基于语义的高维数据聚类技术

刘 铭,王晓龙,刘远超
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 本文提出一种有效处理高维数据的聚类算法,算法首先通过构造特征链将文档集合划分为多个类别,同时在相似度计算及权值调整时考虑相似特征的影响以凝聚语义相似的文档,并动态调整文档权重使分布不平衡的文档得到充分训练.实验表明:该算法在高维空间能够获得较好的聚类结果,类内相似度高,类间区分性好,迭代次数较少.

关键词: 语义相似度计算; 自组织映射; 特征链; 权值调整
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2009) 05-0925-05

Clustering Technology for High Dimensional Data Based on Semantics

LIU Ming, WANG Xiao-long, LIU Yuan-chao
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: A novel clustering algorithm for high dimensional data is proposed in this paper. This algorithm first partitions input document set into some clusters by constructing feature chains. Simultaneously it also considers the effects of similar features in similarity computation and weight adjustment to agglomerate documents with semantic similarities, and dynamically adjusts weights of documents to make unbalanced documents well trained. Experiment results demonstrate that it can obtain relatively better clustering results with high intra-cluster agglomeration and inter-cluster distinctness, and also has less iterative times.

Key words: semantic similarity computation; self organizing mapping; feature chain; weight adjustment

1 引言

文本聚类是信息检索和文本挖掘领域的核心问题,其目的是将相似的文档凝聚到一类.目前针对文本聚类已经进行了大量的研究,提出了许多高性能的聚类算法[1,2],但是大多数聚类算法仅在少量数据形成的低维特征空间中拥有较好的聚类结果.随着网络的发展,人们接触的信息与日俱增,只有对大规模高维数据的聚类处理才具有现实意义.然而随着文档规模的增大,高维特征空间中存在大量相似特征,但是传统相似度计算方法仅仅考虑了文档间的同现特征,而忽略了相似特征的影响.同时发现随文档数增多,文档的分布也越发不均衡.如果我们在自组织聚类中将每篇文档均按照相同的重要性进行聚类,聚类结果会严重偏向含有较多文档数的文档类,会将该文档类分裂为多个相似的小类别,形成"过适应"现象,而那些文档数较少的文档类却错误地分散于多个不相关的类别中,形成"欠适应"现象[3].针对上述问题,本文提出一种基于语义的高维数据聚类算法(SHSOM),算法首先通过计算特征同现词向量间的交叉熵获得特征的上下文语义相似度,在此基础上将相似的特征构造成特征链并通过反映不同信息的特征链将文档集合划分为多个类别;使用自组织映射思想对文档类进行调整以提高划分的准确性,同时在相似度计算及权值调整时考虑了位于相同特征链内的相似特征的影响以凝聚具有语义相似性的文档.为解决文档分布不均匀的问题,算法在训练阶段逐步调整文档的权值,使错误划分的文档得到充分训练从而提高了聚类结果的有效性.

2 类别初始划分

本文首先对待聚类文档进行分词及停用词过滤获得文档的特征集合,然后通过计算特征的统计量选择大于一定权值的特征构造特征空间,具体方法可参见文献[4].以词汇链技术凝聚特征空间中的相似特征,并通过反映不同信息的特征链将文档集合划分为多个初始类别.

2.1 特征相似度计算

汉语语言学家认为:"一个词的语法功能就是指词的(语法)分布".词的上下文环境是一种典型的分布,当两个词的上下文分布基本相同时,其语义相似度是很大

收稿日期:2008-04-29;修回日期:2008-08-06
基金项目:国家自然科学基金重点项目(No. 60435020);国家 863 高技术研究发展计划(No. 2006AA01Z197, No. 2007AA01Z172)

的[5]. 本文即通过统计特征在语料中的上下文同现词及同现概率获得特征的同现词向量,通过计算不同特征的同现词向量间的交叉熵反映不同特征的语义相似度.

$$SimF(f_t, f_s) = 1 - H(FV(f_t), FV(f_s)) \quad (1)$$

公式(1)为特征 f_t 和 f_s 的相似度计算公式,其中 $H(FV(f_t), FV(f_s))$ 为 f_t 和 f_s 对应的同现词向量间的交叉熵,具体含义可参见公式(2).

$$H(FV(f_t), FV(f_s)) = -\sum_{i=1}^{n}\frac{(p_i + q_i)}{2}\log_2\frac{(p_i + q_i)}{2} + \frac{1}{2}\left[\sum_{i=1}^{n}p_i\log_2 p_i + \sum_{i=1}^{n}q_i\log_2 q_i\right] \quad (2)$$

公式(2)为特征 f_t 和 f_s 的同现词向量间的交叉熵公式. 由公式(2)可见:交叉熵的值位于 0～1 之间,且随着特征上下文同现词向量的差异越大其值越大,这说明交叉熵能够作为衡量特征的上下文是否相同的量度[6].

2.2 特征链构造及类别划分

使用上述方法得到不同特征的相似度后,就可以通过词汇链技术将相似的特征组织为一条特征链,并通过反映不同信息的特征链将文档集合划分为多个初始类别.

下面介绍了特征链的构造过程:

(1) 设特征空间为 *FSpace*,特征链集合为 *Chain*.

(2) 顺序扫描 *FSpace*,设此时正在扫描 *FSpace* 中的第 t 个特征 f_t.

(3) 按公式(3)计算 f_t 和 *Chain* 中每条链的相似度,检测最大相似度是否超过阈值,如超过,则将 f_t 插入到与此特征具有最大相似度的链中,否则新建一条仅包含 f_t 的特征链.

(4) 如不到 *FSpace* 的末尾,则按步骤[2][3][4]处理第 $t+1$ 个特征,否则结束.

$$SimChain(f_t, chain_j) = \max_{cw_i \in chain_j}(SimF(f_t, cw_i)) \quad (3)$$

公式(3)中我们以特征 f_t 和特征链 $chain_j$ 所包含的特征词的最大相似度作为 f_t 和 $chain_j$ 的相似度,cw_i 为 $chain_j$ 所包含的第 i 个词. 相似度阈值的设定可参见文献[7].

按如上方法获得的每条特征链均近似描述了文档集合的某一类别信息,本文即通过计算文档和特征链的相似度将文档集合划分为多个初始类别. 划分时以特征链作为类别代表,将文档映射到与此文档具有最大相似度的特征链代表的类别中.

公式(4)为文档 doc_k 和特征链 $chain_j$ 的相似度计算公式,设 doc_k 和 $chain_j$ 的特征集合的交集为 IF_{kj},大小为 $|IF_{kj}|$. $W(f, chain_j)$ 为 IF_{kj} 中的特征 f 在 $chain_j$ 中的权值,其为 f 在文档集合中各文档的权值的平均值.

$$Sim(doc_k, chain_j) = \begin{cases} 0 & \text{If } |IF_{kj}| < SW; \\ \sum_{f \in IF_{kj}} \frac{W(f, doc_k) \times W(f, chain_j)}{W(f, doc_k) + W(f, chain_j)} \times \log_2(|IF_{kj}|) & \text{If } |IF_{kj}| \geq SW; \end{cases} \quad (4)$$

可以看出如果文档和特征链的特征集合的交集小于 *SW*,则认为文档和特征链是不相关的. 如果大于 *SW*,则从两方面衡量文档和特征链之间的相似度. 以第二个乘号为界,第一部分计算了特征在文档和特征链中权值的差异性;第二部分计算了文档和特征链的特征集合的交集大小. 上文给定的阈值 *SW* 为文档与特征链是否相似的判别标准. 实验发现:如果文档和特征链的特征集合的交集大于文档特征集合大小的 1/3,则文档和特征链反映的信息较为相似,因此本文设 *SW* 的值为 $|FSet(doc_k)|/3$, $FSet(doc_k)$ 为文档 doc_k 的特征集合.

3 自组织训练阶段

本文采用自组织映射算法对文中 2.2 节形成的文档类进行调整. 自组织映射聚类(SOM)是一种无导师的自组织自学习网络,其以神经元(NEURON)作为文档类的代表,能够根据文档的分布逐步收敛到最佳的类别划分[8].

以上文形成的每条特征链对应于一个神经元,然后将神经元首尾相连组织为图 1 所示的扇形结构.

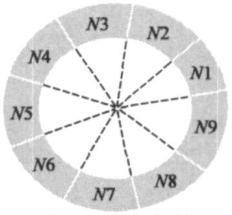


图1 扇形结构组织神经元

使用 SOM 进行文档聚类大致可分为以下几步:

(1) 首先对输出层各神经元的权向量赋小的随机数,并归一化.

(2) 从文档集合中随机选取文档作为 SOM 网络的输入.

(3) 计算输入文档与各神经元的相似度,相似度最大的神经元获胜.

(4) 调整获胜神经元及其邻域内神经元的权值,权值调整一般采用随时间单调下降的退火函数.

传统 SOM 算法在步骤 3 相似度计算和步骤 4 神经元权值调整中均仅考虑了文档与神经元的同现特征,而忽略了相似特征的影响. 然而在进行文档聚类时,由于相似特征描述的含义大致相同,其在神经元中权值的差距也不应很大,因此本文在相似度计算和权值调整时考虑了相似特征(即位于相同特征链内的特征)的影响.

3.1 相似度计算

本文以公式(5)计算文档 doc_k 和神经元 n_l 的相似度并通过其选择获胜神经元进行调整.

$$Dist(doc_k, n_l) = \begin{cases} \sum_i (W(f_t, doc_k) - W(f_t, n_l))^2, & \text{If } f_t \in doc_k; \\ \sum_s (SimF(f_t, f_s) \times W(f_t, doc_k) - W(f_s, n_l))^2 & \\ \quad \text{If } f_s \notin doc_k \&\& f_t \in doc_k \&\& f_s, f_t \text{ 在相同特征链内}; & \\ \sum_s (\frac{\sum_{t=1}^{m} SimF(f_t, f_s) \times W(f_t, doc_k)}{m} - W(f_s, n_l))^2 & \\ \text{If } f_s \notin doc_k \&\& f_t \in doc_k (t = 1 \sim m) \&\& f_s, f_t \text{ 在相同特征链内} \end{cases} \quad (5)$$

公式(5)分为三部分,第一部分仅处理了出现在文档 doc_k 中的特征 f_t,第二部分为与 f_t 相似且不在 doc_k 中出现的特征 f_s 的处理方法. 然而现实应用中很可能出现 doc_k 的特征集合中有多个特征与 f_s 相似,此时按照公式(5)的第三部分计算 f_s 的权值. 而 doc_k 和 n_l 的相似度为上述三部分相似度的叠加. 由实验部分图 2 和图 3 可见:当我们将相似特征融入相似度计算公式后,可以寻找具有语义相似性的文档和神经元从而获得了更好的聚类结果.

3.2　神经元权值调整

公式(6)介绍了神经元 n_l 中特征权值的调整方法. 其中第一部分描述了在文档 doc_k 中出现的特征 f_t 的权值调整方法,详细介绍可参见文献[8]. 第二部分和第三部分描述了不在文档 doc_k 中出现的特征 f_s 的权值调整方法,f_s 的权值赋值方式与公式(5)相同.

$$\begin{cases} W(f_t, n_l)(t+1) = W(f_t, n_l)(t) + a(t)\, h(t)\, (W(f_t, doc_k) \\ \quad - W(f_t, n_l)(t)), \text{If } f_t \in doc_k; \\ W(f_s, n_l)(t+1) = W(f_s, n_l)(t) + a(t)\, h(t)\, (SimF(f_t, f_s) \\ \quad \times W(f_t, doc_k) - W(f_s, n_l)(t)), \\ \text{If } f_s \notin doc_k \&\& f_t \in doc_k \&\& f_s, f_t \text{ 在相同特征链内}; \\ W(f_s, n_l)(t+1) = W(f_s, n_l)(t) + a(t)\, h(t) \\ (\frac{\sum_{t=1}^{m} SimF(f_t, f_s) \times W(f_t, doc_k)}{m} - W(f_s, n_l)(t)), \\ \text{If } f_s \notin doc_k \&\& f_t \in doc_k (t = 1 \sim m) \&\& f_s, f_t \text{ 在相同特征链内} \end{cases} \quad (6)$$

3.3　文档权值调整

如文献[3]所述:如果文档的分布不均匀,那么自组织聚类中会形成含有较多文档数的文档类被分裂为多个相似小类别的"过适应"现象,和文档数较少的文档类被分散到不相关文档类中的"欠适应"现象. 针对上述问题,本文采用文档权值调整技术,根据文档被错误划分的次数调整文档的权值,这样可以尽量避免上述"过适应"和"欠适应"现象.

按公式(5)计算文档与神经元的相似度,如果文档和每个神经元的相似度均相差不多则说明此文档位于神经元形成的文档类的边缘位置,是错误划分的文档. 设文档 doc_k 被错误划分的次数为 $count_k$,如果 doc_k 在某次训练时为错误划分的文档,则将 $count_k$ 加 1,并按公式(7)改变 doc_k 的权值,同时将公式(5)和公式(6)中 f_t 在 doc_k 中的权值 $W(f_t, doc_k)$ 增大 $AW(doc_k)$ 倍,这样能在一定程度上消除文档分布不均匀对算法性能的影响.

$$AW(doc_k)(t+1) = AW(doc_k)(t) \times (1 + \log_2(1 + \frac{count_k}{10})) \quad (7)$$

本文以 MQE 作为算法收敛的判别条件,如果 MQE 小于阈值则停止算法运行(实验中设定阈值为 0.01). 如文献[2]所述,MQE 能够衡量聚类结果的平均凝聚程度.

$$MQE = \frac{\sum_{l=1}^{C} \sum_{D_k \in C_l} \frac{\| D_k - V_l \|^2}{| C_l |}}{C} \quad (8)$$

公式(8)中 C 代表聚类后的类别数,C_l 代表映射到第 l 个神经元的文档类,V_l 代表第 l 个神经元的特征向量,D_k 代表第 k 篇文档的特征向量.

4　实验及分析

本文从 1998 年新闻语料中随机选择十万篇新闻作为测试语料,该语料涵盖了经济、政治、文化、体育等多个类别,是平衡语料. 表 1 列出了测试文档集的部分特征链,并对较长的特征链进行了截取.

表 1　测试文档集中特征链的部分结果

序号	特征链
1	冠军 赛事 比赛 训练 全能 决赛 竞技 预赛 总局 奥委会 运动员 裁判 得分 执教 教练 抽签 俱乐部 国家队 晋级 火炬 开幕 ……
2	市场经济 供应 价格 需求 供求 通货膨胀 销售 收入 商品 零售 资金 投资 货币 基金 消费 理财 证券 外资 盈利 营销 行情 ……
3	电脑 配件 内存 硬件 光盘 笔记本 主板 软件 检索 操作系统 升级 容量 显示器 驱动器 芯片 网络 因特网 指令 视频 接口 光纤 ……
4	陆地 口岸 区域 地形 城市 国内 首都 乡村 城镇 海岸 基地 绿洲 县城 地段 空间 海域 内陆 海外 沙漠 岛屿 水域 港口 ……
5	教师 员工 学校 学生 课堂 报告 成绩 中专 函授 辍学 学业 学历 学费 高考 学龄 毕业 留级 报考 教研 课程 辅导 证书 导师 ……

由表 1 可见:通过计算特征同现词向量间的交叉熵可以很好的获得不同特征的上下文语义相似度,使每条特征链均凝聚了相似特征,并且不同的特征链反映了不同的类别信息.

由于本文采用的测试语料集包含的文档数过多,因此无法手工标注文档应归宿的类别,也就无法计算

准确率、召回率等值. 这样即使用其它方法计算算法的准确率,设算法得到的聚类结果中共有 z 个类别,通常 $z \geq 1000$. 分别将每个文档类划分为多个子类别,使各子类别内的文档均反映相同的信息,并以具有最多文档数的子类别作为文档类所描述信息的代表. 设文档类 C_i 中具有最多文档数的子类别为 $SubC_{i\max}$,将其它子类别包含的文档视为 C_i 的不相关文档,则 C_i 的准确率为 $|SubC_{i\max}|/|C_i|$,以所有文档类的准确率的平均值作为算法的准确率.

图 2 列出了是否调整相似特征后算法 SHSOM 的聚类时间,图 3 列出了是否调整相似特征后算法 SHSOM 的聚类准确率.

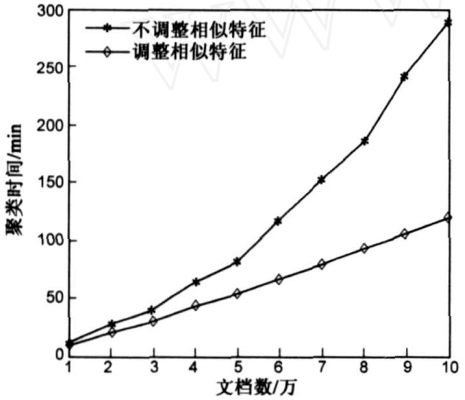


图2 是否调整相似特征的聚类时间

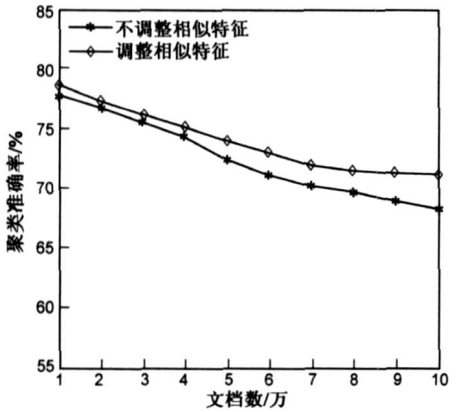


图3 是否调整相似特征的聚类准确率

由图 2 可见:由于对神经元中相似特征进行了调整,加大了权值调整的范围,这样只需经过少量迭代训练即可使神经元获得大幅度的调整,因此算法的收敛时间较短. 由图 3 可见:由于考虑了相似特征的影响,算法可以将含有相似特征的具有语义相似性的文档也凝聚到同一文档类中,提高了聚类的准确率.

本文采用层次聚类、K-means[9]、GSOM[1]、GHSOM[2]与 SHSOM 进行对比,将各算法对于不同文档数的聚类准确率及聚类时间列于图 4 和图 5 中.

由图 4 可见:经典聚类算法,如层次聚类、K-means、GSOM、GHSOM 随文档数增多算法的性能逐渐下降,其中层次聚类、K-means 的性能下降的最快,这是因为随文档数增多,文档的特征集合仅占特征空间的很小一部分,这样大部分文档的相似度趋近于 0. 而 K-means 和层次聚类是在原始特征空间中进行聚类,这样大量相似度趋近于 0 的文档会严重降低算法对文档的划分能力. 相比于上述算法,GSOM 和 GHSOM 较适用于高维数据聚类,这是因为其将高维空间投影到二维平面上,但是随文档数增多,大量文档具有语义相似性且文档的分布也极不平衡,而 GSOM 和 GHSOM 不能对上述问题进行处理从而降低了算法的准确性. 分析图 4 可得:算法 SHSOM 在文档数较少时即拥有较高的准确率,并且随文档数增多算法的性能并不失真,即 SHSOM 对高维数据和低维数据均有较高的准确率. 这是由于算法 SHSOM 在相似度计算和特征权值调整时考虑了相似特征的影响,能够凝聚具有语义相似性的文档,同时通过文档权值调整技术改善了文档分布不平衡对算法性能的影响,因此算法在高维空间和低维空间的性能均很好.

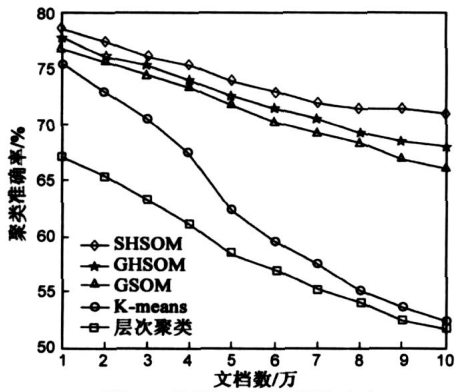


图4 不同算法的聚类准确率

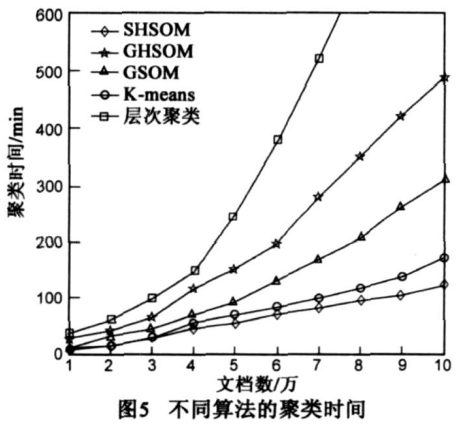


图5 不同算法的聚类时间

由图 5 可见:K-means 和 GSOM 的时间复杂度相差不多,这是因为上述算法均为线性时间复杂度,其中 K-means 的时间复杂度为 $O(klN)$,k 为类别数,l 为循环数,N 为文档数,而 GSOM 以自组织映射算法为基础,其时间复杂度为 $O(k'mN)$,k' 为神经元数,m 为循环数,N 为文档数. 通常 GSOM 中的 k'、m 均比 K-means 中的 k、

l 大,因此时间复杂度稍高[10]. GHSOM 在自组织映射算法的基础上又进行了层次划分,因此时间复杂度较 GSOM 稍高. 层次聚类算法的时间复杂度最高为 $O(n^2)$. 算法 SHSOM 同样基于自组织映射的思想,但是由于算法首先通过特征链技术获得比较合理的初始类别划分,同时在相似度计算和权值调整时考虑了相似特征的影响,加大了权值调整的范围,从而大大减少了算法的迭代次数,因此图 5 中 SHSOM 在处理低维数据时即拥有较好的时间性能,且随文档数增多,SHSOM 的聚类时间的升高幅度比较平缓,即 SHSOM 的时间复杂度受文档数的影响并不严重.

5　结论

本文提出一种基于语义的高维数据聚类算法,算法首先通过计算同现词向量间的交叉熵获得不同特征的语义相似度,然后将相似的特征凝聚为特征链并依此将文档集合划分为多个初始类别. 将按上述方法得到的特征链对应于自组织算法中的神经元,并通过调整神经元中特征的权值逐步提高类别划分的准确性. 针对高维特征空间中存在大量具有语义相似性的文档,算法在相似度计算及权值调整时考虑了相似特征的影响,使其能够凝聚具有语义相似性的文档. 针对高维特征空间中文档分布不均匀的问题,算法根据文档被错误划分的次数,对不同文档赋予不同的权值,该方法可以有效避免聚类结果的"过适应"和"欠适应"现象. 分析实验结果可以得出:算法在高维特征空间中拥有较高的聚类准确率和较短的运行时间.

参考文献:

[1] Damminda A, Saman K H. Dynamic self-organizing maps with controlled growth for knowledge discovery [J]. IEEE Transactions on Neural Networks, 2000, 11(3): 601 - 614.

[2] Rauber A, Merkl D. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data [J]. IEEE Transactions on Neural Networks, 2002, 13(6): 1331 - 1341.

[3] 吴郢,阎平凡. 结构自适应自组织神经网络的研究[J]. 电子学报, 1999, 27(7): 55 - 58.
Wu Y, Yan P F. A study on structural adapting self-organizing neural network [J]. Acta Electronica Sinica, 1999, 27(7): 55 - 58. (in Chinese)

[4] Xu Y D, Xu Z M, et al. Using multiple features and statistical model to calculate text units similarity [A]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics [C]. China: IEEE Press, 2005. 3834 - 3839.

[5] 陈清才,王晓龙. 一种基于词矢量的汉语语义量化模型[J]. 计算机研究与发展, 2001, 38(2): 207 - 212.
Chen Q C, Wang X L. A word-vector-based quantization model of chinese word sense [J]. Journal of Computer Research and Development, 2001, 38(2): 207 - 212. (in Chinese)

[6] 孙茂松,左正平,等. 基于 K-近似的汉语词类自动判定[J]. 计算机学报, 2000, 23(2): 166 - 170.
Sun M S, Zuo Z P, et al. Part-of-speech identification for unknown chinese words based on k-nearest-neighbors strategy [J]. Chinese Journal of Computers, 2000, 23(2): 166 - 170. (in Chinese)

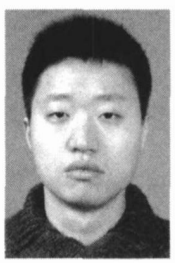
[7] Gonenc E, Ilyas C. Using lexical chains for keyword extraction [J]. Information Processing and Management, 2007, 43(6): 1705 - 1714.

[8] Kohonen T, Kaski S, et al. Self organization of a massive document collection [J]. IEEE Transactions on Neural Networks, 2000, 11(3): 574 - 585.

[9] 刘远超,王晓龙,等. 一种改进的 K-means 文档聚类初值选择算法[J]. 高技术通讯, 2006, 16(1): 11 - 15.
Liu Y C, Wang X L, et al. An adapted algorithm of choosing initial values for k-means document clustering [J]. High Technology Letters, 2006, 16(1): 11 - 15. (in Chinese)

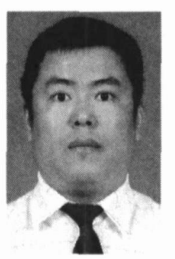
[10] Shahpurkar S S, Sundareshan M K. Comparison of self-organizing map with k-means hierarchical clustering for bioinformatics applications [A]. International Joint Conference on Neural Networks [C]. Hungary: IEEE Press, 2004. 1221 - 1226.

作者简介:



刘　铭　男,1981 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院博士研究生. 研究方向为聚类分析、文本挖掘.
E-mail: mliu@insun.hit.edu.cn

王晓龙　男,1955 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院教授,博士生导师. 研究方向为信息检索、文本挖掘.



刘远超　男,1971 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院副教授,硕士生导师. 研究方向为自然语言处理、人工智能.