调,以维持较高的准确率,尽管可能会有部分坛的由于网址结构过于单一而导致本应同类别的网址被划分到不同类别中,使得召回率降低,但对于一般的网站,基本可以维持较高的召回率,同时也使得准确率得到保证,确保构造的解析器中不会将噪声网址包含在内.

将表 3 和表 4 中的数据绘制为折线图,其结果如图 1,图 2,图 3 所示.

综合实验结果,可以看出 USC 方法中的聚类部分具有如下优点:

1. 准确度高,构造解析器的簇中不含任何噪声,确保使用解析器进行大规模提取时的正确性.

2. 适应性强,对各类论坛均有较高的召回率,确保拥有足

够的样本数据构造解析器并使得解析器可以提取到更多符合要求的网页网址.

表 3　USC 聚类结果

Table 3　Result of clustering in USC

| 序号 | 网址数 | 主题帖网址数 | 提取数 | 提取主题帖数 | 召回率/% | 准确率/% | F 值/% |
|---|---|---|---|---|---|---|---|
| 1 | 522 | 287 | 287 | 287 | 100.00 | 100.00 | 100.00 |
| 2 | 762 | 214 | 211 | 211 | 98.60 | 100.00 | 99.29 |
| 3 | 481 | 202 | 202 | 202 | 100.00 | 100.00 | 100.00 |
| 4 | 368 | 131 | 128 | 128 | 97.71 | 100.00 | 98.84 |
| 5 | 1102 | 671 | 671 | 671 | 100.00 | 100.00 | 100.00 |
| 6 | 627 | 291 | 282 | 282 | 96.91 | 100.00 | 98.43 |
| 7 | 633 | 201 | 198 | 198 | 98.51 | 100.00 | 99.25 |
| 8 | 831 | 455 | 449 | 449 | 98.68 | 100.00 | 99.34 |
| 9 | 487 | 210 | 190 | 190 | 90.48 | 100.00 | 95.00 |
| 10 | 859 | 282 | 282 | 282 | 100.00 | 100.00 | 100.00 |

表 4　文献 [16] 的聚类结果

Table 4　Result of clustering in reference [16]

| 序号 | 网址数 | 主题帖网址数 | 提取数 | 提取主题帖数 | 召回率/% | 准确率/% | F 值/% |
|---|---|---|---|---|---|---|---|
| 1 | 522 | 287 | 230 | 230 | 82.58 | 100.00 | 90.46 |
| 2 | 762 | 214 | 80 | 80 | 38.19 | 100.00 | 55.27 |
| 3 | 481 | 202 | 90 | 90 | 48.67 | 100.00 | 65.47 |
| 4 | 368 | 131 | 50 | 50 | 43.09 | 100.00 | 60.22 |
| 5 | 1102 | 671 | 530 | 530 | 79.93 | 100.00 | 88.85 |
| 6 | 627 | 291 | 210 | 210 | 75.52 | 100.00 | 86.05 |
| 7 | 633 | 201 | 130 | 130 | 64.86 | 100.00 | 78.69 |
| 8 | 831 | 455 | 390 | 390 | 87.04 | 100.00 | 93.07 |
| 9 | 487 | 210 | 210 | 210 | 100.00 | 100.00 | 100.00 |
| 10 | 859 | 282 | 190 | 190 | 69.73 | 100.00 | 82.17 |

### 4.2　关键字打分

KSF 方法采用语义分割和 TF-IDF 加权方法提取出网页文本部分的关键词,对关键词所在位置进行打分,构造出解析规则,故能否从网页中正确提取到主题帖的正文信息取决于

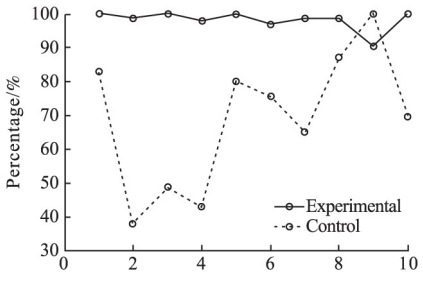


图 1　召回率

Fig. 1　Recall value

解析规则的构造情况. 本实验中需要的对文本进行进行分词,为简化实验流程,实验中的分词程序将使用开源的分词程序包 jieba. 下面,本实验将对使用 KSF 方法构造解析规则的过程进行评估.

#### 4.2.1　实验数据

我们利用网络爬虫,在每个网络论坛中爬取若干论坛页面,去除非主题帖页面后,所有论坛剩余主题帖页面数总计为

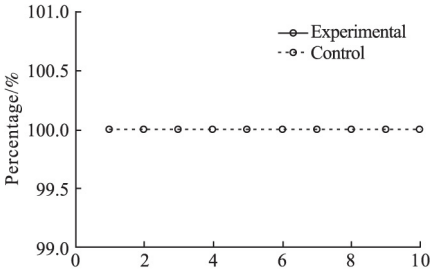


图 2　准确率

Fig. 2　Precision value

3417,每次从同一论坛的页面中随机抽取 5 个页面进行训练,根据解析所得规则判断是否正确分离出所需内容,每个论坛将重复进行 10 次实验.

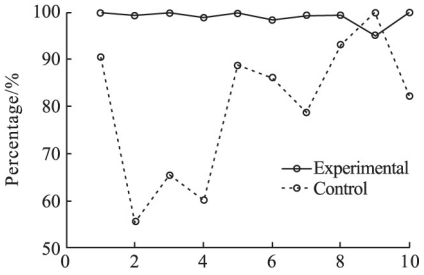


图 3　F 值

Fig. 3　F value

#### 4.2.2　评价指标

本实验重在检验解析规则的正确性,为简化处理,以标签 class 属性值作为解析规则,故评价指标为解析所得 class 属性值恰为论坛正文区标签 class 属性值的比例和程序用时.

#### 4.2.3　结果分析

这一部分实验我们使用文献 [2] 中的 SCEED 算法作为对照实验,实验的结果如表 5 和表 6 所示.

表 5　正文提取算法综合对比

Table 5　Brief comparison between KSF and SCEED

| 算法名称 | 准确率/% | 平均用时/s |
|---|---|---|
| KSF | 90.77 | 28.24 |
| SCEED | 72.31 | 27.43 |

在用时方面,KSF 算法相比 SCEED 算法要慢,进行中文分词需要读取中文数据库,并对字符串序列进行划分和匹配. SCEED 方法没有这些步骤,因而速度相对较快,但准确率较差.

综合实验数据,KSF 方法具有以下特点:

1. 鲁棒性较强,在有较多噪声的主题帖页面中仍可正常进行解析工作.

2. 通用性强,解析出的规则可以应用于同论坛内的所有主题帖页面.

3. 效率高,对一般论坛中单一网页的解析速度视内容量

从 2 秒到 9 秒不等.

表 6 正文提取算法效果对比

Table 6 Detailed comparison between KSF and SCEED

| 序号 | 正确数/个 | | 平均用时/s | |
|---|---|---|---|---|
| | KSF | SCEED | KSF | SCEED |
| 1 | 10 | 10 | 28.61 | 39.33 |
| 2 | 10 | 8 | 44.02 | 33.47 |
| 3 | 7 | 0 | 38.80 | 34.67 |
| 4 | 9 | 10 | 9.62 | 7.47 |
| 5 | 8 | 5 | 37.31 | 34.41 |
| 6 | 9 | 10 | 25.52 | 22.54 |
| 7 | 10 | 10 | 21.53 | 22.41 |
| 8 | 10 | 10 | 12.14 | 11.86 |
| 9 | 9 | 3 | 18.29 | 16.05 |
| 10 | 8 | 6 | 11.98 | 17.10 |
| 11 | 10 | 10 | 4.61 | 3.35 |
| 12 | 10 | 6 | 18.25 | 19.48 |
| 13 | 8 | 6 | 96.46 | 94.41 |

## 5 总结与展望

本文首先提出了网址的结构向量表示法,使得对网址的处理可以采用类似向量的的方法进行,此外还给出网址相异度函数的定义,这是一种结合目录结构特性的相异度计算方法,可以将其推广应用到更多目录结构的表示中. 此外,本文第二节提出的基于网址结构的聚类方法 USC 是对传统论坛网页分类方法的改进,该方法充分利用网络论坛在网址结构上的相似性和结构化特性,无需读取网页内容,即可对同论坛下的网页直接进行分类,同时兼顾性能与质量. 最后,本文结合语义分词技术,TF-IDF 加权统计方法,打分评估方法提出的关键词打分筛选方法,可以快速对论坛中的网页解析出通用的提取规则,应用于后期大规模正文提取中,以满足持久化信息提取的需求.

本文在使用 KNN-DPC 进行聚类时,虽然可以通过决策图直观地选出簇中心,然而仍然缺乏一种适合于所有情况的自动化簇中心选择方法,在接下来的工作中将通过曲线分析寻找自动确定簇数的方法. 此外,本文中为便于描述而采用较为简单的解析规则,在实际应用过程中,除可替换为正则表达式,还可根据需求完全自定义解析规则的行驶,如何合理构造具有较强鲁棒性的解析规则,仍是一个亟待研究的问题.

**References:**


[1] China Internet Network Information Center. 38th China statistical report on Internet development [R]. http://www.cnnic.net.cn/,2016.

[2] Xi Jia-zhen,Guo Yan,Li Qiang,et al. A content extraction method for short web pages [J]. Journal of Chinese Information Processing,2016,30(1):8-15.

[3] Sun Fei,Song Dan-dan,Liao Le-jian. Dom based content extraction via text density [C]. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM,2011:245-254.

[4] Wang Hai-yan,Cao Pan. Information extraction from massive web pages based on node property and text content [J]. Journal on Communications,2016,37(10):9-17.

[5] Krishna Murthy A. XML URL classification based on their semantic structure orientation for web mining applications [J]. Procedia Computer Science,2015,46:143-150.

[6] Selma Ayşe özel. A Web page classification system based on a genetic algorithm using tagged-terms as features [J]. Expert Systems with Applications,2011,38(4):3407-3415.

[7] Kan Min-yen,Thi Hoang Oanh Nguyen. Fast webpage classification using URL features [C]. Proceedings of the 14th ACM International Conference on Information and Knowledge Management,ACM,2005:325-326.

[8] Sun Ai-xin,Lim Ee-peng,Ng Wee-keong. Web classification using support vector machine [C]. Proceedings of the 4th International Workshop on Web Information and Data Management,ACM,2002:96-99.

[9] Sarac E,Ozel S A. An ant colony optimization based feature selection for web page classification [J]. Scientific World Journal,2014,649260.

[10] Li Xiao-li,Liu Ji-min,Shi Zhong-zhi. A Chinese web page classifier based on support vector machine and unsupervised clustering [J]. Chinese Journal of Computers,2001,24(1):62-68.

[11] Bontcheva K,Derczynski L,Funk A,et al. TwitIE: an open-source information etraction pipeline for microblog text [C]. RANLP,2013:83-90.

[12] Xiong Zhong-yang,Lin Xian-qiang,Zhang Yu-fang,et al. Content extraction method combining web page structure and text feature [J]. Computer Engineering,2013,39(12):200-203,210.

[13] Yang Liu-qin,Li Xiao-dong,Geng Guang-gang. Study of web pages content extraction based on layout similarity [J]. Application Research of Computers,2015,32(9):2581-2586.

[14] Xie Juan-ying,Gao Hong-chao,Xie Wei-xin. K-nearest neighbors optimized clustering algorithm by fast search and finding the density peaks of a dataset [J]. Scientia Sinica Informationis,2016,46(2):258-280.

[15] Rodriguez A,Laio A. Clustering by fast search and find of density peaks [J]. Science,2014,344(6191):1492-1496.

[16] Liu Chun-mei,Guo Yan,Yu Xiao-ming,et al. Information extraction research aimed at open source web pages [J]. Journal of Frontiers of Computer Science & Technology,2017,11(1):114-123.


**附中文参考文献:**


[1] 中国互联网络信息中心. 第 38 次中国互联网络发展状况统计报告 [R]. http://www.cnnic.net.cn/,2016.

[2] 郁家贞,郭 岩,黎 强,等. 一种短正文网页的正文自动化抽取方法 [J]. 中文信息学报,2016,30(1):8-15.

[4] 王海艳,曹 攀. 基于节点属性与正文内容的海量 Web 信息抽取方法 [J]. 通信学报,2016,37(10):9-17.

[10] 李晓黎,刘继敏,史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器 [J]. 计算机学报,2001,24(1):62-68.

[12] 熊忠阳,蔺显强,张玉芳,等. 结合网页结构与文本特征的正文提取方法 [J]. 计算机工程,2013,39(12):200-203,210.

[13] 杨柳青,李晓东,耿光刚. 基于布局相似性的网页正文内容提取研究 [J]. 计算机应用研究,2015,32(9):2581-2586.

[14] 谢娟英,高红超,谢维信. K 近邻优化的密度峰值快速搜索聚类算法 [J]. 中国科学:信息科学,2016,46(2):258-280.

[16] 刘春梅,郭 岩,俞晓明,等. 针对开源论坛网页的信息抽取研究 [J]. 计算机科学与探索,2017,11(1):114-123.