(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107402998 A
(43)申请公布日 2017.11.28

(21)申请号 201710598015.6

(22)申请日 2017.07.20

(71)申请人 山东师范大学
地址 250014 山东省济南市文化东路88号

(72)发明人 王红 刘锐

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 张勇

(51)Int.Cl.
*G06F 17/30*(2006.01)

权利要求书3页 说明书12页 附图1页

(54)发明名称

一种基于网址结构的网络论坛页面聚类方法及设备

(57)摘要

本发明涉及一种基于网址结构的网络论坛聚类方法及设备，该方法涉及数据挖掘领域，是为解决大规模网页分类问题而提出的。该方法从网址总体中抽取部分样本，利用网络论坛网址高度结构化的特性，对每个网址进行结构划分，构造结构向量，使用本发明提出的距离函数评估结构向量之间的距离，接着使用密度峰值聚类方法对样本结构向量进行聚类分析，提取出每簇的特征结构，构造用于描述簇中所有样本网址的解析器，用于对总体中剩余网址进行解析和分类。经实验验证，本方法具有较高的准确度和执行效率。


CN 107402998 A

1.一种基于网址结构的网络论坛页面聚类方法,其特征是:该方法包括以下步骤:

(1)按照网页所属域名对所有网页进行初步分组,对于初步分组后的每一组网页进行抽样组成样本,并在样本中插入带标记的待筛选网页形成样本网页;

(2)将样本网页的除域名外的网址根据符号进行分割,对分割后的每一部分网址的类别和内容进行编号,构造出结构块;

(3)将同一网址的各个结构块按顺序依次排列,构成该网址的结构向量;计算样本网页中的任意两个结构向量的相异度,和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度;分别作为横坐标和纵坐标构造决策图,确定簇中心,采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇;

(4)根据步骤(3)的决策图构造出解析规则,将解析规则应用于初步分组后的每一组网页中的非样本网页,进行网页聚类筛选,并采用评价指标进行评价。

2.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法,其特征是:所述步骤(2)中构造的结构块,用于定量表示网页分割后每部分网址的结构;其构造的具体步骤为:

将样本网页的除域名外的网址根据符号进行分割,判断分割后的每一部分网址的类别和内容是否已有编号;

若某类别或内容已有编号,则采用此编号;

否则,赋予该部分网址的类别和内容一个新编号;

重复上述步骤,直至构成所有样本网页的结构块。

3.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法,其特征是:所述步骤(3)中网址的结构向量,将结构块组合以表示完整网址的结构;一个结构向量S(u)由若干结构块编号元组p(u,i)组成:

$$p(u,i)=(t(u,i),v(u,i)) \quad (1)$$

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

其中,u为网址,i为网址中第i个结构块,t(u,i)为类别编号,v(u,i)为值即内容编号,p(u,i)为结构块编号元祖,N为总结构块数,S(u)为结构向量。

4.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法,其特征是:所述步骤(3)中样本网页中的任意两个结构向量的相异度的计算方法为:

$$D(u_m,u_n)=\left\lfloor\frac{N!}{\min\{i|p(u_m,i)\neq p(u_n,i)\}}\right\rfloor \quad (3)$$

其中,$u_m$,$u_n$为不同的网址,p(u,i)为结构块编号元组,i为网址中第i个结构块,N为总结构块数,$D(u_m,u_n)$为网址$u_m$,$u_n$的相异度。

5.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法,其特征是:所述步骤(3)中计算网页样本中最小较高密度结构向量相异度的具体步骤为:

对网页样本中的每个结构向量,分别计算其局部密度;

对于网页样本中的任一结构向量,判断其局部密度与其他结构向量的局部密度,在局部密度大于其局部密度的结构向量中,比较该结构向量与其他大于其局部密度的结构向量的相异度,选择最小的相异度作为网页样本中该结构向量的最小较高密度结构向量相异

度；

所述步骤(3)中结构向量的局部密度$\rho_i$为：

$$\rho_i = \sum_{j \in KNN(i)} \exp(-d_{ij}) \qquad (4)$$

其中，KNN(i)为结构向量i的K近邻集合，$d_{ij}$为结构向量i与结构向量j的相异度，$\rho_i$为结构向量i的局部密度；

所述步骤(3)中最小较高密度结构向量相异度$\delta_i$为：

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \qquad (5)$$

其中，$\delta_i$为结构向量i的最小较高密度结构向量相异度。

6.如权利要求5所述的一种基于网址结构的网络论坛页面聚类方法，其特征是：所述步骤(3)中构造决策图的具体步骤为：

分别以每个结构向量的$\rho_i$值和$\delta_i$值为横纵坐标做散点图；

选择$\rho_i$值和$\delta_i$值均较高的若干结构向量作为聚类的簇中心；

采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇。

7.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法，其特征是：所述步骤(3)中采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇的具体步骤为：

对于每个簇中心，将其K近邻的结构向量加入该簇，并加入队列；

每次取队列队首的结构向量，对于其K近邻的结构向量，判断其是否被分配至任何簇，若未被分配至任何簇，则加入队首所在簇并加入队列；

重复上述步骤直至确认除簇中心的结构向量外的其他结构向量的归属。

8.如权利要求1所述的一种基于网址结构的网络论坛页面聚类方法，其特征是：所述步骤(4)中构造出解析规则的具体步骤为：

在决策图中，选出样本网页中插入的带标记的待筛选网页所在的簇；

在选出的簇中，对于该网页结构向量中的每个位置，记录所有该位置的结构块的类别和内容，当出现超过5种不同的内容时，不再记录内容，仅记录类别；

得到解析规则；

所述步骤(4)中采用评价指标进行评价时，分别采用召回率R，准确率P和F值进行聚类评价的指标：

$$R = \frac{N_{et}}{N_t} \qquad (5)$$

$$P = \frac{N_{et}}{N_e} \qquad (6)$$

$$F = \frac{2PR}{P+R} \qquad (7)$$

其中，$N_{et}$为提取待筛选网页数，$N_t$为待筛选网页数，$N_e$为提取数。

9.一种存储设备，其中存储有多条指令，所述指令适于由处理器加载并执行：

(1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行

抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

(2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

(3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

(4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

10.一种终端设备，包括：

处理器，适于实现各指令；以及

存储设备，适于存储多条指令，所述指令适于由处理器加载并执行：

(1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

(2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

(3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

(4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

# 一种基于网址结构的网络论坛页面聚类方法及设备

## 技术领域

[0001] 本发明属于网络数据挖掘的技术领域，尤其涉及一种基于网址结构的网络论坛页面聚类方法及设备。

## 背景技术

[0002] 网址是用于唯一确定一个网页的基本特征。而页面分类对网络数据挖掘具有重要意义，是对不同种类页面进行后续处理前的最重要的准备工作。目前，对网页进行分类的方法有根据语义结构进行分类；使用遗传算法，以网页标签和属性为分类特征进行分类；利用上下文特征，使用支持向量机进行分类。使用蚁群算法根据优选特征进行分类等等。但是，在实际中，论坛页面之间的共性并不显著，使得网页特征提取具有随意性；此外，网络论坛中页面众多，上述方法均难以满足大规模分类的速度需求。目前，基于论坛页面网址结构，构造结构向量进行聚类分析的方法尚未出现。

[0003] 综上所述，在现有技术中针对网络论坛页面如何有效进行网页分类，提高网页分类的准确度与效率的问题，尚缺乏有效的解决方案。

## 发明内容

[0004] 本发明为了解决上述问题，提供一种基于网址结构的网络论坛页面聚类方法及设备。本发明根据网址构造结构向量，并计算结构向量之间的相异度，使得可以使用聚类分析方法对网页进行分类，针对网络论坛页面有效实现网页分类，提高网页分类的准确度与效率。

[0005] 本发明的第一目的是提供一种基于网址结构的网络论坛页面聚类方法。

[0006] 为了实现上述目的，本发明采用如下一种技术方案：

[0007] 一种基于网址结构的网络论坛页面聚类方法，该方法包括以下步骤：

[0008] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0009] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0010] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0011] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0012] 进一步的，所述步骤(2)中构造的结构块，用于定量表示网页分割后每部分网址的结构；其构造的具体步骤为：

[0013] 将样本网页的除域名外的网址根据符号进行分割,判断分割后的每一部分网址的类别和内容是否已有编号;

[0014] 若某类别或内容已有编号,则采用此编号;

[0015] 否则,赋予该部分网址的类别和内容一个新编号;

[0016] 重复上述步骤,直至构成所有样本网页的结构块。

[0017] 进一步的,所述步骤(3)中网址的结构向量,将结构块组合以表示完整网址的结构;一个结构向量S(u)由若干结构块编号元组p(u,i)组成:

[0018] $p(u,i)=(t(u,i),v(u,i))$ (1)

[0019] $S(u)=\{p(u,i)\,|\,i=1,2,\cdots,N\}$ (2)

[0020] 其中,u为网址,i为网址中第i个结构块,t(u,i)为类别编号,v(u,i)为值即内容编号,p(u,i)为结构块编号元祖,N为总结构块数,S(u)为结构向量。

[0021] 进一步的,所述步骤(3)中样本网页中的任意两个结构向量的相异度的计算方法为:

[0022] $$D(u_m,u_n)=\left\lfloor\frac{N!}{\min\{i|p(u_m,i)\neq p(u_n,i)\}}\right\rfloor \quad (3)$$

[0023] 其中,$u_m$,$u_n$为不同的网址,p(u,i)为结构块编号元组,i为网址中第i个结构块,N为总结构块数,D($u_m$,$u_n$)为网址$u_m$,$u_n$的相异度。

[0024] 进一步的,所述步骤(3)中计算网页样本中最小较高密度结构向量相异度的具体步骤为:

[0025] 对网页样本中的每个结构向量,分别计算其局部密度;

[0026] 对于网页样本中的任一结构向量,判断其局部密度与其他结构向量的局部密度,在局部密度大于其局部密度的结构向量中,比较该结构向量与其他大于其局部密度的结构向量的相异度,选择最小的相异度作为网页样本中该结构向量的最小较高密度结构向量相异度。

[0027] 进一步的,所述步骤(3)中结构向量的局部密度$\rho_i$为:

[0028] $$\rho_i=\sum_{j\in KNN(i)}\exp(-d_{ij}) \quad (4)$$

[0029] 其中,KNN(i)为结构向量i的K近邻集合,$d_{ij}$为结构向量i与结构向量j的相异度,$\rho_i$为结构向量i的局部密度;

[0030] 所述步骤(3)中最小较高密度结构向量相异度$\delta_i$为:

[0031] $$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij}) \quad (5)$$

[0032] 其中,$\delta_i$为结构向量i的最小较高密度结构向量相异度。

[0033] 进一步的,所述步骤(3)中构造决策图的具体步骤为:

[0034] 分别以每个结构向量的$\rho_i$值和$\delta_i$值为横纵坐标做散点图;

[0035] 选择$\rho_i$值和$\delta_i$值均较高的若干结构向量作为聚类的簇中心;

[0036] 采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇;

[0037] 进一步的,所述步骤(3)中采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇的具体步骤为:

[0038] 对于每个簇中心，将其K近邻的结构向量加入该簇，并加入队列；

[0039] 每次取队列队首的结构向量，对于其K近邻的结构向量，判断其是否被分配至任何簇，若未被分配至任何簇，则加入队首所在簇并加入队列；

[0040] 重复上述步骤直至确认除簇中心的结构向量外的其他结构向量的归属。

[0041] 进一步的，所述步骤(4)中构造出解析规则的具体步骤为：

[0042] 在决策图中，选出样本网页中插入的带标记的待筛选网页所在的簇；

[0043] 在选出的簇中，对于该网页结构向量中的每个位置，记录所有该位置的结构块的类别和内容，当出现超过5种不同的内容时，不再记录内容，仅记录类别；

[0044] 得到解析规则。

[0045] 进一步的，所述步骤(4)中采用评价指标进行评价时，分别采用召回率R，准确率P和F值进行聚类评价的指标：

[0046] $$R = \frac{N_{et}}{N_t} \qquad (5)$$

[0047] $$P = \frac{N_{et}}{N_e} \qquad (6)$$

[0048] $$F = \frac{2PR}{P + R} \qquad (7)$$

[0049] 其中，$N_{et}$为提取待筛选网页数，$N_t$为待筛选网页数，$N_e$为提取数。

[0050] F值越高，则说明信息提取效果越好。

[0051] 本发明的第二目的是提供一种基于网址结构的网络论坛页面聚类方法的存储设备。

[0052] 为了实现上述目的，本发明采用如下一种技术方案：

[0053] 一种存储设备，其中存储有多条指令，所述指令适于由处理器加载并执行：

[0054] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0055] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0056] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0057] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0058] 本发明的第三目的是提供一种基于网址结构的网络论坛页面聚类方法的终端设备。

[0059] 为了实现上述目的，本发明采用如下一种技术方案：

[0060] 一种终端设备，包括：

[0061] 处理器，适于实现各指令；以及

[0062] 存储设备，适于存储多条指令，所述指令适于由处理器加载并执行：

[0063] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0064] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0065] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0066] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0067] 本发明的有益效果：

[0068] 本发明的一种基于网址结构的网络论坛页面聚类方法及设备，根据网址构造结构向量，并计算结构向量之间的相异度，使得可以使用聚类分析方法对网页进行分类，具有高准确度和执行效率。尤其针对共性不显著的论坛页面，本发明构造结构向量进行聚类分析，满足大规模分类的速度需求。

## 附图说明

[0069] 图1为本发明整体方法的流程图。

## 具体实施方式：

[0070] 应该指出，以下详细说明都是例示性的，旨在对本申请提供进一步的说明。除非另有指明，本发明使用的所有技术和科学术语具有与本申请所属技术领域的普通技术人员通常理解的相同含义。

[0071] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制根据本申请的示例性实施方式。如在这里所使用的，除非上下文另外明确指出，否则单数形式也意图包括复数形式，此外，还应当理解的是，当在本说明书中使用术语“包含”和/或“包括”时，其指明存在特征、步骤、操作、器件、组件和/或它们的组合。

[0072] 在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面结合附图与实施例对本发明作进一步说明。

[0073] 实施例1：

[0074] 正如背景技术所介绍的，本发明为了解决上述问题，提供一种基于网址结构的网络论坛页面聚类方法及设备。本发明根据网址构造结构向量，并计算结构向量之间的相异度，使得可以使用聚类分析方法对网页进行分类，针对网络论坛页面有效实现网页分类，提高网页分类的准确度与效率。

[0075] 为了实现上述目的，本发明采用如下一种技术方案：

[0076] 一种基于网址结构的网络论坛页面聚类方法，该方法包括以下步骤：

[0077] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页

进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0078] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0079] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0080] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0081] 在本实施例中，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。本实施例旨在筛选出主题帖页面。

[0082] 如图1所示，

[0083] 步骤一：对网页进行初步分组。同一论坛内的网址在结构上呈现高度同一性，而不同论坛的网址不具有此性质。在本实施例中，按照网页所属域名对所有网页进行初步分组，以便在后续操作中对不同论坛单独构造解析规则。在后续步骤中，本实施例仅描述对其中一个论坛的处理过程，其他论坛过程与之类似。

[0084] 步骤二：抽样。对于初步分组后的每一组网页进行抽样组成样本，构造解析规则不需要同时处理所有网页，故从所有网页中抽取部分网页进行处理。

[0085] 步骤三：指定主题帖页面。在样本中插入带标记的待筛选网页形成样本网页；为了方便在分类后的簇中快速确定主题帖页面所在的簇，本实施例事先将一个带有标记的主题帖页面插入到样本中形成样本网页。

[0086] 步骤四：构造结构块。经过初步分组，所有页面都隶属于同一论坛，故可以忽略所有网址的域名部分。将样本网页的网址的剩余部分根据符号进行分割，对于每一部分，为其类别和内容分别赋予一个编号，若某类别或内容已有编号，则采用此编号，否则赋予新编号。以此构成该部分的结构块。

[0087] 步骤五：构造结构向量。将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量。

[0088] 构造结构向量的具体方法如下：

[0089] 每个网址都可以转化为一个结构向量，而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：

[0090] $p(u,i)=(t(u,i),v(u,i))$ (8)

[0091] $S(u)=\{p(u,i)\mid i=1,2,\cdots,N\}$ (9)

[0092] 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)为结构块编号元祖，N为总结构块数，S为结构向量。

[0093] 若有如下5条网址在数据集中：

[0094] 1.http://example.com/query.php?id=001&grade=100

[0095] 2.http://example.com/query.php?id＝001&grade＝99

[0096] 3.http://example.com/query.php?id＝002&grade＝100

[0097] 4.http://example.com/query.php?id＝002&grade＝99

[0098] 5.http://example.com/query.php?id＝003

[0099] 分别简记他们为$u_1$到$u_5$,则对类型可以有如下编号：

[0100] 表1类型编号

[0101]

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

[0102] 对值有如下编号：

[0103] 表2值编号

[0104]

| 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 | | |

[0105] 根据式8和式9,先对每个位置构造结构块,后将结构块进行拼合,构造的结构向量为：

[0106] $S(u_1)=[(1,1),(1,2),(1,3),(2,4),(1,5),(2,6)]$

[0107] $S(u_2)=[(1,1),(1,2),(1,3),(2,4),(1,5),(2,7)]$

[0108] $S(u_3)=[(1,1),(1,2),(1,3),(2,8),(1,5),(2,6)]$

[0109] $S(u_4)=[(1,1),(1,2),(1,3),(2,8),(1,5),(2,7)]$

[0110] $S(u_5)=[(1,1),(1,2),(1,3),(2,9),(0,0),(0,0)]$

[0111] 实际上,为了节约编号,每个位置可以采用完全独立的编号,本例中为简化说明而采用的统一编号。

[0112] 步骤六:计算相异度。对样本中任意两个结构向量,计算其相异度,具体计算公式为

[0113] $N_s(u_m,u_n)=\min\{i\,|\,p(u_m,i)\neq p(u_n,i)\}$ (10)

[0114] $$D(u_m,u_n)=\left|\frac{N!}{N_s(u_m,u_n)!}\right| \quad (11)$$

[0115] 其中,$u_m$,$u_n$代表不同的网址,$p(u,i)$为结构块编号元组,N为总结构块数,$D(u_m,u_n)$为网址$u_m$,$u_n$的相异度。

[0116] 步骤七:计算局部密度。对样本中每个结构向量,分别计算其局部密度,具体计算公式为

[0117] $$\rho_i = \sum_{j \in KNN(i)} \exp(-d_{ij}) \quad (12)$$

[0118] 其中，KNN(i)为点i的K近邻集合，$d_{ij}$为结构向量i与结构向量j的相异度，$\rho_i$为结构向量i的局部密度。

[0119] 步骤八：计算最小较高密度结构向量相异度。对样本中每个结构向量，其最小较高密度结构向量相异度即是该结构向量与局部密度大于之，且与之相异度最小的结构向量的相异度，具体计算公式为

[0120] $$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \quad (13)$$

[0121] 其中，$\delta_i$为结构向量i的最小较高密度结构向量相异度。

[0122] 步骤九：作决策图。分别以每个结构向量的ρ值和δ值为横纵坐标做散点图，选择ρ值和δ值均较高的若干结构向量作为聚类的簇中心。

[0123] 步骤十：确定其他结构向量归属。采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；对于每个簇中心，将其K近邻结构向量加入该簇，并加入队列；每次取队首结构向量，对于其K近邻结构向量，若未被分配至任何簇，则加入队首所在簇并加入队列。

[0124] 步骤十一：构造规则。选出带有标记的页面所在的簇，对于结构向量中的每个位置，记录所有该位置的结构块的类别和内容，当出现超过5种不同的内容时，不再记录内容，仅记录类别。其结果为解析规则。

[0125] 步骤十二：应用规则。将解析规则应用于非样本网页，筛选出所有与解析规则相匹配的网页，得到主题帖的分类结构。

[0126] 判断得到的主题帖是否正确，若正确输出网页，否则，返回步骤四重新进行聚类分析。

[0127] 该方法从网址总体中抽取部分样本，利用网络论坛网址高度结构化的特性，对每个网址进行结构划分，构造结构向量，使用本发明提出的距离函数评估结构向量之间的距离，接着使用密度峰值聚类方法对样本结构向量进行聚类分析，提取出每簇的特征结构，构造用于描述簇中所有样本网址的解析器，用于对总体中剩余网址进行解析和分类。本实施例的整体主代码如下：

[0128]

**基于网址结构的网络论坛页面聚类方法**

**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:     α ← specified topic page url in samples
4:     samples ← samples w/ same type of α
5:     vectors ← Array<StructureVector>
6:     for all url in samples do
7:         vectors.add(structure vector of url)
8:     end for
9:     clusters ← CLUSTERING(vectors, K)
10:    topics ← cluster to which α belongs
11:    Rule ← resolve rule from topics
12:    topics ← topic urls identified by Rule
13:    return topics
14: end procedure
```

[0129] 上述代码为本实施例基于网址结构的网络论坛页面聚类方法的整体代码，输入值为网页统一资源定位符urls和近邻数K，输出值为统一资源定位符urls；

[0130] 上述代码第二行为步骤二，从所有网页中抽取部分网页进行处理。步骤一为网页数据预处理过程。

[0131] 上述代码第三行-第四行为步骤三，将一个带有标记的主题帖页面插入到样本中。

[0132] 上述代码第六行-第八行为步骤四和步骤五，构造结构块，并由结构块构造结构向量。

[0133] 上述代码第九行为步骤六到步骤十，进行网页聚类分析，得到簇。

[0134] 上述代码第十行为步骤十一，选出带有标记的页面所在的簇；上述代码第十一行为步骤十一，得到解析规则。

[0135] 上述代码第十二行为步骤十二，将解析规则应用于非样本网页，筛选出所有与解析规则相匹配的网页，得到主题帖的分类结构。

**KNN-PDC聚类分析方法**

**Input** *points*: distance function is available
**Input** $K$: number of neighbors
**Output** *clusters*: result clusters

```
1: procedure CLUSTERING(points, K)
2:     for all i in points do
3:         for all j in points do
4:             d_ij ← distance between i and j
5:             if i in KNN(i) then
6:                 ρ_i ← ρ_i + exp(−d_ij)
7:             end if
8:         end for
9:     end for
10:    for all i in points do
11:        δ_i ← min distance from i to j where ρ_j > ρ_i
12:    end for
13:    DecisionGraph ← plot of δ_i and ρ_i
14:    centers ← points in DecisionGraph w/ high ρ_i and δ_i
15:    clusters ← BFS(points, centers, K)
16:    return clusers
17: end procedure
```

[0136]

[0137] 上述代码为本实施例聚类分析方法的代码，即基于网址结构的网络论坛页面聚类方法的整体代码的第9行的函数代码，描述了步骤六到步骤十的过程。

**广度优先结构向量归属确定方法**

**Input** *points*: distance function available
**Input** *centers*: centers with high $\rho_i$ and $\delta_i$
**Input** $K$: number of neighbors
**Output** *clusters*: result clusters

```
1: procedure BFS(points, centers, K)
2:     clusters ← Array<Cluster>
3:     for all c in centers do
4:         allocate c and KNN(c) to an empty cluster
5:         Queue.push(KNN(c))
6:         while Queue.isEmpty() = False do
7:             q ← Queue.pop()
8:             for all r in KNN(q) do
9:                 if r has not been allocated and d_qr ≤
   mean({d_rj | j ∈ KNN(r)}) then
10:                    allocate r to cluster w/ center c
11:                    Queue.push(r)
12:                end if
13:            end for
14:        end while
15:    end for
16:    return clusters
17: end procedure
```

[0138]

[0139] 上述代码为本实施例广度优先结构向量归属确定方法的代码，即基于网址结构的网络论坛页面聚类方法的整体代码的第9行的函数代码，聚类分析方法的代码的第15行的函数代码，描述了步骤十的过程。

[0140] 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：

[0141] $$R = \frac{N_{et}}{N_t} \qquad (14)$$

[0142] $$P = \frac{N_{et}}{N_e} \qquad (15)$$

[0143] $$F = \frac{2PR}{P+R} \qquad (16)$$

[0144] 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。F值越高，则说明信息提取效果越好。

[0145] 本例中对十个论坛的聚类结果如表3所示。

[0146] 表3聚类方法结果

[0147]

| 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 |

[0148] 从表中可得，本发明提出的网页聚类方法的准确率均为100％，召回率均在90％以上，F值均在95％以上，能够准确地对网页进行分类，其结果令人满意。

[0149] 实施例2：

[0150] 本发明的第二目的是提供一种基于网址结构的网络论坛页面聚类方法的存储设备。

[0151] 为了实现上述目的，本发明采用如下一种技术方案：

[0152] 一种存储设备，其中存储有多条指令，所述指令适于由处理器加载并执行：

[0153] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0154] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0155] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0156] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0157] 实施例3：

[0158] 本发明的第三目的是提供一种基于网址结构的网络论坛页面聚类方法的终端设备。

[0159] 为了实现上述目的，本发明采用如下一种技术方案：

[0160] 一种终端设备，包括：

[0161] 处理器，适于实现各指令；以及

[0162] 存储设备，适于存储多条指令，所述指令适于由处理器加载并执行：

[0163] (1)按照网页所属域名对所有网页进行初步分组，对于初步分组后的每一组网页进行抽样组成样本，并在样本中插入带标记的待筛选网页形成样本网页；

[0164] (2)将样本网页的除域名外的网址根据符号进行分割，对分割后的每一部分网址的类别和内容进行编号，构造出结构块；

[0165] (3)将同一网址的各个结构块按顺序依次排列，构成该网址的结构向量；计算样本网页中的任意两个结构向量的相异度，和网页样本中最小较高密度结构向量相异度即任意一个结构向量与大于其局部密度且与其相异度最小的结构向量的相异度；分别作为横坐标和纵坐标构造决策图，确定簇中心，采用广度优先结构向量归属确定法确定非簇中心结构向量的归属簇；

[0166] (4)根据步骤(3)的决策图构造出解析规则，将解析规则应用于初步分组后的每一组网页中的非样本网页，进行网页聚类筛选，并采用评价指标进行评价。

[0167] 本发明的有益效果：

[0168] 本发明的一种基于网址结构的网络论坛页面聚类方法及设备，根据网址构造结构向量，并计算结构向量之间的相异度，使得可以使用聚类分析方法对网页进行分类，具有高准确度和执行效率。尤其针对共性不显著的论坛页面，本发明构造结构向量进行聚类分析，满足大规模分类的速度需求。

[0169] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

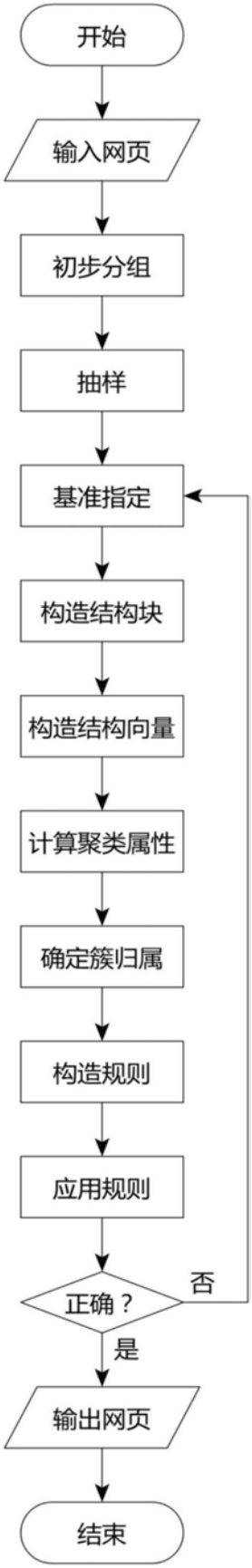
开始
输入网页
初步分组
抽样
基准指定
构造结构块
构造结构向量
计算聚类属性
确定簇归属
构造规则
应用规则
正确？
否
是
输出网页
结束

图1