基于改进谱聚类方法的搜索引擎排序算法

白 亮[1] 于天元[1] 刘 湜[2] 老松杨[1] 杨 征[1]
(国防科学技术大学信息系统与管理学院 长沙 410073)[1] (61599部队 北京 100011)[2]

摘 要 搜索引擎的性能优劣主要由排序结果决定。针对网页文本特性改进了谱聚类方法,提出了一种融合网页内容和链接质量的排序算法。利用改进的谱聚类方法对网页内容进行分类,并与评价链接质量的PageRank值进行加权融合,计算得到排序结果。实验结果表明,相对于传统的PageRank,HITS,TF-IDF等排序算法,所提算法返回的排序结果具有更高的相关性。

关键词 搜索引擎,排序算法,谱聚类,PageRank
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.042

Ranking Algorithm of Search Engine Using Improved Spectral Clustering

BAI Liang[1] YU Tian-yuan[1] LIU Shi[2] LAO Song-yang[1] YANG Zheng[1]
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)[1]
(Unit 61599, Beijing 100011, China)[2]

Abstract The performance of a search engine is determined by its ranking algorithm. A novel ranking algorithm was proposed which combines the webpage content and its hyperlinks. Spectral clustering is used for analyzing the webpage content and the PageRank value is used for scoring the quality of hyperlinks. Then, final ranking results are generated based on the content relevant value and hyperlink relevant value. Experimental results show that the proposed ranking algorithm is better than traditional ranking algorithms such as TF-IDF, PageRank and HITS.

Keywords Search engine, Ranking algorithm, Spectral clustering, PageRank

1 引言

随着互联网的迅猛发展,网络信息呈爆炸式增长,而搜索引擎能够根据用户查询条件返回具有相关性的查询结果,但大部分用户只浏览前30条,甚至前10条结果[1],如果搜索引擎中排在前面的结果与用户查询无关,用户满意度便会下降。因此,搜索引擎排序算法已经成为信息技术领域的研究热点。

传统的排序算法按照时间顺序可分为如下4类[2,3]:1)基于分类目录的排序算法,以人工分类网页为基础对初期互联网进行搜索排序;2)基于文本检索的排序算法,此类排序算法主要使用了发展已久的信息检索模型,其中TF-IDF算法和BM25算法沿用至今;3)基于链接整合分析的排序算法,其代表就是应用广泛的PageRank和HITS算法,目前的大部分搜索引擎仍然以这两种排序算法为基础;4)以用户为中心的排序算法,目的是满足不同用户的查询偏好,例如用搜索引擎查询同一关键词时,根据用户意图和偏好返回不同的结果,这也是目前搜索引擎技术发展的热门方向。

在上述的搜索引擎排序算法中,基于文本检索的排序算法只关注网页内容,会导致搜索效率低下等问题;而基于链接整合的排序算法只关注网页链接,会导致返回的内容与用户查询不相关等问题。而以用户为中心的排序算法在链接方法的基础上融入了用户信息,针对不同的用户返回不同的排序结果。在实际的应用中,如果在基础排序算法中能够返回与用户查询相关度高的结果,然后再加入用户信息,便能够有效增加查询结果的相关性,提升搜索引擎的质量。

因此,本文综合考虑搜索效率与结果相关性问题,将文本内容和文本链接有效结合,提出了一种基于网页内容和链接质量的排序算法。针对获取网页内容先验知识的困难,利用改进的谱聚类方法对网页内容进行分析,基于网页链接结构对初始查询结果集进行拓展,并计算拓展结果集与用户查询的距离作为网页内容相似度,然后结合衡量网页链接质量的PageRank值,最终得到每个网页的相似度得分并根据该得分返回排序结果。

2 改进的谱聚类方法

文献[4]的研究表明,聚类能更准确地定位搜索结果,在搜索引擎中,应用较多的聚类方法主要包括后缀树聚类、k均值聚类和层次聚类等。相对于这些聚类方法,谱聚类在高维

到稿日期:2015-09-05 返修日期:2016-03-14 本文受国家自然科学基金资助项目(61201339,61571453),湖南自然科学基金资助项目(14JJ3010)资助。

白 亮(1978-),男,博士,副教授,主要研究方向为多媒体信息处理与应用,E-mail:xabpz@163.com;于天元(1992-),男,硕士生,主要研究方向为多媒体信息处理;刘 湜(1981-),女,博士,助理研究员,主要研究方向为信息传播与信息检索;老松杨(1968-),男,博士,教授,主要研究方向为多媒体信息处理与应用,E-mail:songyanglao@sina.com;杨 征(1978-),男,博士,副教授,主要研究方向为多媒体信息系统与虚拟现实,E-mail:yz_nudt@hotmail.com。

数据(例如文本、图像、视频)的处理方面有着较多的优势[5]。近年来谱聚类在数据分析、语音识别、视频分类、图像处理、文字识别等领域得到了成功应用[6]。

针对文本聚类问题,谱聚类方法将聚类问题转化成图的最优划分问题,文本数据由图的顶点 V 表示,两个文本之间的相似度由边 E 的权重表示,则文本数据集可被表示为一个无向加权图 $G=(V,E)$。图 G 的最优化划分即可将图中的点进行分类,使得图中的类内相似度最大,类间相似度最小,此时点的类别划分也就是对应文本的类别划分。但是在处理无先验知识的网页文本时,该方法存在相似度度量难以选取、最佳聚类数目无法自动确定和增量文本无法处理等难题。为解决上述问题,本文提出一种面向网页文本聚类的改进谱聚类方法。

2.1 基于密度的相似度度量

基于谱聚类的文本聚类需满足两个假设[7]:局部一致性,即空间位置上距离较近的数据有较高的相似性;全局一致性,即位于同一流形上的数据有较高的相似性。

谱聚类中经常使用的高斯核函数只能够反映局部一致性而没有考虑全局一致性,不能完全反映分布复杂的数据集。有效表征全局一致性必须考虑文本数据在空间上的密度。

因此,定义基于密度的线段长度如式(1)所示:

$$L(x,y)=\rho^{dist(x,y)}-1 \tag{1}$$

其中,$dist(x,y)$表示两点之间的欧氏距离,ρ 是一个大于 1 的伸缩因子。可以通过调节 ρ 的大小来调整两点之间的基于密度的距离,使得密度较大区域内多点距离之和小于密度较小区域内的两点距离,达到反映全局一致性的目的。令边集合为 $E=\{L(a,b)\}$,$v=\{v_1,v_2,\cdots,v_l\}\in V$ 表示图中长度 $l=|v|$ 的连接数据点 v_1 和 v_2 的路径,其中边 $(v_k,v_{k+1})\in E$,$1\leqslant k\leqslant l-1$,则数据点 x_i 与 x_j 的距离为:

$$D(x_i,x_j)=\min\sum_{k=1}^{l-1}L(v_k,v_{k+1}) \tag{2}$$

该距离度量放大了类间数据间距,缩短了类内数据间距。基于此,定义基于密度的相似性度量:

$$W(x_i,x_j)=\frac{1}{D(x_i,x_j)+1} \tag{3}$$

与高斯核函数相比,式(3)的参数的敏感度较小,并且充分考虑了全局一致性。

2.2 最佳聚类数目的确定

如上文所述,用图来表示文本数据,因此对文本的聚类即可视为对图中点的聚类。假设文本数据集中共有 m 个 n 维文本数据,可以将该数据集用一个 $m\times n$ 的矩阵 W 表示,其中行向量表示一个文本,列向量表示一个文本特征项的权值,用 x_i 表示第 i 个文本数据向量。下面定义几个变量。

数据集中所有文本数据的平均值定义为:

$$\bar{x}=\frac{1}{m}\sum_{i=1}^{m}x_i \tag{4}$$

数据集的某一类中所有文本数据的平均值定义为:

$$\mu_j=\frac{1}{|c_j|}\sum_{x_i\in C_j}x_i \tag{5}$$

其中,$|c_j|$表示类别 c_j 中文本数据的数量。

数据集的总体方差为:

$$S^t=\sum_{i=1}^{m}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}} \tag{6}$$

数据集类内方差为:

$$S_w^t(k)=\sum_{j=1}^{k}\sum_{x_i\in C_j}(x_i-\mu_j)(x_i-\mu_j)^{\mathrm{T}} \tag{7}$$

数据集类间方差为:

$$S_b^t(k)=\sum_{j=1}^{k}|c_j|(\overline{x_i}-\mu_j)(\overline{x_i}-\mu_j)^{\mathrm{T}} \tag{8}$$

我们的目标是使得同一类的文本相似度较小而不同类的文本相似度较大,即数据集的类内方差最小且类间方差最大。而在上述各式中,总方差 S^t 是一个常数,因此目标函数为:

$$\begin{cases}\min S_w^t(k)\\ \max S_b^t(k)\end{cases} \tag{9}$$

事实上,将式(7)和式(8)带入式(9),可得:

$$S_w^t(k)+S_b^t(k)=S^t \tag{10}$$

因此,上述两个目标函数的解是一致的。利用 C-H 指数定义方差比标准[8,9],如式(11)所示,使 $S_{k,m}$ 到达第一个局部最大值的 k 值即为最佳类别数。

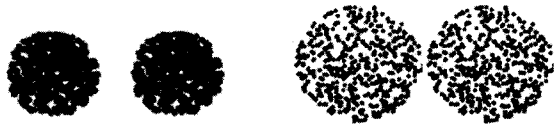
$$S_{k,m}=\frac{(m-k)S_b^t(k)}{(k-1)S_w^t(k)} \tag{11}$$

由上述描述可知,为了找到最佳类别数,要不断迭代运行聚类算法。显然,如果该方法应用在谱聚类算法上,排序算法的效率将会更低,所以本文采用聚类效率较高的 k-means 算法作为寻找最佳类别数的基本算法,以避免采用复杂的优化算法寻找初始聚类中心的问题,降低了计算复杂度,加快了聚类速度。

2.3 基于类别相合性的增量谱聚类方法

由于网页内容信息更新周期很快,使用聚类方法得到的类别特征可能与新的网页文本不匹配,因此需要重新计算抽取类别信息,通常采用重新聚类或者增量聚类方法[10]。但是,每次都进行重新聚类不仅会浪费计算资源,而且会造成信息更新的不及时,导致搜索引擎无法提供最新的搜索信息。

针对增量数据,关键问题是可能有大量数据出现在两类之间使得两类之间有了合并的可能性,但是仅仅依靠类的中心距离来判断两类之间能否合并是不合适的。例如图 1 中(a)和(b)所示的类中心距离是相等的,但是(b)合并的可能性应该要比(a)合并的可能性大,这是因为(b)中两类的分布情况更倾向于类的合并。因此,引入类的相合性来度量两类之间的相似性程度[10]。



(a) (b)

图 1 类中心距离相同而分布不同的两个聚类

对于新文本数据,判断其与各类的连接度,如果连接度大于某一阈值,则可以将该文本分至该类;否则将文本独自分为一类。基于该原则对增量文本进行聚类处理。但是,一旦增量文本被处理后,聚类结果就不能进行调整,即一旦某个文本被错误分类,那么这个错误就会一直延续下去,使得类的信息与真正类信息相差越来越远,大大地降低了聚类的准确性。因此,应该重新分配不确定分类的文本,对聚类结果进行调整和修正。当计算文本与类的连接度时,不仅要选取最大连接度,还要考虑次大连接度,当两者差值较小时判定该文本的分类是不确定的,此时先将本分类但是不更改类信息,以防止由于一个文本分类错误导致整个分类错误。当处理的增量文本

数据达到一定数量后，考虑重新对该类文本进行分类，并进行类间合并。

基于上述分析，定义两种类别特征信息：类中心向量和类的均值，计算公式分别如式(12)和式(13)所示：

$$cen_j=\frac{\sum_{x_i\in C_j}x_i}{|C_j|} \tag{12}$$

$$\overline{C_j}=\frac{\sum_{j=1}^{|C_j|}\sum_{x_i\in C_j}D(x_i,cen_j)}{|C_j|} \tag{13}$$

其中，$D(x_i,cen_j)$表示类内任意文本数据与类中心的距离，计算方式如式(14)所示。

$$D(x_i,C_j)=\min(D(x_i,x_j)),x_j\in C_j \tag{14}$$

其中，两点之间的距离的计算方式已于 2.1 节给出。定义类 C_i 中的文本 x_i 与类 C_j 之间的连接度为：

$$j(x_i,C_j)=e^{-\frac{|D(x_i,c_j)-D(x_i,c_i)|}{r_i}} \tag{15}$$

其中，r_i 表示类 C_i 的半径，数学表达为：

$$r_i=(\frac{\sum_{i=1}^{n}D(x_i,\overline{C_i})}{|C_i|})^{\frac{1}{2}} \tag{16}$$

本文提出的基于相合度的增量谱聚类方法如算法 1 所示。

算法 1 基于相合度的增量谱聚类算法

输入：初始类别信息(包括各类的中心和均值)，新增文本

输出：新增文本加入的类别和更新后的类别信息

第 1 步 计算新文本数据 x_i 与各类的连接度；

第 2 步 如果最大连接度 $\max j(x_i,C_j)>\beta$ 并且最大连接度与第二最大连接度的差 $\max j(x_i,C_j)-\text{secmax } j(x_i,C_k)>\alpha$，则 x_i 将加入到类 C_j 中，并更新类的特征信息；

第 3 步 如果最大连接度 $\max j(x_i,C_j)>\beta$ 并且最大连接度与第二最大连接度的差 $\max j(x_i,C_j)-\text{secmax } j(x_i,C_k)<\alpha$，则暂时将 x_i 加入到类 C_j 中，并给出标记但是不更新类信息；

第 4 步 如果最大连接度 $\max j(x_i,C_j)<\beta$，则将 x_i 分为一个新的类。

当进行了一定数量的文本增量聚类后，对于已经标记的暂时存放的文档进行重新分类。重新计算最佳聚类数 k，如果 k 比当前类别小，则合并最大相合度的类；如果 k 比当前类别大，则重新进行聚类。

3 基于改进谱聚类的搜索引擎排序算法

通过改进谱聚类对原始文本数据集进行聚类后，需要根据用户查询得到初始结果集，而用户查询词是有可能跨类存在的，例如“林肯”这个词，它可能指的是汽车的品牌也可能指美国总统，所以在维度上可能会出现两类相交的现象。因此，不能够单纯地从文本间距的角度考虑用户查询类别的划分。本文采用条件概率解决这个问题，设 q 为用户查询向量，q_i 为用户查询向量的分量，则用户查询属于某一类别的概率可由式(17)计算得到：

$$P(C_j|q)=\frac{p(C_j)*P(q|C_j)}{P(q)}\propto p(C_j)*\prod_i P(q_i|C_j) \tag{17}$$

定义 $P=(p_1,p_2,\cdots,p_k)$ 来表示查询 q 与各类相关的概率，可以认为概率越大，查询与该类的相关度越大。按照概率在每个类中都选出相应数量的结果作为内容分析的结果集，并以文本与查询距离的倒数作为该文本在当前查询下的权值。

进而，融合链接质量(即 PageRank 值)来确定最终的排序结果。考虑到现有方法是完全根据内容对网页与查询条件进行相似度匹配，在聚类情况不稳定的情况下，可能会有一些重要的相关网页因侧重点不同而被分到了其他类别，可以通过链接信息构建与这部分信息的关联。具体方法如下：首先利用布尔查询对整个文本数据集进行查询，如果查询到的文本不在现有的初始结果集中，那么将该文本加入结果集并计算与查询向量之间的距离。然后将初始结果集按网络结构拓展一层，计算拓展结果集中文本与查询向量的距离，即内容相关度。再将拓展结果集的内容相关度和 PageRank 值进行归一化，进行加权得到每个文本与查询的相关度。最后按照文本相关度排序，由大到小返回查询结果。

最终的文本得分计算公式如下：

$$Score(x_i,q)=a*CR(x_i)+b*PR(x_i) \tag{18}$$

其中，a 和 b 为设定的有关网页内容和链接权值，满足加和为 1，$Cr(x_i)$ 表示归一化后的网页文本 x_i 的内容相关度，$PR(x_i)$ 表示归一化后的网页文本 x_i 的 *PageRank* 值。

4 实验结果与分析

对排序算法进行评价的最重要的指标就是相关度，对于网页文本来说，一个网页内容是否与用户查询相关只能由用户决定，即使对于同一用户的同样的查询，在不同情况下用户的真正意图也是不同的，对除用户以外的人来说，判断网页与用户查询是否相关是比较困难的。因此，本文采用主观评价的方法对提出的排序算法进行性能分析，实验中邀请了 10 名志愿者对 10 组用户查询进行实验，并依据志愿者主观意愿判断查询效果。对于实验中的一些基本参数做出如下说明：连接度阈值被设定为初始聚类中文本与类的连接度的最小值；在计算网页最终得分时，由于很难判断内容相关度和链接重要性对结果的影响大小，因此将 a 和 b 各设为 0.5。

4.1 评价准则

本文采用的排序算法评价指标包括：

(1) $p@n$

由于用户只关注前 n 个排序结果，因此对前 n 个结果计算查询准确率更有意义，该指标的计算方式如下：

$$p@n=\frac{\text{前 } n \text{ 个结果中与查询相关的结果数}}{n} \tag{19}$$

(2) *MAP*

虽然查准率和 $p@n$ 指标已经能够衡量前 n 个结果的准确率，但是它们还不能衡量结果的位置的影响，首先定义平均精度(*AP*)，对于任意查询 q，

$$AP=\frac{\sum_n P@n*I(n)}{\text{前 } n \text{ 个结果中相关文档数量}} \tag{20}$$

其中，

$$I(n)=\begin{cases}1, & \text{第 } n \text{ 个文档与查询相关}\\ 0, & \text{第 } n \text{ 个文档与查询不相关}\end{cases} \tag{21}$$

MAP 则是所有查询的平均精度的均值。

(3) *NDCG*

NDCG 对传统的评价标准做出了改进，主要体现在两方面：1)相关程度应该有区别，完全相关的文档的得分应该更高；2)文档的排序位置越靠后，文档的得分应该越小。

在这个评价指标里定义了几个变量，首先将排序结果评

级，评级越高则说明文档越重要。当评级为 i 时，变量 $Gain$ 的计算方式如下：

$$Gain=2^i-1 \tag{22}$$

然后定义变量 CG：

$$CG[0]=Gain[0] \tag{23}$$

$$CG[j]=CG[j-1]+Gain[j] \tag{24}$$

其中，j 为文档的序列号，考虑到排序的位置，定义变量 DCG 如下：

$$DCG[0]=Gain[0] \tag{25}$$

$$DCG[j]=DCG[j-1]+\frac{Gain[j]}{\log_2(j+1)} \tag{26}$$

若该排序不是最优排序，则计算出 max DCG，故 $NDCG$ 为：

$$NDCG[j]=\frac{DCG[j]}{\max DCG[j]} \tag{27}$$

可以看出，$NDCG$ 为一个向量，为了便于比较算法之间的差异，用 $NDCG$ 的平均值来表示指标的最终结果。

(4)效率

搜索引擎的响应时间可以作为评价搜索引擎的指标之一，虽然搜索引擎的整体响应时间可能会受到网络状况、主机性能、用户设备的性能等因素的影响，但是用户总是希望快速地返回查询结果，因此也将排序算法的效率作为评价准则之一。

4.2 实验结果与分析

为使实验具有一般性，本实验将 Alexa 网站评价的互联网中前 100 强网址作为爬虫的初始集合，最终爬取了共 120 万个原始网页文本数据。基于该数据集选定了 10 组关键词，其中包括表示时间的“2014 年”，近期的热点词汇“奶茶”和“世界杯”，人名“丁俊晖”，一词多义的“奶茶”和“老虎”，英文简写“cctv”和“suv”，缩写“男篮”，网站名称“新浪”和搜索中用户经常查询的词汇“热门”。

通常用户不希望逐个浏览查询结果，其最希望的是：最重要的信息排在最前面！因此，实验中只对各类算法的前 30 个结果进行比较分析。

对于上述评价指标，由于实验环境的影响，用归一化后的响应时间表示效率指标。由 100 名志愿者对上述关键词进行查询，对查询结果与自身期望结果进行相关度判断，进而计算相关度均值，实验结果如表 1 所列。

表 1 各个算法的评价指标

排序算法	p@10	p@30	MAP	NDCG	归一化后的响应时间
TF-IDF	0.743	0.473	0.529	0.465	**0.503**
PageRank	0.728	0.425	0.496	0.537	0.704
HITS	0.571	0.611	0.522	0.454	1.000
基于谱聚类的排序算法	0.728	0.692	0.651	0.541	0.901
基于改进谱聚类的排序算法	**0.848**	**0.811**	**0.758**	**0.542**	0.740

从实验结果可以发现，除归一化的响应时间以外，较其他经典算法，本文提出的算法取得了更好的效果。从前 10 个结果的准确性来说，HITS 由于根据网络链接结构做了初始集合的拓展，结果的准确性明显要比其他几种算法差一些，但是也正是由于该特点，HITS 算法在前 30 个结果的准确性表现良好。而本文算法在考虑链接的基础上做了内容上的分析，可以发现结果的准确性和相关度有了双重保证，这一点在 MAP 指标上也有所体现。$NDCG$ 指标体现了算法的排序结果是否合理，与用户的真正需求有直接的关系，此处取平均值表现了每种算法的平均排序合理程度，可以看出几种算法的差距其实并不大。从算法效率上看，TF-IDF 算法的效率最高，而本文的谱聚类算法的响应时间在 PageRank 算法与 HITS 算法之间，因此虽然本文算法离线计算复杂度较高，但是在线计算量较小，可以让用户接受。

结束语 本文提出了一种基于谱聚类和 PageRank 相结合的排序算法。针对无先验知识的网页文本聚类问题，从相似度度量选取、最佳聚类数目确定、初始聚类中心确定和增量文本处理 4 个方面对谱聚类方法进行了改进。利用改进的谱聚类方法对网页进行聚类以达到对网页内容进行归类的目的，在计算最终的排序结果时，利用条件概率确定与用户查询相关的文本类别，并考虑查询与各文本类中文本的距离，保证了与查询相关概率大的类返回更多结果，而又不会使相关概率小的类不返回结果，即保证了结果的全面性，最终考虑 PageRank 值，并借鉴 HITS 算法的拓展根集的做法，在保证准确率的同时进一步提高了查询结果的全面性。通过与 TF-IDF，PageRank 和 HITS 3 种经典排序算法的对比实验验证了所提方法的有效性。

在下一步的工作中，将重点解决以下几个方面的问题：

(1)在实验中，效率问题是该算法应用于实际的主要困难之一。在文本矩阵维度较大的情况下，谱聚类在实现过程中的时间开销很大，再加上考虑密度的文本相似度计算，确定最佳聚类数目，该算法在离线计算部分的时间复杂度较高。

(2)本文构建的搜索引擎的网络爬虫搜索大规模数据集的效率问题也是下一步工作的改进方向之一。本实验中，网络爬虫的爬取效率比较低，由于不是全网搜索，应该考虑搜索时不只用简单的广度优先策略，而应该考虑有偏好的搜索方法。

(3)在该算法的基础上融入用户信息，如此便能够返回令用户更满意的结果，真正构建出基于聚类方法的第四类搜索引擎。然而，用户信息和数据是我们难以获取的，因此实现起来也有很大的难度

参 考 文 献

[1] Li Xiao-ming, Yan Hong-fei, Wang Ji-ming. Search engine-principle, technology and system[M]. Beijing: Science Press, 2009: 169-182, 14-15(in Chinese)
李晓明，闫红飞，王继明. 搜索引擎—原理、技术与系统[M]. 北京：科学出版社，2009：169-182，14-15

[2] Chen Kai. On the study of ranking algorithms in relation to search engine[D]. Wuhan: Wuhan University of Technology, 2011(in Chinese)
陈凯. 搜索引擎有关排序算法研究[D]. 武汉：武汉理工大学，2011

[3] Dong Shu-ling. Research and improvement of ranking algorithms on search engine[D]. Liaoning: Liaoning Technical University, 2012(in Chinese)
董书玲. 搜索引擎排序算法的研究与改进[D]. 辽宁：辽宁工程技术大学，2012

[4] Wang Jia-le. Textual clustering study in search engine[J]. Busi-

ness Economy,2014(3):101-102(in Chinese)
王佳乐.搜索引擎的文本聚类研究[J].商业经济,2014(3):101-102
[5] Ulrike L. A tutorial on spectral clustering [J]. Statistics and Computing,2007,17(4):395-416
[6] Jia H,Ding S,Xu X,et al. The latest research progress on spectral clustering[J]. Neural Computing and Applications,2014,24(7/8):1477-1486
[7] Ding C,He X,Zha H,et al. Spectral min-max cut for graph partitioning and data clustering[C]// Proc. of the IEEE Intl. Conf. on Data. Mining. 2001:107-114
[8] David G,Averbuch A. Spectral CAT:categorical spectral clustering of numerical and nominal data[J]. Pattern Recognition,2012:416-433
[9] Calinski T,Harabasz J. A dendrite method for cluster analysis [J]. Communications in Statistics,1974,3:1-27
[10] Lin C R,Chen M S. A robust and efficient clustering algorithm based on cohesion self-merging[C]// Proc of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2002:582-587

(上接第 210 页)

唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究[J].中文信息学报,2007,21(6):88-94
[3] Pang B,Lee L,Vaithyanathan S. Thumbs up?:sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics,2002:79-86
[4] Liu S M,Chen Jun-huan. A multi-label classification based approach for sentiment classification[J]. Expert Systems with Applications,2015,42(3):1083-1093
[5] Li Dong,Wei Fu-ru,Liu Shu-jie,et al. A statistical parsing framework for sentiment classification[J]. Computational Linguistics,2015,14(2):293-336
[6] Zhang Dong-wen,Xu Hua,Su Zeng-cai,et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications,2015,42(4):1857-1863
[7] Chawla N V,Japkowicz N,Kotcz A. Editorial:Special issue on learning from imbalanced data sets[J]. SIGKDD Explorations Newsletters,2004,6(1):1-6
[8] Wang Su-ge,Li De-yu,Zhao Li-dong,et al. Sample cutting method for imbalanced text sentiment classification based on BRC [J]. Knowledge-Based Systems,2013,37:451-461
[9] Su Jin-shu,Zhang Bo-feng,Xu Xin. Advances in machine learning based text categorization[J]. Journal of Software,2006,17(9):1848-1859(in Chinese)
苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848-1859
[10] Japkowicz N,Stephen S. The Class Imbalance Problem:A Systematic Study[J]. Intelligent Data Analysis,2002,6(5):429-449
[11] Chandrashekar G, Sahin F. A survey on feature selection methods [J]. Computers & Electrical Engineering,2014,40(1):16-28
[12] Kubat M, Matwin S. Addressing the curse of imbalanced training sets:one-sided selection[C]// Proceedings of the 14th International Conference on Machine Learning. 1997:179-186
[13] Wang B X,Japkowicz N. Imbalanced data set learning with synthetic samples[C]// Proc. IRIS Machine Learning Workshop. 2004:19
[14] Zhu Ming,Tao Xin-min. The SVM classifier for unbalanced data based on combination of RU-Undersample and SMOTE [J]. Information Technology,2012,1:39-43
[15] Yan Jun, Liu Ning, Zhang Ben-yun, et al. OCFS: optimal orthogonal centroid feature selection for text categorization[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2005:122-129
[16] Wang Su-ge,Li De-yu,Song Xiao-lei,et al. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification[J]. Expert Systems with Applications,2011,38(7):8696-8702
[17] Dai Liu-ling,Huang He-yan,Chen Zhao-xiong. A comparative study on feature selection in Chinese text categorization [J]. Journal of Chinese Information Processing,2004,18(1):26-32 (in Chinese)
代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004,18(1):26-32
[18] Mladenic D,Grobelnik M. Feature selection for unbalanced class distribution and naive bayes[C]// ICML. 1999:258-267
[19] Wasikowski M,Chen Xue-wen. Combating the small sample class imbalance problem using feature selection[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1388-1400
[20] Yin Liu-zhi,Ge Yong,Xiao Ke-li,et al. Feature selection for high-dimensional imbalanced data[J]. Neurocomputing,2013,105:3-11
[21] Ren Yong-gong,Yang Rong-jie,Yin Ming-fei,et al. Information-gain-based text feature selection method[J]. Computer Science,2012,39(11):127-130(in Chinese)
任永功,杨荣杰,尹明飞,等.基于信息增益的文本特征选择方法[J].计算机科学,2012,39(11):127-130
[22] Ogura H,Amano H,Kondo M. Comparison of metrics for feature selection in imbalanced text classification[J]. Expert Systems with Applications,2011,38(5):4978-4989
[23] Zheng Zhao-hui,Wu Xiao-yun,Srihari R. Feature selection for text categorization on imbalanced data[J]. ACM SIGKDD Explorations Newsletter,2004,6(1):80-89
[24] Fan R E,Chen P H,Lin C J. Working set selection using second order information for training support vector machines[J]. The Journal of Machine Learning Research,2005,6:1889-1918
[25] He Hai-bo,Garcia E. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Engineering,2009,21(9):1263-1284