(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101763404 B
(45)授权公告日 2012.03.21

(21)申请号 200910219439.2

(22)申请日 2009.12.10

(73)专利权人 陕西鼎泰科技发展有限责任公司
地址 710072 陕西省西安市科技路 30 号合力紫郡大厦 B-2001 室

(72)发明人 赵安军 王磊 王礼 杨宗良

(74)专利代理机构 西北工业大学专利中心 61204
代理人 顾潮琪

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

(56)对比文件
CN 101571868 A,2009.11.04, 全文.
王晓勇等. 因特网文本智能挖掘的模糊聚类算法研究. 《计算机仿真》.2009, 第 26 卷(第 7 期), 第 216-219 页.
耿新青,王正欧. TGFCM:基于模糊聚类的中文文本挖掘的新方法. 《计算机工程》.2006, 第 32 卷(第 5 期), 第 7-9 页.

审查员 张蕾

权利要求书 2 页 说明书 5 页

(54)发明名称

基于模糊聚类的网络文本数据检测方法

(57)摘要

本发明公开了一种基于模糊聚类的网络文本数据检测方法,先对提取的网络内容进行预处理;对需要聚类的预处理后网络内容进行特征提取后对网络内容进行聚类,设定初始聚类数。在聚类过程中,一个聚类数对应一个隶属度矩阵,每个隶属度矩阵都有一个平均信息熵值,平均信息熵基于密度函数选择初始聚类中心,算法迭代过程中修改聚类数,当平均信息熵达到最小值时,所对应的聚类数为最佳聚类数。最后将聚类结果返回给用户。本发明具有高效的智能聚类效果,并且可以根据应用的不同,调整聚类的精度,兼顾聚类的速度。

CN 101763404 B

1. 基于模糊聚类的网络文本数据检测方法，其特征在于包括下述步骤：

(1) 首先对原始的网络文本进行分词，然后计算每个词出现的频率，删除所有出现频率超过 10 的功能词；

(2) 应用向量空间模型作为网络内容特征的表示方法，将网络内容文档中的词条项在整个网络内容文档中出现的频次作为该词条项的权重，将所有的词条项以及词条项所占的权重作为网络内容空间的一个特征向量，将网络内容空间作为一组正交词条向量所组成的向量空间；词条向量 $V(d) = (t_i, w_i(d); i = 1,2,\ldots,n)$，其中，n 表示降维和分词后整个网络内容文档的词条数目，d 表示此网络内容文档，t_i 为词条项，$w_i(d)$ 为词条在此网络内容文档中所占的权重；

(3) 对网络内容进行模糊聚类，具体步骤如下：

步骤 1，设定初始聚类数 c，$c \geq 2$；将迭代次数 b 设置为零，并且选择指数权重 m 和迭代停止阈值 ε，m 在 1.5 到 2.5 之间，ε 在 0.1 到 0.001 之间；

步骤 2，对于网络内容空间中具有 n 个样本的数据集合 $X = \{x_l, l = 1,2,...,n\}$，在 x_l 处的密度函数定义为：$D_l^0 = \sum_{k=1}^{n} \frac{1}{1 + f_d \bullet \|x_l - x_k\|^2}$，其中，$f_d = 1/r_d^2$，$r_d$ 为类密度有效邻域半径，$r_d = \alpha \bullet \frac{1}{2} \mathop{\wedge}_{k=1}^{n} \mathop{\vee}_{l=1}^{n} \|x_l - x_k\|$，α 取值范围为 [0,1]；令 $D_1^* = \max\{D_l^0; l = 1,2,...,n\}$，$x_l^*$ 是对应 D_1^* 的样本点，并且取为第一个聚类中心；设 $D_k^* = \max(D_l^{k-1}; l = 1,2,...,n)$，$x_k^*$ 是对应 D_k^* 的样本点，k = 1,2,...,c-1，$D_l^k = D_l^{k-1} - D_k^* \frac{1}{f_d \bullet \|x_l - x_k^*\|}$，$x_k^*$ 作为第 k 个初始聚类中心；

步骤 3，通过公式 $u_{ij}^b = 1 \Big/ \sum_{k=1}^{c} (\frac{d_{ij}^b}{d_{kj}^b})^{\frac{2}{m-1}}$ 计算隶属度，其中 u_{ij}^b 为在第 b 次迭代中样本 j 属于类 i 的隶属度，c 为聚类数，d_{ij} 表示第 j 个元素到第 i 个聚类中心的欧式距离；对计算获取隶属度增加一个权值，形成新的隶属度，改进后的隶属度 $U_{ij}^b = \lambda u_{ij}^b + (1-\lambda) u_{ij}^{b2}$，λ 的取值为 [0,1]；

步骤 4，根据上述计算的隶属度 u_{ij}^b 以及通过权值形成的改进后隶属度 U_{ij} 对聚类中心进行更新，更新后的聚类中心 $C_i^{b+1} = \frac{\sum_{j=1}^{n} (U_{ij}^{b+1})^m \cdot x_j}{\sum_{j=1}^{n} (U_{ij}^{b+1})^m}$，i = 1,2,...,c，并且判定迭代停止阈值的条件 $\|C_i^b - C_i^{b+1}\| < \varepsilon$ 是否满足，如果满足，输出隶属度矩阵和形成的聚类中心，否则令 b = b+1，并转向步骤 3；

步骤 5，计算平均信息熵 $H = \sum_{i=1}^{C} \sum_{j=1}^{n} \{[U_{ij} \times \log b(U_{ij}) + (1-U_{ij}) \times \log b(1-U_{ij})]/n\}$，以步骤 4 的输出为输入，当平均信息熵达到最小值时，所对应的聚类数即为最佳聚类数，聚类过程结束，保存最终聚类数目 c 以及聚类中心 C_i，i = 1,2...,c；否则，令 c = c+1 并转向步骤 2；

(4) 将聚类结果返回给用户，聚类结果包括聚类中心的数目以及聚类中心。

2. 根据权利要求 1 所述的基于模糊聚类的网络文本数据检测方法，其特征在于：所述的初始聚类数 c 选取为 2。

3. 根据权利要求 1 所述的基于模糊聚类的网络文本数据检测方法，其特征在于：所述的选择指数权重 m 为 1.9，迭代停止阈值 ε 选择 0.01。

基于模糊聚类的网络文本数据检测方法

技术领域

[0001] 本发明涉及一种数据检测方法，尤其是一种网络文本数据的检测方法。

背景技术

[0002] 网络内容中有 80％左右的信息是文本形式，所以对文本数据挖掘技术的研究成为数据挖掘中的一个日益流行且十分重要的研究课题。网络内容聚类是将网络内容中相似的文本分为一组的全自动处理过程，它是一个无监督学习过程。聚类的目的是将物理或抽象的对象，按对象间的相似性进行区分和分类。聚类方法按对数据划分的形式可分为 ：划分时有明确的边界称为硬划分，即将数据划分到一个确定的类 ；没有明确的边界的划分称为模糊划分，即将给定数据以隶属度的形式表示属于哪几个类。

[0003] 我国文本智能分类的研究起始于 20 世纪 80 年代，大体经历了可行性探讨、辅助分类系统、自动分类系统三个阶段。中文文本分类还处于在试验研究阶段，正确分类率约为 70％ -90％，正在逐渐向商业化的软件应用靠拢，并已经尝试开发了一批自动分类系统，例如清华大学吴军研制的自动分类系统、山西大学刘正瑛等人开发的金融自动分类系统、上海交大的西风文本自动分类系统。如何找到合理的应用并且在实践中逐步改善算法，提高性能成为文本分类算法的当务之急。通过文献检索发现，目前国内外常用的文本分类方法大多数是基于文本内容的相似度对文本进行分类。诸如基于概念的文档分类算法、K- 最近邻接参照分类算法 (K-NN)、贝叶斯分类算法、基于语义网络的概念推理网分类算法以及决策树和支持向量机 (SVM) 等方法。基于这些方法的网络内容分类系统大都是基于平面的分类，即多采用基于词或词串信息的动态聚类方法和基于特征属性的分类技术来实现，挖掘的深度不够，执行速度慢，聚类的准确度较低。

发明内容

[0004] 为了克服现有技术挖掘的深度不够、执行速度慢、聚类的准确度较低等不足，本发明提供一种基于模糊聚类的网络文本数据检测方法，能够有效提高网络安全审计中对于文本分类的精度与可靠性，从而改善网络内容中目标文本的获取效率，实现网络内容的智能检索。

[0005] 本发明解决其技术问题所采用的技术方案是 ：首先对提取的网络内容进行预处理 ；其次，对需要聚类的预处理后网络内容进行特征提取 ；然后，对网络内容进行聚类，设定初始聚类数。在聚类过程中，一个聚类数对应一个隶属度矩阵，每个隶属度矩阵都有一个平均信息熵值，平均信息熵基于密度函数选择初始聚类中心，算法迭代过程中修改聚类数，当平均信息熵达到最小值时，所对应的聚类数为最佳聚类数。最后，将聚类结果返回给用户。

[0006] 本发明具体包括以下步骤 ：

[0007] (1) 网络内容预处理 ：如果以原始的网络内容作为特征向量提取的对象，那么，网络内容的特征向量维数会相当大，因此，必须进行降维的处理。降维的方法采用特征抽取方

式，首先对原始的网络文本进行分词，然后，计算每个词出现的频率，删除所有出现频率超过 10 的功能词，从而降低网络内容特征提取时所获取特征向量的维度。由于特征向量维数降低，不但能加快聚类算法计算的速度，而且还能提高分类结果的精度和避免重复匹配问题。

[0008] (2) 网络内容特征提取：应用向量空间模型作为网络内容特征的表示方法。在该模型中，网络内容空间被看作是由一组正交词条向量所组成的向量空间。所述的词条向量是指将每次捕获到网络流的网络内容作为一篇网络内容文档，经过步骤 (1) 的网络内容预处理后，将网络内容文档中的词条项在整个网络文档中出现的频次作为该词条项的权重，将所有的词条项以及词条项所占的权重作为网络内容空间的一个特征向量。词条向量表示为 $V(d) = (t_i, w_i(d); i = 1,2,\ldots,n)$，其中，n 表示降维和分词后整个网络文档的词条数目，d 表示此网络文档，t_i 为词条项，$w_i(d)$ 为词条在此网络内容文档中所占的权重，被定义为 t_i 在 d 中出现的频率。

[0009] (3) 模糊聚类：现有技术的模糊聚类方法存在对孤立点数据比较敏感，须预先指定聚类数目和模糊加权指数的缺陷。为降低孤立点对聚类结果的影响，本发明对数据对象的隶属度增加一个权值，使隶属度的值高的数据对象对聚类中心位置的影响增大，隶属度小的数据对象降低它们对聚类中心的影响。模糊聚类的具体步骤如下：

[0010] 步骤 1，设定初始聚类数为 c，初始聚类数大于等于 2 即可，一般选取 2；将迭代次数 b 设置为零，并且选择指数权重 m 和迭代停止阈值 ε，指数权重 m 的选择范围在 1.5 到 2.5 之间，在此方法中，选择指数权重 m 为 1.9，迭代停止阈值 ε 的选择范围在 0.1 到 0.001 之间，在这里考虑到算法的执行速度和聚类的精度，ε 选择 0.01。

[0011] 步骤 2，由于聚类结果受到初始聚类数目和初始聚类中心的影响，本发明采用基于密度函数选择初始聚类中心的方法。对于网络内容空间中具有 n 个样本的数据集合 $X = \{x_l, l = 1,2,\ldots,n\}$，在 x_l 处的密度函数定义为：$D_l^{(0)} = \sum_{k=1}^{n} \frac{1}{1 + f_d \bullet \|x_l - x_k\|^2}$，其中，$f_d = 1/r_d^2$，$r_d$ 为类密度有效邻域半径，$r_d = \alpha \bullet \frac{1}{2} \bigwedge_{k=1}^{n} \bigvee_{l=1}^{n} \|x_l - x_k\|$，α 与样本集合分布特性有关，取值范围为 [0,1]，在邻域半径 r_d 之外的数据点对 x_l 的密度的计算影响很小。密度函数越大，表示在点 x_l 的周围聚集的样本点越多，说明点 x_l 处的密度越大，从而 $D_l^{(0)}$ 的值越高。令 $D_1^* = \max\{D_l^0; l = 1,2,\ldots,n\}$，$x_1^*$ 是对应 D_1^* 的样本点，并且取为第一个聚类中心。设 $D_k^* = \max(D_l^{k-1}; l = 1,2,\ldots,n)$，$x_k^*$ 是对应 D_k^* 的样本点，k = 1，2，...，c-1，$D_l^k = D_l^{k-1} - D_k^* \frac{1}{f_d \bullet \|x_l - x_k^*\|}$，$x_k^*$ 作为第 k 个初始聚类中心。

[0012] 步骤 3，计算隶属度。通过公式 $u_{ij}^b = 1 \Big/ \sum_{k=1}^{c} \left(\frac{d_{ij}^b}{d_{kj}^b}\right)^{\frac{2}{m-1}}$ 计算隶属度。其中，u_{ij}^b 为在第 b 次迭代中样本 j 属于类 i 的隶属度，b 为迭代次数，m 为指数权重，c 为聚类数，d_{ij} 表示第 j 个元素到第 i 个聚类中心的欧式距离。为降低孤立点对聚类结果的影响，对计算获取隶属度增加一个权值，形成新的隶属度，使隶属度值高的数据对象对聚类中心位置的影响增大，对于隶属度小的数据对象则降低它们对聚类中心的影响。隶属度的改进公式为：改进后的

隶属度$U_{ij}=\lambda u_{ij}+(1-\lambda)u_{ij}^2$，λ 的取值为 [0,1]，λ 取值与聚类精度和算法执行速度有关，使用时可以根据聚类的精度和聚类时间进行调整。当 λ = 1 时，$U_{ij}=u_{ij}$，当 $u_{ij}=0$，$U_{ij}=0$，当 $u_{ij}=1$，$U_{ij}=1$。在 [0,1] 区间的隶属度在改进后有一定程度的减少。在算法迭代过程中，隶属度值越小，改进后隶属度相应减少地越明显，隶属度小的数据对象对聚类中心的影响降低了；隶属度越大，改进后的隶属度相应减少的较小，这样就相对的提高隶属度值高的数据对象对于聚类的中心位置的影响。

[0013] 步骤 4，更新聚类中心。根据上述计算的隶属度 u_{ij} 以及通过权值形成的改进后隶属度 U_{ij} 对聚类中心进行更新，更新公式为：$C_i^{b+1}=\frac{\sum_{j=1}^{n}(U_{ij}^{b+1})^m\cdot x_j}{\sum_{j=1}^{n}(U_{ij}^{b+1})^m}$，i = 1,2,...,c，$C_i^{b+1}$ 为更新后的聚类中心，m 为权重指数。并且通过 $\left\|C_i^b-C_i^{b+1}\right\|<\varepsilon$ 判定迭代停止阈值的条件是否满足，如果满足，输出隶属度矩阵和形成的聚类中心，否则令 b = b+1，并转向步骤 3。

[0014] 步骤 5，本发明以隶属度的平均信息熵作为评判聚类数目的标准，平均信息熵的定义为 $H=\sum_{i=1}^{C}\sum_{j=1}^{n}\left\{[u_{ij}\times Ib(u_{ij})+(1-u_{ij})\times Ib(1-u_{ij})]/n\right\}$，其中，C 为设定的初始聚类数目，n 为聚类的样本数目，b 为迭代次数，u_{ij} 为样本 j 属于类 i 的隶属度，I 表示熵的计算。当平均信息熵达到最小值时，所对应的聚类数即为最佳聚类数。以步骤 4 的输出为输入，按照上述的最佳聚类数的评判标准判定是否满足，如果满足聚类数评判标准，聚类过程结束，保存最终聚类数目 c 以及聚类中心 C_i，i = 1,2...c,。否则，令 c = c+1 并转向步骤 2。

[0015] (4) 聚类结果输出。将聚类结果返回给用户，聚类结果包括聚类中心的数目以及聚类中心。

[0016] 本发明的有益效果是：本发明是在性能良好的网络内容特征提取技术、基于密度函数获取初始聚类中心技术、优化的隶属度计算技术以及聚类数的评判标准确定技术的基础上研发的。与已有的相应技术相比，该技术具有高效的智能聚类效果，并且可以根据应用的不同，调整聚类的精度，兼顾聚类的速度。

[0017] 下面结合实施例对本发明进一步说明。

具体实施方式

[0018] 本发明具有网络内容预处理、网络内容特征提取、模糊聚类以及聚类结果输出四个部分的功能。其中网络内容预处理完成对多维的网络内容文档特征向量进行降维处理，进行特征抽取；网络内容特征提取完成对所捕获网络流中网络内容的处理，包括网络内容文档的建立，文档的特征向量表示；模糊聚类是本发明的核心，采用基于密度函数选择初始聚类中心，平均信息熵作为评判聚类数目的标准，设定初始聚类数，在算法的迭代过程修改聚类数，当平均信息熵达到最小值时的聚类数即为最佳聚类数，完成对网络内容文档的聚类。聚类结果输出将聚类结果返回给用户，包括聚类中心的数目以及类别信息。

[0019] 基于本发明开发了原型系统，该系统执行包括以下步骤：网络内容预处理、网络内容特征提取、设定初始聚类参数、选择初始聚类中心、隶属度计算、聚类中心更新、聚类结果评价以及聚类结果输出。

[0020] 本发明具体包括以下步骤：

[0021] 第一步，将待聚类的网络内容分割成 1000 篇文本，对每篇文档进行标点分析，把它们分成单句；并删除出现频率超过 10 次的功能词，对每个单句利用文本分析工具 PatCount 对其中每个词进行词法分析，对每个单句利用 n-gram 方法得到所有由三个以内词组成的词条短语，在这里 n-gram 方法所述的 n 为 3。

[0022] 第二步，应用向量空间模型作为网络内容特征的表示方法，将网络内容文档中的词条项在整个网络文档中出现的频次作为该词条项的权重，将所有的词条项以及词条项所占的权重作为网络内容空间的一个特征向量。统计所获取的词条短语数目以及各词条短语在网络文本中出现的频次，将 1000 篇经过第一步处理过的网络文本表示成文本向量，由此组成维数为 3768 的网络文本特征向量 $V(d) = (t_i, w_i(d); i = 1,2,...,3768)$，d 表示 1000 篇网络文档集合，$t_i$ 为集合当中的一个词条项，$w_i(d)$ 为此词条在此网络内容文档中所占的权重，被定义为 t_i 在 d 中出现频率。

[0023] 第三步包含以下步骤：

[0024] 步骤 1：设定初始的聚类数为 2，将迭代次数设置为 0，并且选择指数权重为 1.9 和迭代停止阈值为 0.01；

[0025] 步骤 2：根据上述设定的初始聚类数，以 3768 维的网络文本特征向量为输入计算 2 个初始的聚类中心。对于网络内容空间中具有 3768 个样本的数据集合 $X = \{x_1, l = 1, 2, ..., 3768\}$，在 x_1 处的密度函数定义为：$D_l^{(0)} = \sum_{k=1}^{3768} \frac{1}{1 + f_d \bullet \|x_l - x_k\|^2}$，其中，$f_d = 1/r_d^2$，$r_d$ 为类密度有效邻域半径，$r_d = \alpha \bullet \frac{1}{2} \mathop{\wedge}_{k=1}^{3768} \mathop{\vee}_{l=1}^{3768} \|x_l - x_k\|$，α 与样本集合分布特性有关，在这里取为 0.9。令 $D_1^* = \max\{D_l^0; l = 1,2,...,3768\}$，$x_1^*$ 是对应 D_1^* 的样本点，并且取为第一个聚类中心。设 $D_2^* = \max(D_l^1; l = 1,2,...,3768)$，$x_2^*$ 是对应 D_2^* 的样本点，$D_l^1 = D_l^0 - D_1^* \frac{1}{f_d \bullet \|x_l - x_1^*\|}$，$x_2^*$ 作为第 2 个初始聚类中心。

[0026] 步骤 3：，计算隶属度，通过公式 $u_{ij}^b = 1 \Big/ \sum_{k=1}^{c} (\frac{d_{ij}^b}{d_{kj}^b})^{\frac{2}{1.9-1}}$ 计算隶属度。其中，u_{ij}^b 为在第 b 次迭代中样本 j 属于类 i 的隶属度，b 为迭代次数，c 为聚类数，d_{ij} 表示第 j 个元素到第 i 个聚类中心的欧式距离。为降低孤立点对聚类结果的影响，对计算获取的数据对象的隶属度增加一个权值，形成新的隶属度，使隶属度值高的数据对象对聚类中心位置的影响增大，对于隶属度小的数据对象则降低它们对聚类中心的影响，改进隶属度公式为：$U_{ij} = \lambda u_{ij} + (1 - \lambda) u_{ij}^2$，这里 λ 取值 0.8。

[0027] 步骤 4：根据上述计算的隶属度以及通过权值形成的新的隶属度对聚类中心进行更新，判断本次更新的聚类中心和上一次聚类中心的差是否小于迭代停止阈值，如果满足，输出隶属度矩阵和形成的聚类中心，否则，迭代次数加 1，跳转到步骤 3 重新计算隶属度，循环上述过程。本次试验中的上述循环过程执行 3 次，执行时间 1 分钟；

[0028] 步骤 5：以步骤 4 的输出为输入，计算平均信息熵是否最小，如果最小，此时获得的聚类数目为当前的聚类数减 1，聚类过程结束，保存最终聚类数目 c 以及聚类中心 C_i，i = 1，

2...,c。否则转向步骤 2,并且给当前的聚类数加 1,重新计算初始的聚类中心,循环上述过程,本次试验的上述循环过程执行 4 次,执行时间为 4 分钟。

[0029] 第四步,将聚类结果返回给用户,包括聚类中心的数目以及各个聚类中心。本实施例获取的聚类数为 5,各个聚类中心类别分别是计算机、金融、交通、体育以及军事五大类。

[0030] 本方法经过原型系统的具体实施,效果较好。采用基于密度函数选择初始聚类中心,平均信息熵作为评判聚类数目的标准,在算法的迭代过程修改聚类数,当平均信息熵达到最小值时的聚类数即为最佳聚类数,完成对网络内容文档的聚类。这些方法的使用使得分类的准确性有了较大的提高,并且在执行的速度方面也有一定的改善。