文章编号:1007-6735(2010)05-0507-04

基于密度优化的KNN算法的研究

陈东晓， 陈庆奎
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：通过对样本网页文本的影响因子特征提取，构建向量空间模型，同时利用OPTICS算法密度无关性，改进了KNN算法.实验表明，该算法增强了结果的稳定性，并产生质量较高的聚类结果.

关键词：KNN算法；Web特征；奇异值分解；OPTICS算法
中图分类号：TP 301.6 **文献标志码**：A

Density optimization based KNN algorithm

CHEN Dong-xiao， CHEN Qing-qui
(*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: A vector space model by abstracting document features from Web pages of samples by using influencial factors was established. The advantage of density independence of OPTICS algorithm was taken for, improving the KNN algorithm to overcome the shortcoming of inadaptability to clustering of low-density point. Experiments show that the algorithm can enhance the stability of the results and produce higher-quality clustering results.

Key words: *KNN algorithm*; *Web features*; *singular value decomposition*; *OPTICS algorithm*

Web主题聚类[1]是目前文本数据挖掘研究的热点之一.将物理或抽象对象的集合分成相似的对象类的过程称为聚类.聚类的目的是使同一个组内的数据对象具有较高的相似度.Web页面是Internet发布信息最常见的代理，用户为了得到自己需要的资源和知识，也许要花费几个小时、甚至需要更长的时间从庞大的知识库中搜索结果 .如何快速准确地从巨大的信息资源中找到所需信息成为困扰网络用户的一大难题.

在Web主题提取技术领域，国内外已进行了大量的研究.文献[2]提出了将网页划分为内容块的思想，以页面中Table，Div等作为处理元素将页面分割为块.对于同一Web模板生成的对象，按照差异性部分提取各个对象的特征.文献[3]阐述了一种模板与机器识别相结合的Web信息自动提取方法，此方法采用一组启发式规则自动识别HTML文本中不同属性信息之间的分隔符，再将它们配置到模板中，然后根据模板分析相同类型的网页.

1 基于Web样式完成特征选择

特征提取是进行文本数据源提取的过程，它通过Web页面区域和文本样式等手段结合的方式来达到特征选择的目的.目前国内对于特征提取的研究主要集中在文档结构解析和文档内容解析两个方面.在实际运用中，则是通过对固定模板Web的配置和特定结构网站解析达到的.在用类似思想做成的特征提取系统中，如果要解析不同电子商务网站

收稿日期：2009-10-26
作者简介：陈东晓(1984-)，男，硕士研究生. E-mail：cdxkfe@163.com

的Web主题特征,就必须定义对应规则的模板和解析规则.

本文在研究相关资料的基础上,提出了内容解析和结构解析结合的方法,即在解析文本结构的同事,结合文本结构的特性,对于解析的主题关键词进行权值解析.

Web文档实际是一种纯文本文档,其中包括了大量的HTML标签和文档内容.文档内容体现了Web内容的主题倾向,在当今信息众多的Web页面中,文档内容的重要性实际上和Web文档结构有着重要的关系.依据W3C的定义[4],DOM(document object model)模型的属性包括标记名(nodeName)、节点类型(nodeType)、节点内容(data)、父节点对象模型(parentNode)、子节点对象模型(firtstChild…lastChild)及兄弟节点对象模型(priviousSibling…nextSibing)等.依照上述的定义,可以清晰地将Web文档解析成DOM树,而在DOM树的叶子节点上体现出Web文档内容.可以结合HTML标签和文档内容给出DOM树节点影响因子.

定义1 (影响因子)DOM模型中HTML标签属性和内容的相互作用关系,用Influence(node)表示,Influence(node)$\in[0,1]$.

计算节点影响因子 I 要综合考虑DOM模型的属性的功能,在这里初步提取了几种标签值,对于标签值和权值进行划分.可构造影响度因子初值向量.同时结点影响度因子具有传递性,即某结点的影响度因子值应向其子结点传递模型中的标签及影响因子,如表1所示.

表1 标签及其影响因子

Tab.1 Label and influence

HTML标签	I
Title	0.90
META	0.85
H1,H2,H3,…	0.80
B,EM,STONG,I	0.75
其他标签	0.70

2 空间向量模型及特征提取

与传统的数据库的简单数据相比较,文档的高维数据是人类所使用的自然语言,计算机难以处理其语义.向量空间模型VSM (vector space model)[5]是近年来应用最多且效果较好的文档表示方法.

特征词文档矩阵:根据空间向量模型的概念,构建空间向量模型 $V=DS$.其中,D 内的向量模型由文档 $d_{i(}(i=1,2,3,\cdots,n)$ 构成,那么包含 d 个文档和 t 个特征词的可观测文档集 V 可以表示成 $t\times d$ 阶的特征词-文档矩阵.

$$V = DS \tag{1}$$

建立可观测的文档集 V 后,由于任意一个文档都是由有限词汇组成,而不是 t 个词汇的全集,所以,潜在语义分析(LSA)是将每个文档中特征词的出现视为非随机现象,而LSA的关键算法是利用奇异值分解(SVD)的方法实现矩阵维数的降低,从而在向量空间模型上达到简化稀疏矩阵的目的.

任何一个矩阵 $\boldsymbol{X}_{m\times n}$,$\boldsymbol{X}$ 的秩记为 r,均可分解为两个正交矩阵和一个对角矩阵的乘积

$$\boldsymbol{X} = \boldsymbol{TSD}^{\mathrm{T}} \tag{2}$$

$\boldsymbol{T}_{m\times r}=(t_1,t_2,\cdots,t_r)$ 为正交矩阵,其中,t_1,$t_2,\cdots,t_r$ 为 $\boldsymbol{X}$ 的左奇异向量,并且是 $\boldsymbol{XX}^{\mathrm{T}}$ 的特征向量,$\boldsymbol{S}_{r\times r}=\mathrm{diag}(\partial_1,\partial_2,\cdots,\partial_r)$ 是对角矩阵,∂_1,$\partial_2,\cdots,\partial_r$ 是 $\boldsymbol{X}$ 的所以奇异值,同时也是 $\boldsymbol{XX}^{\mathrm{T}}$ 所有特征值的平方根,并且满足关系 $\partial_1>\partial_2>\partial_r>0$;$\boldsymbol{D}_{n\times r}=(d_1,d_2,\cdots,d_r)$ 为正交矩阵,其中,$d_1,d_2,\cdots,d_r$ 为 $\boldsymbol{X}$ 的右奇异向量,并且是 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征向量,所以,矩阵 $\boldsymbol{X}$ 可以表示为

$$\boldsymbol{X} = \partial_1 t_1 d_1^{\mathrm{T}} + \partial_2 t_2 d_2^{\mathrm{T}} + \cdots + \partial_r t_r d_r^{\mathrm{T}} \tag{3}$$

首先计算各个文档的特征权重以及集合的相似度矩阵.特征权重的计算采用经典的 TF×IDF 算法[6],进行格式化处理

$$w_t = \frac{TF\log_2(N/DF_t+0.01)}{\sqrt{\sum_{t=1}^{m}[TF\log_2(N/DF_t+0.01)]^2}\left[1+\sum_{j\in L}\Big/\sum_{i=1}^{r}a_i\right]} \tag{4}$$

式中,TF 为词 t 在文档 D 中出现的频率;N 为训练文档的总数;DF 为训练文档中出现词 t 的文档数;r 为特殊标记的数目.

文档之间的相似度采用余弦法则,即计算文档向量之间的内积

$$\cos(D_i,D_j) = \frac{\sum_{t=1}^{m}(w_{i,t}w_{j,t})}{\sqrt{\sum_{t=1}^{m}w_{i,t}\sum_{t=1}^{m}w_{j,t}}} \tag{5}$$

3 原始KNN算法基本流程

K均值算法[7]是平面划分法中最具代表性的算

法.一般 K 均值文本聚类算法以 k 为参数，将 n 个文档对象分为 k 个簇.首先随机地选择 k 个初始文本对象，每个对象代表一个簇的平均值或中心.对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平均值.这个过程不断重复，直到聚类中心趋于稳定为止.判断文本相似或不相似的方法是度量向量之间的角度，两个文档向量之间的角度越小，文档之间的相似度越大，角度越大，相似度越小.角度度量通常用夹角余弦值来表示.

4　KNN 算法改进流程

OPTICS 算法总是主动搜寻对象密集的区域并能够快速抵达并处理区域核心部位的对象.而且识别出的高密度区域不受形状、密度和大小的限制，并能够提供较为理想的类信息.但是，对于相对稀疏的区域，次序信息表现出不合理性，不能识别低密度区域的点的所属类别.可以利用 OPTICS 算法对于输入参数不敏感的特性，先用 OPTICS 算法对于数据进行初步筛选，这样可以克服距离聚类算法发现“类圆型”特征点的弱点.现介绍该算法.

输入参数：特征集 SourceData，核心距离，可达距离.

Step 1　对于特征集中未被处理的数据源点 P.

Step 2　将点 P 的处理标志位重写成 FALSE，并将点 P 加入 ControlList.

Step 3　循环处理 ControlList 的第一个非空元素 m，计算 C_dist = core-distance(O)，并将 O 的处理标位置否.

Step 4　当 O 是核心对象的情况下，更新 O 的可达距离对象的内容，并根据对象是否在 ControlList 中更新 ControlList 列表.

Step 5　当 ControlList$\in\phi$，进入 Step 1.

利用 OPTICS 算法，完成搜寻高密度点对象，对于稀疏点对象的聚类分析，采用 K 临近准则将未完成计算的对象进行分组，从而完成文档对象的分类.

$$V(C_i \mid D) = \sum_{K=1}^{K} \cos(D, D_k) P(C_i \mid D_k) \quad (6)$$

其中，cos 余弦向量法函数是特征抽取中的式(5)，$D_1, D_2, \cdots, D_k$ 是 OPTICS 算法执行完后，已经分类的聚类文档与待计算文档相似度最大的 K 个文档.

$$P(C_i \mid D_k) \begin{cases} 1 & D_k \in C_i \\ 0 & D_k \notin C_i \end{cases} \quad (7)$$

5　Web 聚类试验及结果分析

为了检验算法改进后的有效性，本文对特征提取优化和 KNN 算法的改进两个方面进行了试验比较，试验开发平台 WindowsXP SP2，Visual C ++ 6，硬件配置赛扬 2.1 GHz，512 MB 内存.为了测试算法的有效性，作者进行了验证.

5.1　特征优化对比

在基于潜在语义分析环节，主要考虑奇异值分解中的压缩率，压缩率 α 是维数 r 与原矩阵维数 n 的比率.当压缩率越低的时候，体现出特征值抽取对于噪声的排除高.当压缩率较高的时候，特征值数量的保留率较大.

机器学习中常用的评估标准有分类正确率(classification accuracy)、查准率(precision) 与查全率(recall)、查准率与查全率的几何平均数、信息估值(information score)及兴趣度(interestingness)等.

查全率 R = 分类的正确文本数/应有文本数.

查准率 P = 分类的正确文本数/ 实际分类的文本数.

奇异值分解对于特征提取的影响如表 2 所示.A 为原始向量数，B 为压缩后数量，C 为向量出错数.

表 2　奇异值分解对于特征提取的影响

Tab.2　Influence of SVD

α/%	A	B	C	R/%	P/%
100	1 558	1 558	96	97.2	78.3
55	1 438	465	103	90.3	85.2
33	1 375	135	132	81.0	89.4

通过表 2 可以看出，随着压缩率的降低，查全率虽然降低，但是，降低的幅度相对于压缩率降低来说不是十分明显.此外，随着压缩率的提高，查准率反而提高了，这表明奇异值分解后所进行的特征提取，对于聚类算法的初始值选取是有益的.

5.2　算法优化对比

在聚类算法比较方面，对于改进型 KNN 算法和原始 KNN 算法在数据精度方面进行比较.由于查全率和差准率在数据精度方面只能体现各自的特点，对于本试验，必须综合考虑影响.因此，引进一种新的度量标准 F 测试值.

$$F = \frac{2RP}{R + P} \quad (8)$$

为了检测二次聚类算法的质量，本文选取了京东商城的商品模块作为数据源.由于网页来源于现实的网站，网页数据具有一定的复杂性和实时性.通过分词处理，无效词语过滤后，建立词汇信息表.在进行数据页面的收集的时候，本文主要从京东的几个子商品系统进行数据的采集.如表3所示.

表3 文档数据集

Tab.3 Document data set

文档号	文档数量	分类情况	类别数
D1	2 060	手机通讯，百货类，汽车用品，办公耗材	4

为了检测文档聚类的效果，首先要对原始的文档数据集合进行数据预处理，根据居于页面结构文本块的算法，可以首先构造文档结构模板对象来对文档进行结构解析.在进行网页模块划分处理后，计算出TF－IDF值计算，建立VSM模型.表4是经过聚类算法的初步划分后，所得到的网页聚类结构和特征词语集合.

表4 聚类主题结果

Tab.4 Clustering theme result

类 别	文档数(个)	聚类文档数(个))	主题词语示例
手机通讯	1 032	899	移动，手机，电池，频段，网络
百货类	411	360	家用，零售，超市，服装，家电
汽车用品	380	450	车载，配件，内饰，养护，改装
办公耗材	237	351	复印，耗材，文化，打印，办公

二次划分算法是在初步划分的基础上，获得热点集合的簇后，进行二次聚类划分.根据一次聚类后的结果，可以分析出由于初步聚类不能发现类似圆形的数据特征点，有些特征词语在内容上存在重复的情况，从而使得聚类结果的数据出现了一定的偏差.图1为基于二次划分和一次划分的聚类算法的比较.H为准确率，F代表综合测量值.

从实验结果可以看出，应用OPTICS-KNN对特征矩阵进行数据聚类后，查全率和查准率有一定幅度的提高，从而提高了分类的精确度.

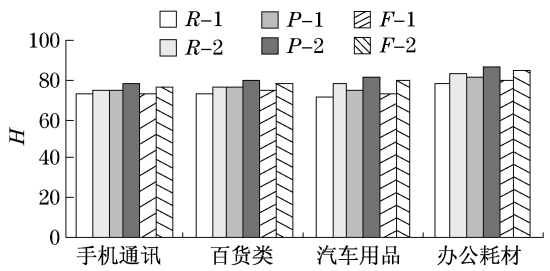

图1 二次划分算法和一次聚类算法比较

Fig.1 Comparision of the second division algorithm and a clustering algorithm

6 结束语

给出了一种改进KNN算法的文本聚类算法.利用KNN算法对于高密度文档对象较好的聚类特性和OPTICS算法的文档对象距离无关性得到较好的聚类效果.此外，本文在特征提取方面，利用潜在语义分析，降低了数据的噪声和高维数据的稀疏行，为聚类算法的执行减小了算法的计算复杂度.下一步，作者将试验研究的主要方向集中在聚类算法结合计算方面，研究如何提高提高聚集K阈值的初始数值和中心点位置.

参考文献：

[1] [加]HAN J，KAMBER M.数据挖掘概念与技术[M].范明，孟小峰，译.北京：机械工业出版社，2001.

[2] 刘坤朋，罗可.改进的模糊C均值聚类算法[J].计算机工程与应用，2009，(21)：105－108.

[3] 韩客松，王永成，陈桂林.无词典高频字串快速提取和统计算法研究[J].中文信息学报，2001，15(2)：22－24.

[4] VASSILIADIS P，SIMITSIS A. Conceptual modeling for ETL processes[C]//Proceedings of the Fifth ACM International Workshop on Data Warehousing and OLAP，2002.

[5] Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis[J]. IEEE Transactions on Neural Networks，1999，10(3)：626－634.

[6] 凌云，刘军，王勋.多层次Web文本分类[J].情报学报，2005，24(6)：684-689.

[7] 黄兰，郭志敏，习万球.利用聚类技术对图书馆读者社群的研究分析[J].计算机工程与设计，2007，28(22)：5 552－5 555.