基于区域划分的kNN文本快速分类算法研究

胡 元[1,2] 石 冰[1]
(山东大学计算机科学与技术学院 济南 250101)[1] (中国人民解放军77675部队 林芝 860000)[2]

摘 要 kNN方法作为一种简单、有效、非参数的分类方法,在文本分类中广泛应用。为提高其分类效率,提出一种基于区域划分的kNN文本快速分类算法。将训练样本集按空间分布情况划分成若干区域,根据测试样本与各区域之间的位置关系快速查找其k个最近邻,从而大大降低kNN算法的计算量。数学推理和实验结果均表明,该算法在确保kNN分类器准确率不变的前提下,显著提高了分类效率。

关键词 文本分类,kNN算法,聚类,k-均值算法
中图法分类号 TP311 **文献标识码** A

Fast kNN Text Classification Algorithm Based on Area Division

HU Yuan[1,2] SHI Bing[1]
(School of Computer Science and Technology, Shandong University, Jinan 250101, China)[1]
(77675 Troop, PLA, Linzhi 860000, China)[2]

Abstract As a simple, effective and non-parametric classification algorithm, kNN method has been widely used in text classification. In order to improve the efficiency of classification, we proposed a fast kNN text classification algorithm based on area division. We divided the training set into several parts based on their area distribution, and then according to the relative positions between test patterns and those parts, easily found out k nearest neighbours of the test patterns in the training set. This will sharply cut down the amount of calculation of kNN algorithm. Mathematical reasoning and the experimental results both show that this algorithm significantly improves the efficiency of classification while keeping the same accuracy rate of kNN classifier algorithm.

Keywords Text classification, K-nearest neighbor algorithm, Clustering, K-means algorithm

1 引言

kNN算法[1]作为一种简单、有效和非参数的分类方法,在文本分类[2,3]中得到广泛使用,并且取得了很好的效果。kNN算法是一种惰性学习方法,它存放所有的训练样本,直到测试样本需要分类时才建立分类,当训练集较大时,会导致计算量很大。

目前,大量的研究集中在提高算法的执行效率上,方法主要有3种:一是建立低维的特征向量空间来降低计算开销,文献[4-7]对该方法进行了讨论;二是引入快速搜索算法或建立高效索引来加快寻找最近邻的速度,文献[8-13]对该方法进行了讨论;三是通过缩减训练样本来减少相似度的计算量,文献[14-20]对该方法进行了讨论。另外,还有基于急切学习法思想[21]的支持向量机分类算法[22]。但是,上述这些方法对kNN分类器的改进仍然存在欠缺,有的算法是以牺牲准确率为代价的,有的算法对效率的提升并不理想。

本文提出一种最近邻快速搜索算法,该方法将训练集按空间分布情况划分成若干区域,通过各区域与测试样本之间的位置关系快速搜索测试样本的k个最近邻,在确保分类准确率不变的情况下,显著提高了kNN分类器的执行效率。

2 基于区域划分的kNN文本快速分类算法

2.1 区域划分方法

如图1所示,根据某种划分方法,可以很容易地将数据空间中的样本划分到不同的区域中。将训练集样本划分到多个区域后,测试样本的k个最近邻一定落在测试样本的最临近区域中,只要将这些包含测试样本的k个最近邻的最临近区域作为新的训练集,就能有效减少kNN分类器的计算量,提高算法的执行效率。

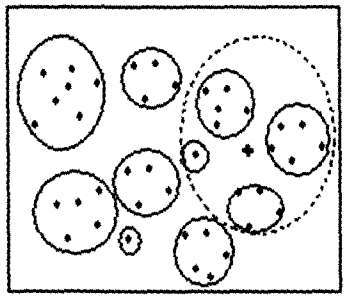


图1 区域划分示意图

到稿日期:2011-12-15 返修日期:2012-03-05
胡 元(1978-),男,硕士生,主要研究方向为数据挖掘、信息检索,E-mail:190820011@qq.com;石 冰(1957-),男,博士,教授,主要研究方向为数据库理论、数据挖掘、信息检索。

k-均值算法[21]作为一种经典划分聚类[21]算法，具有许多优点：(1)简单、快速；(2)对大数据集的处理，该算法是相对可伸缩的和高效率的；(3)聚类得到的结果簇是密集的类圆形区域，而簇与簇之间区别明显。因此，本文选取 k-均值方法作为区域划分的方法。

虽然 k-均值方法得到的区域具有显著的类球形特点，但是，由于受孤立点及数据分布不匀等的影响，我们还是很难对聚类区域边界情况有一个明确的认识。为此，将划分得到的聚类区域的中心向量以及区域中样本到区域中心的最大距离作为区域中心和区域半径，保存为初级分类器。这样，由区域中心和区域半径(即初级分类器)确定的超球体在数据空间上刚好包含了区域中的所有样本点。

2.2 基本概念

根据划分得到的区域特点，给出如下几个定义。

定义 1 给定一个区域 A 和该区域的所有样本 $D=\{d_1, d_2, \cdots, d_n\}$，则这些样本确定的区域中心为：

$$C_A=\frac{\sum_{i=1}^{n} d_i}{n} \tag{1}$$

式中，$d_i=\{w_{i1}, w_{i2}, \cdots, w_{ik}\} \in D$，是样本的空间向量表示。

定义 2 设 $dist(X,Y)$ 表示两个样本 X 和 Y 间的距离，若给定一个区域 A 和该区域内的所有样本 $D=\{d_1, d_2, \cdots, d_n\}$，则区域半径是区域中样本到区域中心距离的最大值，区域半径的计算公式为：

$$r_A=\max(dist(d_1, C_A), \cdots, dist(d_n, C_A)) \tag{2}$$

定义 3 给定一个区域 A 和测试样本 t，则样本 t 到区域 A 的距离是指样本 t 到区域边界的距离：

$$d_A=dist(t, C_A)-r_A \tag{3}$$

式中，当 $d_A>0$ 时，样本 t 在区域外；当 $d_A=0$ 时，样本 t 在区域边界上；当 $d_A<0$ 时，样本 t 在区域内。

2.3 算法描述

基于区域划分的 kNN 文本快速分类算法的思想如下：在训练阶段，用 k-均值方法将训练集样本划分到多个区域中，并将区域中心和区域半径保存为初级分类器；在对每一个测试样本分类前，应用初级分类器选取包含测试样本 k 个最近邻的若干最临近区域组成新的训练集，并利用每个最临近区域更新 k 最近邻集合；分类时，直接利用更新后的 k 个最近邻对测试样本进行类别判定。具体算法如下：

算法 1 基于区域划分的 kNN 文本快速分类算法

输入：训练集 D，测试集 S，近邻数 k，区域数 m；

输出：测试集中每个文本 T 的类别 C；

步骤：

1. 利用 k-均值算法进行区域划分，并将区域中心及区域半径保存为初级分类器 $A\{a_1, a_2, \cdots, a_m\}$；
2. 计算文本 T 到各个区域的距离 $d[a_i]$，并按照距离值的升序将距离值和对应区域的索引存入数组；
3. 取出数组中文本 T 的最临近区域索引 a_i；
4. 如果最近邻集合中样本数目不足 k 时，用区域 a_i 中的样本补充；否则，将区域 a_i 中的样本按照到文本 T 的距离升序排序后，更新文本 T 的最近邻集合；
5. 计算文本 T 与其 k 个最近邻间的最大距离 d_{kNN}；
6. 区域索引 a_i 加入新训练集；
7. 从数组中取出样本 T 到下一个最临近区域的距离 d_A；
8. 如果最近邻集合样本数据不足 k，或者 $d_{kNN}>d_A$ 时，返回 3；
9. 根据最近邻集合中的 k 个最近邻对文本 T 进行类别判定，输出文本 T 的类别 C。

2.4 算法分析

步骤 1 的时间复杂度为 $O(d^2 v)$，其中 d 为训练集样本数目，v 为文本向量维数，对于给定的训练集，其时间复杂度是常量级的；步骤 2 中计算每个测试样本到各区域距离的时间复杂度为 $O(nmv)$，将距离值排序保存的时间复杂度为 $O(nm^2)$，步骤 2 总的时间复杂度为 $O(nmv+nm^2)$，其中 n 为测试样本的数目，m 为划分得到的区域数目；步骤 3 的时间复杂度为 $O(n)$；步骤 4 中使用查找到的新训练集中的最临近区域更新 kNN 集合的时间复杂度约为 $O(nw(d'v+d'^2))=O(nw(dv/m+d^2/m^2))$，其中 w 为查找到的新训练集中最临近区域的数目，d' 为区域中平均样本的数目；步骤 5 中计算测试样本到 k 个最近邻距离的时间复杂度为 $O(nkwv)$，其中 k 为最近邻的数目；步骤 6－9 的时间复杂度为 $O(n)$。算法总的时间复杂度为 $O(nw(dv/m+d^2/m^2)+nwkv+nmv+nm^2)$，通常 $k<<d$，$1\leqslant w\leqslant m\leqslant d$。

在最坏情况下：当区域总数 $m=1$，$w=1$，或 $m=d$，$w=k$ 时，考虑到 $k<<d$，算法的时间复杂度又可表示为：$O(nw(dv/m+d^2/m^2)+nwkv+nmv+nm^2)=O(ndv+nd^2)$。

经典 kNN 分类算法计算测试样本与每一个训练文件的距离并比较其与当前选出的 k 个近邻距离的大小，算法的时间代价约为 $n(k+1)dv$。改进算法在步骤 2 和步骤 4 中，其对距离值排序的时间代价分别为 $nm^2/2$ 和 $nwd'^2/2$，故改进算法的时间代价约为 $ndv+nd^2/2$。

因此，不访设经典 kNN 算法的时间代价为：

$$f_1=n(k+1)dv \tag{4}$$

基于区域划分的 kNN 文本快速分类算法的时间代价为：

$$f_2=n(dv+d^2/2) \tag{5}$$

令

$$f=\frac{f_1}{f_2}=\frac{n(k+1)dv}{n(dv+d^2/2)}=\frac{kv+v}{v+d/2} \tag{6}$$

则，当 $f=\frac{kv+v}{v+d/2}>1$ 时，改进方法分类效率要高于经典 kNN 算法，此时有不等式 $v>\frac{d}{2k}$ 成立。

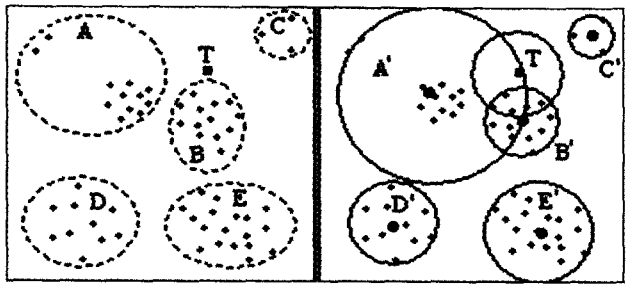
可见，当特征维数较大使得不等式 $v>\frac{d}{2k}$ 成立时，改进算法效率至少能够达到经典 kNN 分类算法效率的 $\frac{k+1}{1+d/(2v)}$ 倍，这一点也通过实验得到了验证。

选择合适的参数 m，可以使新训练集中每个区域至少包含一个最近邻，即 $w\leqslant k$，且不等式 $k<<m<<d$ 成立。算法的时间复杂度为：$O(nw(\frac{dv}{m}+\frac{d^2}{m^2})+nwkv+nmv+nm^2)=O(nmv+nm^2)$，此时，改进算法的时间代价约为 $nmv+nm^2/2$。又 $m<<d$，显然，$nmv+nm^2/2<<ndv+nd^2/2$。因此，在理想情况下，改进算法的时间代价要远小于经典 kNN 算法。

综上所述，基于区域划分的 kNN 分类算法在数据空间维数较大的分类应用中要优于经典 kNN 分类算法，且当 m 取值满足 $k<<m<<d$ 时，改进算法的效率明显高于经典 kNN 算法。可对 m 取不同值，进行多次实验选择最合适的参数取值。

2.5 算法准确性证明

设图 2(a)为某训练集的一次聚类划分，A,B,C,D,E 为划分得到的 5 个区域，训练集样本各不相同，T 为测试样本。图 2(b)为初级分类器描述的区域间位置关系，图中粗圆点代表区域中心，实心圆 A',B',C',D',E' 分别代表以对应区域中心为球心、区域内样本到中心的最大距离为半径的超球体。实心圆 T 表示以测试样本为球心、以测试样本到其 $k(k=5)$ 个最近邻的最远距离为半径的超球体。



(a)聚类划分示意图　　(b)初级分类器示意图

图 2　聚类区域及对应初级分类器

下面以图 2 为例证明改进算法的准确性。

最初，测试样本的最临近区域为区域 A'，将区域 A'(即聚类 A)中的样本作为新训练集，在剩余的区域中，测试样本的当前最临近区域为区域 B'。显然，在区域 A'构成的新训练集中，测试样本到其 k 个最近邻的距离均大于测试样本到区域 B'的距离，这表明在区域 B'中可能存在距离测试样本更近的训练样本。

此时，将区域 B'追加至新训练集，在剩余区域中，测试样本的当前最临近区域为区域 C'。显然，在区域 A'和区域 B'构成的新训练集中，测试样本到其 k 个最近邻的距离均小于测试样本到其最临近区域 C'的距离，这表明在剩余区域中不可能存在距离测试样本更近的训练样本。

最终，由区域 A'和区域 B'构成的新训练集中，包含了测试样本的 k 个最近邻。又由于训练集样本和测试样本的分布可以是任意的，因此在整个分类过程中，基于区域划分的 kNN 文本快速分类算法搜索到的测试样本的 k 个最近邻都是准确的，改进算法与经典算法具有相同的分类准确率。

由第 2.4 节和第 2.5 节可以看出：基于区域划分的 kNN 文本快速分类算法在确保 kNN 分类器准确率不变的前提下，显著提高了分类效率。该结论也在后续的实验中得到了验证。

3　实验及结果分析

3.1　实验设置

本文实验均在 IBM 兼容机上进行，系统配置如下：CPU 为 2.6GHz、内存为 1GB、操作系统为 Windows2000、编译器为 VC++6.0，其它工具软件有智动网页内容采集器 V1.8.0、中科院分词系统 ICTCLAS5.0 和批量文件重命名软件。

实验所需数据集是从新浪网新闻中心采集获得的，在军事、体育、财经、娱乐、汽车和数码 6 种类别网页中，从每一种类别中采集 1000 个样本，共 6000 个样本。其中训练集文本由每个类别中选出的 500 个样本构成，共计 3000 个样本，测试集文本为剩余的 3000 个样本。在数据预处理时，将特征项按信息增益(Information Gain，IG)[24] 降序排列，分别选择前 50 项、100 项、200 项、500 项和 1000 项作为最终的文本特征，并使用 TDIDF[25] 方法将所有文本向量化。

为了验证基于区域划分的 kNN 文本快速分类算法的有效性，本文设置了如下实验：

实验 1　基于区域划分的 kNN 文本快速分类算法准确性验证实验。通过实验验证改进算法与经典算法具有相同分类准确率。

实验 2　基于区域划分的 kNN 文本快速分类算法执行效率实验。通过实验讨论改进算法效率随参数 m 和 k 的变化情况。

实验 3　文本向量维数不同时改进算法效率比较实验。通过实验讨论改进算法效率随文本向量维数的变化情况。

3.2　实验数据及分析

观察实验数据可以发现：参数相同时，不同数据表中改进前后算法的执行时间均存在一定的差异，这是由于不同实验过程中系统运行环境不同以及初始区域中心选择的随机性造成的。虽然，在参数相同时，算法的执行时间和裁剪率存在一定差异，但是，在参数不同时，改进前后分类算法各自执行时间的变化趋势并没有发生改变；并且，参数相同时，算法的宏平均查准率、宏平均查全率、宏平均 F1 值和微平均等参数均相同。因此，实验相关数据以及根据其得出的结论都是有效的。

3.2.1　改进算法准确性验证实验

如表 1 所列，针对不同的 k 和 m 取值(表 1 中仅列出少量数据)，改进前后算法的各项性能指标完全一致，这表明基于区域划分的 kNN 文本快速分类算法的分类性能是完全可靠的。

表 1　算法改进前后准确率比较表(1000 维)

k	算法	宏平均查准率(%)	宏平均查全率(%)	宏平均 F1(%)	微平均(%)
1	改进前	91.53	91.17	91.15	91.17
	改进后(m=1)	91.53	91.17	91.15	91.17
2	改进前	92.18	92.00	91.97	92.00
	改进后(m=200)	92.18	92.00	91.97	92.00
3	改进前	92.15	91.90	91.89	91.90
	改进后(m=1000)	92.15	91.90	91.89	91.90
4	改进前	92.58	92.40	92.39	92.40
	改进后(m=2659)	92.58	92.40	92.39	92.40

3.2.2　改进算法执行效率实验

在表 2 和表 3 中，时间比＝经典算法执行时间：改进算法执行时间，该数据可以描述改进后算法的效率提升情况。

表 2　算法改进前后执行效率比较表($k=4$，1000 维)

算法	区域总数 m	裁剪率(%)	执行时间(s)	时间比
改进前	0	0	2612	1
改进后	1	0	1361	1.92
	10	0.03	571	4.57
	50	0.61	425	6.15
	100	2.79	362	7.22
	200	5.74	330	7.92
	500	16.20	380	6.87
	1000	34.32	449	5.82
	2000	74.07	580	4.50
	2659	99.86	673	3.88

从表 2 可以看出，$m=1$ 时，改进算法的效率最低，但与经典 kNN 分类算法相比，分类效率仍提高了 1.92 倍。这与式

(6)的 $f=\frac{kv+v}{v+d/2}=\frac{4\times1000+1000}{1000+3000/2}=2$ 是相符的。当 $m=200$ 时，改进算法效率达到最高，与经典 kNN 算法相比，效率提升了近 8 倍。可见，通过选择合适的参数值，改进算法较大幅度地提高了 kNN 分类器的分类效率。

改进算法提出的初衷是希望通过减少新训练集样本数目来提高效率。由于初级分类器描述的区域间存在重叠且位置关系与聚类间位置关系不一致，使得利用初级分类器查找到的最临近区域并不一定就是距离测试样本最近的聚类，裁剪率提升不大，最终使得通过裁剪样本来提升分类效率的效果并不明显。表 2 中，当 $m=200$ 时，改进算法的裁剪率却仅为 5.74%。可见，通过进一步优化区域划分方法来改进算法的执行效率还有很大的提升空间。

表 3　算法改进前后执行效率比较表($m=200$，1000 维)

k	改进前时间	改进后时间	时间比	宏平均 F1 值(%)
1	1081	215	5.03	91.15
4	2798	344	8.13	92.40
10	5932	621	9.55	91.67
20	11609	1164	9.97	90.70

在表 3 中，随着参数 k 的增加，改进算法效率的最大值也在提升。当算法准确率最高的参数 k 取较大值时，改进算法的效率与经典 kNN 分类算法相比将有很大提高。

3.2.3　文本向量维数不同时改进算法效率比较实验

表 4 中，时间比＝经典算法执行时间：改进算法执行时间。

表 4　维数不同时改进算法执行效率比较表($k=4$)

区域总数 m	文本维数 v	改进前时间	改进后时间	时间比	裁剪率(%)	宏平均 F1 值(%)
1	50	116	558	0.2079	0	85.69(85.73)
1	100	290	616	0.4708	0	90.41
1	200	573	778	0.7365	0	92.93
1	500	1377	1017	1.35	0	92.97
1	1000	2713	1424	1.91	0	92.39
200	50	116	17	6.82	70.36	85.69(85.73)
200	100	290	34	8.53	35.02	90.41
200	200	573	63	9.10	17.43	92.93
200	500	1377	166	8.30	5.22	92.97
200	1000	2713	354	7.66	5.63	92.39
2267	50	116	409	0.2836	99.70	85.69(85.73)
2576	100	290	495	0.5859	99.86	90.41
2657	200	573	558	1.03	99.79	92.93
2657	500	1377	601	2.29	99.86	92.97
2659	1000	2713	701	3.87	99.86	92.39

从表 4 可以看出，改进算法在最坏情况即 $m=1$ 时，分类效率随数据空间维数的下降而大幅度下降，并且当数据空间维数满足不等式 $v<d/(2k)=3000/8=375$ 时，改进算法效率要低于经典 kNN 分类算法；但是，在最好情况即 $m=200$ 时，改进算法效率提升相对稳定。当文本向量维数取最小 $v=50$ 时，改进算法效率是经典 kNN 分类算法的近 7 倍。可见，改进算法在维数较小时，通过选择合适的参数 m，同样可以得到很高的分类效率。

改进算法在最好情况下，裁剪率随向量维数的下降呈增加趋势。裁剪率的增加表明，划分得到的区域之间的重叠现象减少，且聚类区域间位置关系与初级分类器中描述的区域间位置关系趋于一致。这也表明，随着文本维数的减少，量化后的训练集样本分布更趋均匀化。这也是改进算法效率未随维数的减少而迅速降低的原因。更深层的原因有待作者后续研究。

3.2.4　相关 kNN 分类器改进算法实验结果的比较

表 5 中，时间比＝经典算法执行时间：改进算法执行时间。

表 5　几种改进算法执行效率比较表

算法来源	节约时间(%)	时间比	准确率变化情况
本文	89.97	9.97	与经典算法相同
文献[5]	78.00	4.55	提高
文献[10]	30.00	1.43	降低
文献[12]	70.00	3.33	提高
文献[13]	72.90	3.69	与经典算法相同
文献[18]	20.00	1.25	提高

表 5 中列出了几种改进算法的实验结果，可以看出，本文提出的基于区域划分的 kNN 文本快速分类算法的效率最高，并且算法的准确率与经典 kNN 分类算法完全相同。

另外，基于区域划分的 kNN 文本快速分类算法可以普遍适用于对各种数据类型下的 kNN 分类算法及其衍生算法进行改进，实用性较强。

结束语　kNN 方法作为一种简单、有效、非参数的分类方法，在文本分类中广泛应用。为了有效提高 kNN 分类器的效率，本文提出了一种基于区域划分的 kNN 文本快速分类算法，并使用数学推理和仿真实验的方法对算法的有效性进行了数学证明和实验验证。从实验结果可以看出，本文提出的改进算法在确保准确率不变的前提下，有效提高了 kNN 分类器的执行效率，较好地实现了初始设定的目标。下一步，作者将对区域划分方法进行深入研究，通过优化区域划分方法进一步提高算法的分类效率。

参 考 文 献

[1] Tan Pang-ning, Steinbach M, Kumar V. 数据挖掘导论[M]. 范明，范宏建，译. 北京：人民邮电出版社，2006：146-198

[2] 李荣陆. 文本分类及其相关技术研究[D]. 上海：复旦大学，2005

[3] Ogura H, Amano H, Kondo M. Feature selection with a measure of deviations from Poisson in text categorization [J]. Expert Systems with Applications, 2009, 36(3): 6826-6832

[4] 张著英，黄玉龙，王翰虎. 一个高效的 KNN 分类算法[J]. 计算机科学，2008，35(3)：170-172

[5] 钱晓东，王正欧. 基于改进 KNN 的文本分类方法[J]. 情报科学，2005，23(4)：550-554

[6] 钟将，孙启干，李静. 基于归一化向量的文本分类算法[J]. 计算机工程，2011，37(8)：47-49

[7] 刘海峰，张学仁，姚泽清，等. 基于类别选择的改进 KNN 文本分类[J]. 计算机科学，2009，36(11)：213-216

[8] Pan J S, Qiao Y L, Sun S H. A fast K nearest neighbors classification algorithm [J]. IEICE Trans FundamElectron Commun Comput Sci, 2004, E87-A(4): 961-963

[9] 孙荣宗，苗夺谦，卫志华，等. 基于粗糙集的快速 KNN 文本分类算法[J]. 计算机工程，2010，36(24)：175-177

[10] 孙荣宗. 一种快速 KNN 文本分类算法[J]. 电脑知识与技术，2010，6(1)：174-175，178

[11] 张国英，沙芸，江慧娜. 基于粒子群优化的快速 KNN 分类算法[J]. 山东大学学报：理学版，2006，41(3)：34-36

[12] 卜凡军，钱雪忠. 基于向量投影的 KNN 文本分类算法[J]. 计算机工程与设计，2009，30(21)：4939-4941

[13] 乔玉龙,潘正祥,孙圣和. 一种改进的快速 k-近邻分类算法[J]. 电子学报,2005,33(6):1146-1149
[14] Hart P E. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory,1968,14(3):515-516
[15] Wilson D L. Asymptotic properties ofnearestneighbor rules using edited data [J]. IEEE Transactions on Systems, Man and Cybernetics,1972,2(3):408-421
[16] Devijver P,Kittler J. Pattern recognition:A statistical ap-proach [M]. Englewood Cliffs:PrenticeHall,1982
[17] 李荣陆,胡运发. 基于密度的 KNN 文本分类器训练样本裁减方法[J]. 计算机研究与发展,2004,41(4):539-545
[18] 熊忠阳,杨营辉,张玉芳. 基于密度的 kNN 分类器训练样本裁剪方法的改进[J]. 计算机应用,2010,30(3):799-801,817
[19] 刘金岭,王朝,谢少峰. 基于聚类中心初始化的文本分类高效算法[J]. 软件导刊,2010,9(4):47-49
[20] 张孝飞,黄河燕. 一种采用聚类技术改进的 KNN 文本分类方法[J]. 模式识别与人工智能,2009,22(6):936-940
[21] Han Jia-wei,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,2007:223-262
[22] 张磊,刘建伟,罗雄麟. 基于 KNN 和 RVM 的分类方法——KNN-RVM 分类器[J]. 模式识别与人工智能,2010,23(3):376-384
[23] 知识发现[M]. 史忠植,译. 北京:清华大学出版社,2002
[24] Youn E,Jeong M K. Class dependent feature scaling method using naive Bayes classifier for textdatamining [J]. Pattern Recognition Letters,2009,30(5):477-485
[25] 罗欣,夏德麟,晏蒲柳. 基于词频差异的特征选取改进的 TF-IDF 公式[J]. 计算机应用,2005,25(9):2031-2033

(上接第 173 页)

名实体识别过程中,具有人名特点但识别率比较高的属性,可以划归为昵称类。QQ 娱乐明星数据源中的 year 和 day 元素结点表示实体的出生年和出生月,但它们之间是分开存放的。由于两个结点间不存在关联信息,无法将它们与日期年月实际联系起来,因此在处理这类相互关联但无实际指定关联信息的数据源属性时,只能采取用户手工映射的方法才能实现。在实际映射过程,也会出现同一属性映射到实体的不同描述信息上去,如 QQ 娱乐明星数据源中的 area(表示明星的所在地域)就映射到了实体的居住地和家乡两个实体描述信息上去。当实体的部分描述信息具有相似的特点或在数据源中的数据包含有实体多方面的描述信息时,这种情况就会出现,因此在映射时部分属性或属性的部分内容可以映射到实体的多个描述信息,但姓名、昵称、血型列,则一般只映射到单个实体描述信息。

结束语 本文提出了一种数据空间命名实体的集成模型及各异质异构数据源中命名实体的集成方法,实现了异构数据源到实体的映射。数据空间命名实体及其描述信息的集成从数据是由实体及其描述信息构成这一特征出发,从数据源中抽取命名实体的主要描述信息,以提高数据源查询能力。数据空间中的实体集成也可以帮助数据空间完成演化,发现数据源间的关系:一方面,根据数据源共同包含的同名实体情况,可以发现拥有相同内容的数据源;另一方面,可以通过数据空间中命名实体在不同数据源的分布情况,发现数据源间的关联。下一步工作是利用数据空间中的实体来发现数据源间的关系,增强数据空间的演化特性。

参 考 文 献

[1] Dong X L,Halevy A. A platform for personal information management and integration[C]//VLDB. 2005
[2] Franklin M,Halevy A,Maier D. From databases to dataspaces:a new abstraction for information management[J]. ACM Sigmod Record,2005,34(4):27-33
[3] Halevy A,Franklin M,Maier D. Principles of dataspace systems [C]//ACM SIGMOD. 2006
[4] Franklin M, Halevy A, Maier D. A first tutorial on dataspaces [J]. Proceedings of the VLDB Endowment,2008,1(2):1516-1517
[5] Dittrich J P,Salles M A V. iDM:a unified and versatile data model for personal dataspace management[C]//VLDB. 2006
[6] Pradhan S. Towards a novel desktop search technique:Database and Expert Systems Applications[C]//Regensburg. Germany, 2007
[7] Ming Z, Mengchi L, Qian C. Modeling heterogeneous data in dataspace[C]//Information Reuse and Integration. IRI 2008. IEEE International Conference,July 2008:404-409
[8] 孟小峰. 数据空间技术研究进展[R]. 北京,2008
[9] Sarma A,Dong X,Halevy A. Data modeling in dataspace support platforms[J]. Conceptual Modeling:Foundations and Applications,LNCS,2009,5600:122-138
[10] Jiang X, Sun X, Zhuge H. A Resource Space Model for Dataspace:Semantics Knowledge and Grid (SKG)[C]//2010 Sixth International Conference. 2010
[11] Yang D,Shen D,Nie T,et al. Layered graph data model for data management of dataspace support platform[J]. Web-Age Information Management,LNCS,2011,6897:353-365
[12] Dong X,Halevy A. Indexing dataspaces[C]//the 2007 ACM SIGMOD International conference on Management of Data. Beijing,China,2007
[13] Sarma A D,Dong X L,Halevy A Y. Uncertainty in Data Integration and Dataspace Support Platforms[M]. Schema Matching and Mapping,part 1,2011:75-108
[14] 刘莉,郭艳艳,吴扬扬. 一种基于基本信息单元的索引[J]. 计算机工程与科学,2011(9):117-122
[15] Chung P W H,Liao Z. Cross-organisation dataspace (COD)-architecture and implementation[C]//International Conference on Computer Science and Software Engineering. Wuhan, China, 2008
[16] 董彦磊,申德荣,寇月,等. 数据空间中数据组织模型以及关联关系发现模型的研究[C]//第 26 届中国数据库学术会议. 南昌,中国,2009
[17] Dong Y,Shen D,Nie T,et al. Discovering Relationships among Data Resources in DataSpace[C]//Web Information Systems and Applications Conference(WISA 2009). 2009
[18] Yang D,Shen D,Nie T,et al. Layered graph data model for data management of dataspace support platform[C]//the 12th International Conference on Web-age Information Management (WAIM '11). Wuhan,China,Springer-Verlag,2011
[19] 孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术,2010(6):42-47