第 35 卷 第 9 期 计 算 机 工 程 2009 年 5 月
Vol.35 No.9 Computer Engineering May 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)09—0222—03 文献标识码：A 中图分类号：TP391

基于 K-Modes 聚类的自适应话题追踪技术

任晓东[1,2]，张永奎[1,2]，薛晓飞[1,2]
(1. 山西大学计算机与信息技术学院，太原 030006；2. 山西大学计算智能与中文信息处理教育部重点实验室，太原 030006)

摘 要：传统自适应话题追踪用向量空间模型表示一个话题模型，通常会对话题模型更新带来错误的反馈。针对传统自适应话题追踪中话题模型的不足，提出基于 K-Modes 聚类的自适应话题追踪方法(K-MATT 方法)，用话题类中心代替话题模型，把命名实体向量空间模型作为话题类中心，在追踪过程中不断迭代更新话题类中心，直到话题类中心稳定。实验证明 K-MATT 方法是有效的。

关键词：话题追踪；K-MATT 方法；话题类中心

Adaptive Topic Tracking Technique Based on K-Modes Clustering

REN Xiao-dong[1,2], ZHANG Yong-kui[1,2], XUE Xiao-fei[1,2]
(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006)

【Abstract】Traditional Adaptive Topic Tracking(ATT) uses VSM to express a topic model and bring mistaken feedback to topic model updating. This paper presents an Adaptive Tracking Technique based on K-Modes clustering(K-MATT) to solve the problems caused by traditional topic model expression. This method uses Topic Kind Center(TKC) to substitute topic model and uses named entities VSM to express TKC, updates TKC in topic tracking until TKC is stable. Experiments prove K-MATT method is effective.

【Key words】Topic Tracking(TT); K-MATT method; Topic Kind Center(TKC)

1 概述

话题追踪(Topic Tracking, TT)是话题识别与追踪(Topic Detection and Tracking, TDT)的 5 个子任务[1]之一。在话题追踪中，现有的话题训练模型不能很好地表示话题的内容，在实际应用中用户对突发性新闻具备的先验知识通常也很少，这就造成初始训练得到的话题模型不够充分和准确。因此，一种具备自学习能力的无指导自适应话题追踪[2](Adaptive Topic Tracking, ATT)逐渐成为 TT 领域新的研究趋势。

现有的自适应话题模型大都用向量空间模型表示。向量空间模型把时间、地点、人物和内容压缩到一个向量中进行表示，不仅不利于表示出上面提到的细微差别，而且对话题的事件的表示不够形象直观。 因此，在自适应话题模型更新过程中对话题特征词更新不准确，从而造成了话题漂移[3]。

2 K-MATT方法

2.1 K-Modes聚类概述

K-Modes 聚类是 Huang[4]在 k-means 聚类方法基础上提出的，是对 k-means 聚类方法的一种扩展。在聚类过程中用 modes 来代替 means。算法类似于 k-means 算法，交替更新聚类中心和划分矩阵，直到代价函数值不再变化。

2.1.1 相异度测量

X,Y 是 2 个由 m 维分类属性描述的 2 类分类对象。X,Y 的相异度测量[4]为 2 个分类对象相应的分类属性的错误匹配总数。其中，错误匹配总数最小的为 2 个对象中最相似的。这个相异度测量运用到简单的匹配算法，如：

$$d_1(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \tag{1}$$

其中，$\delta(x_j,y_j)=\begin{cases}0 & x_j=y_j\\ 1 & x_j\neq y_j\end{cases}$。

2.1.2 分类对象中的 mode

假设 $X=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$ 是一组数据组元组，其中 $X_i=[x_{i1},x_{i2},\cdots,x_{im}]$ 表示具有 m 个属性的数据对象。设向量 $\boldsymbol{Q}=[q_1,q_2,\cdots,q_m]$ 是 $X=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$ 的一个 mode，则相异匹配[4]总数为

$$D(S,\boldsymbol{Q})=\sum_{i=1}^{n}d_1(\boldsymbol{X}_i,\boldsymbol{Q}) \tag{2}$$

其中，$\boldsymbol{Q}$ 不一定是 X 的一个元素。

2.2 话题文档预处理

话题文档预处理主要是对所有保存为 TXT 文本的话题报道，用哈尔滨工业大学的命名实体抽取软件抽取话题文档的命名实体。

2.3 话题模型表示

根据新闻报道的特点，用命名实体向量空间模型[5-6] $X=\{\boldsymbol{x}_{ns},\boldsymbol{x}_{nt},\boldsymbol{x}_n,\boldsymbol{x}_m,\boldsymbol{x}_t,\boldsymbol{x}_{nz},\boldsymbol{x}_{nr}\}$ 表示话题类中心。用 $\boldsymbol{ns}_1,\boldsymbol{w}_{(ns)1}$，$\boldsymbol{nt}_1,\boldsymbol{w}_{(nt)1},\boldsymbol{n}_1,\boldsymbol{w}_{(n)1},\boldsymbol{m}_1,\boldsymbol{w}_{(m)1},\boldsymbol{t}_1,\boldsymbol{w}_{(t)1},\boldsymbol{nz}_1,\boldsymbol{w}_{(nz)1},\boldsymbol{nr}_1,\boldsymbol{w}_{(nr)1}$ 来分别表示一篇报道的地名、机构名、名词、数词、日期、专有名词、人名的词向量模型及各个分类命名实体的词频。在本文的地名命名实体去除了如省、市、县、乡、村、地区的后缀。话题模型表示如下：

$$x_{ns}=\{\boldsymbol{ns}_1/\boldsymbol{w}_{(ns)1};\boldsymbol{ns}_2/\boldsymbol{w}_{(ns)2};\cdots\}$$

基金项目：国家自然科学基金资助项目(60475022)；山西省自然科学基金资助项目(20041041)；山西省回国留学人员基金资助项目(2002004)

作者简介：任晓东(1980－)，男，硕士研究生，主研方向：中文信息处理；张永奎，教授；薛晓飞，硕士研究生

收稿日期：2008-12-01 **E-mail：**2000rxdmail@163.com

$x_{nt}=\{\boldsymbol{nt}_1/\boldsymbol{w}_{(nt)1};nt_2/\boldsymbol{w}_{(nt)2};\cdots\}$

$x_n=\{\boldsymbol{n}_1/\boldsymbol{w}_{(n)1};\boldsymbol{n}_2/\boldsymbol{w}_{(n)2};\cdots\}$

$x_m=\{\boldsymbol{m}_1/\boldsymbol{w}_{(m)1};\boldsymbol{m}_2/\boldsymbol{w}_{(m)2};\cdots\}$

$x_t=\{\boldsymbol{t}_1/\boldsymbol{w}_{(t)1};\boldsymbol{t}_2/\boldsymbol{w}_{(t)2};\cdots\}$

$x_{nz}=\{\boldsymbol{nz}_1/\boldsymbol{w}_{(nz)1};\boldsymbol{nz}_2/\boldsymbol{w}_{(nz)2};\cdots\}$

$x_{nr}=\{\boldsymbol{nr}_1/\boldsymbol{w}_{(nr)1};\boldsymbol{nr}_2/\boldsymbol{w}_{(nr)2};\cdots\}$

如发生在 2008 年的拉萨打砸抢烧事件中的一篇报道如表 1 所示。

表 1　拉萨打砸抢烧事件话题表示模型

NE	命名实体向量及词频表示
ns	拉萨铁路/1，拉萨/6，西藏自治区/1，…
nt	拉萨市建设局/1，拉萨市第二中学/1，西藏自治区商务厅/1，…
n	情况/1，加油站/1，铁路/1，街上/1，…
m	1 235 人/1，2 个/1，…
t	17 日/1，18 日上午/1，3 月 14 日/1，…
nz	
nr	德吉卓嘎/1，张贵国/1，…

2.4　相似度计算

传统的相似度计算方法有余弦相似度计算法等。本文中引用 K-Modes 算法思想利用简单匹配的方法进行相似度测量。

根据话题的文档表示和式(1)变化为以下计算方法:

$$d_1(\boldsymbol{Q},\boldsymbol{X})=\sum_{k=1}^{7}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(q_{ki},x_{kj})$$

$$\delta(x_j,y_j)=\begin{cases}1 & x_j=y_j\\ 0 & x_j\neq y_j\end{cases} \tag{3}$$

其中，$\boldsymbol{Q}$ 表示初始话题类中心；$\boldsymbol{X}$ 表示待识别话题类中心；q_{ki} 表示话题类中心第 k 类命名实体的第 i 个词；x_{kj} 表示一个报道第 k 类命名实体的第 j 个词。

用 $\boldsymbol{Q}$ 和 $\boldsymbol{X}$ 的每个分类值匹配，匹配为 1，其他为 0，然后分别对各个分类匹配值累加。则相似度公式定义为

$$sim(\boldsymbol{Q},\boldsymbol{Y})=\sum_{k=1}^{7}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(q_{ki},x_{kj})(w_{ki}+v_{kj}) \tag{4}$$

其中，w_{ki}，v_{kj} 分别为 q_{ki}，x_{kj} 的词频。

相似度归一化处理:

$$sim(\boldsymbol{Q},\boldsymbol{Y})=\sum_{k=1}^{7}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(q_{ki},x_{kj})(w_{ki}+v_{kj})/(w+v) \tag{5}$$

其中，w，v 分别 $\boldsymbol{Q}$，$\boldsymbol{Y}$ 的词频总数。

2.5　类中心更新算法

本文的初始类中心选用话题的第 1 篇报道。由于本算法总是选择相似度最高的话题进行更新，直到话题类中心趋于稳定，用稳定的类中心对话题进行再次追踪。算法主要对话题类中心 $\boldsymbol{Q}_0$ 进行更新，即 $\boldsymbol{Q}_0\Rightarrow\boldsymbol{Q}_m$，其中，$\boldsymbol{Q}_0$ 为初始话题类中心，$\boldsymbol{Q}_m$ 为更新后话题类中心。

当代价函数 $\max mise(\boldsymbol{Q}_{m-1},\boldsymbol{Q}_m)>\gamma$ 时更新结束。算法如下:

(1)若 $sim(\boldsymbol{Q}_0,\boldsymbol{X}_i)=\max$，则

$\boldsymbol{Q}_1=\{\boldsymbol{Q}_0\cup\boldsymbol{X}_i\}$，$sort(\boldsymbol{Q}_1)$，$sim(\boldsymbol{Q}_0,\boldsymbol{X}_i)=0$

(2)若 $sim(\boldsymbol{Q}_1,\boldsymbol{X}_j)=\max$，则

$\boldsymbol{Q}_2=\{\boldsymbol{Q}_1\cup\boldsymbol{X}_j\}$，$sort(\boldsymbol{Q}_2)$，

$\boldsymbol{Q}_2=\{\boldsymbol{Q}_1\cup\boldsymbol{X}_j\}/2$，$sim(\boldsymbol{Q}_0,\boldsymbol{X}_j)=0$

(3)若 $sim(\boldsymbol{Q}_m,\boldsymbol{X}_k)=\max$，则

$\boldsymbol{Q}_m=\{\boldsymbol{Q}_{m-1}\cup\boldsymbol{X}_k\}$，$sort(\boldsymbol{Q}_m)$，

$\boldsymbol{Q}_m=\{\boldsymbol{Q}_{m-1}\cup\boldsymbol{X}_k\}/2$，$sim(\boldsymbol{Q}_m,\boldsymbol{X}_k)=0$

(4)若 $\max mise(\boldsymbol{Q}_{m-1},\boldsymbol{Q}_m)>\gamma$，则结束更新，否则 goto(3)。

其中，$sort(\boldsymbol{Q})$ 表示以词频对 $\boldsymbol{Q}$ 进行排序；$\boldsymbol{Q}_m=\{\boldsymbol{Q}_{m-1}\cup\boldsymbol{X}_k\}$ 表示合并 $\boldsymbol{Q}_{m-1}$ 与 $\boldsymbol{X}_k$ 的每个分类向量模型，如果命名实体分类项匹配，则对匹配的命名实体分类项进行词频合并。

2.6　代价函数

在本文中代价函数定义为更新前后的 2 个话题类中心相似度，如果比值趋于稳定，则话题类中心更新结束。在这里代价函数表示为

$$\max mise(\boldsymbol{Q}_{m-1},\boldsymbol{Q}_m)=\sum_{k=1}^{7}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(\boldsymbol{Q}_{(M-1)ki},\boldsymbol{Q}_{(M)kj})(w_{ki}+v_{kj})/(w+v) \tag{6}$$

其中，$\boldsymbol{Q}_{m-1}$ 表示更新前话题类中心；$\boldsymbol{Q}_m$ 表示更新后话题类 中心。

2.7　话题类中心更新后对话题的再次跟踪

在话题追踪过程中不同的话题在同一训练语料追踪时的阈值会有所差别，仅用人工经验定义一个固定的阈值会影响不同话题追踪效果。因此，本文用最后一次话题类中心更新时的相似度 $sim(\boldsymbol{Q}_m,\boldsymbol{X}_k)-\theta$ 作为追踪阈值。

$$\lambda=sim(\boldsymbol{Q}_m,\boldsymbol{X}_k)-\theta \tag{7}$$

3　实验结果及分析

3.1　训练语料和测试语料

本文实验使用突发事件语料作为训练语料。语料中共有 2 100 多篇突发事件。本文实验对 58 篇话题进行测试。

3.2　评测机制

为了准确地评价不同系统的性能，本文使用 TDT 的评测方法。由于 TDT 的所有任务都可看作是检测任务，因此它们的性能都可以用误报率和漏报率表示。这 2 个错误率合并成 1 个检测开销，C_{Det} 就是误报率和漏报率加权求和的结果，其计算公式为

$$C_{Det}=C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot P_{non\text{-}target} \tag{8}$$

其中，C_{Det} 是系统的错误识别代价；C_{Miss} 与 C_{FA} 分别是漏报和误报的代价，它们的值通常根据应用预选给定，在本文中分别取 10 和 1，即认为漏报的代价要高很多；P_{Miss} 与 P_{FA} 分别是系统识别的漏报率和误报率，漏报率是指系统没有识别出来的关于某话题的新闻报道的数目与语料中描述该话题的新闻报道的总数之比，而误报率是指对某个话题来说判断错误的新闻报道的数目与语料库中所有没有描述该话题的新闻报道的总数之比；P_{target} 是一个先验的目标出现概率 $P_{non\text{-}target}=(1-P_{target})$，表示关于某个话题的新闻报道出现的可能性，它的值通常也根据具体应用给出。为了使得到的性能指标落在更有意义的范围内，将错误识别代价 C_{Det} 做归一化处理得到 $(C_{Det})_{Norm}$。

$$(C_{Det})_{Norm}=C_{Det}/\min(C_{Miss}\cdot P_{target},C_{FA}\cdot P_{non\text{-}target}) \tag{9}$$

本文采用 $Norm(C_{Det})$ 值作为评价话题追踪系统性能的指标。$Norm(C_{Det})$ 值越小系统的性能越好。

3.3　实验结果

本文实验使用突发事件语料作为测试语料。实验考察代价函数 γ =0.990 时，θ 从 0.30~0.40 的追踪实验结果。本文用 K-MATT 方法与传统的向量空间模型自适应追踪方法作比较。结果如表 2、表 3 所示。

表 2　K-MATT 方法追踪结果

θ	$(C_{Det})_{Norm}$	C_{Det}	P_{Miss}	P_{FA}
0.30	0.35	0.134	0.31	0.144
0.35	0.252 6	0.096	0.21	0.016
0.40	0.152	0.058	0.10	0.018 7

表 3　传统向量空间模型自适应话题追踪结果

θ	$(C_{Det})_{Norm}$	C_{Det}	P_{Miss}	P_{FA}
0.30	0.39	0.242	0.32	0.044 7
0.35	0.303	0.188	0.22	0.051
0.40	0.264	0.164	0.129	0.089

3.4　结果分析

本文实验用 K-Modes 自适应话题追踪算法和传统向量空间模型的自适应追踪算法作了比较。

用突发事件作为实验语料。实验表明在阈值 θ =0.40 时话题追踪效果最好。

如表 2 所示用 K-Modes 自适应算法在突发事件话题追踪过程中误识率较传统话题追踪方法有较大提高。话题追踪损耗 $(C_{Det})_{Norm}$ 较传统向量空间模型自适应算法降低了 42%。由于语料库中矿难突发事件语料比较多，因此在追踪矿难类型突发事件话题时效果不是很好。

在以后的研究中应该考虑不同命名实体对话题追踪的影响。

4　结束语

本文尝试用命名实体向量模型来表示一个话题，把命名实体向量作为自适应话题追踪的更新模板，在自适应话题追踪过程中有效保留了对话题贡献较大的特征项。引用 K-Modes 聚类思想迭代更新话题模板。实验与传统向量空间模型自适应算法作了比较，证明该方法是有效的。

参考文献

[1] 李保利, 俞士汶. 话题识别与话题追踪[J]. 计算机工程与应用, 2003, 39(17): 6-10.

[2] 洪　宇, 张　宇, 刘　挺, 等. 话题检测与追踪的评测及研究综述[J]. 中文信息学报, 2007, 6(21): 77-79.

[3] 王会珍, 朱靖波, 季　铎, 等. 基于反馈学习自适应的中文话题追踪[J]. 中文信息学报, 2006, 3(20): 92-98.

[4] Huang Zhexue. Extensions to the k-Means Algorithm for Clustering Large Data with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 287-290.

[5] Kumaran G, Allan J. Text Classification and Named Entities for New Event Detection[C]//Proc. of the 27th Annual International ACM SIGIR Conference. New York, USA: ACM Press, 2004: 297-304.

[6] 林鸿飞, 宋　丹, 杨志豪. 基于语义框架的话题追踪方法[C]//中国中文信息学会二十五周年学术会议. 北京: 清华大学出版社, 2006: 383-384.

编辑　任吉慧

（上接第 221 页）

曲线和 IGA 算法[5]收敛曲线的比较。可以看出，本算法能在开始的有限代内使结果快速收敛到最优值附近，而收敛的快速性是网络路由选择算法考虑的最重要因素之一，因此，本算法在速度方面具有很大优势。

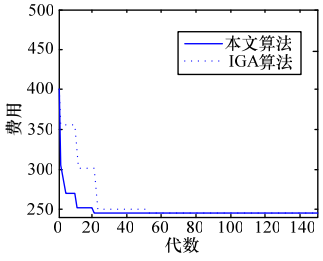

图 2　收敛曲线比较

表 1 给出了在目的节点数变化的情况下，有限代数内本文算法与 IGA 算法在寻优性能方面的比较结果。可以看出，在目的节点数量较少、代数足够多的情况下，本文算法与 IGA 效果相近，但当目的节点数量增加，且迭代数量有限时，本文极值遗传算法的成功率明显高于 IGA。

表 1　IGA 算法和本文算法的寻优成功率比较　(%)

max$G$	IGA 算法		本文极值遗传算法	
	n=3	n=6	n=3	n=6
50	78	34	80	55
100	85	75	86	82
150	98	92	98	95

6　结束语

本文引入极值优化思想，将极值优化与遗传算法相结合，通过极值优化的跳变避免算法陷入局部解，并加快其收敛速度。在算法编码中，充分考虑解空间的完全性，使编码空间和解空间能一一对应。在制定交叉和变异规则时，充分考虑解的可行性，避免操作后出现重复或不可行解的现象。

参考文献

[1] Wang Zheng, Crow C. Quality of Service for Supporting Multimedia Applications[J]. IEEE Journal on Selected Areas in Communications, 1996, 14(7): 1228-1234.

[2] 陈　杰, 张洪伟. 基于自适应蚁群算法的 QoS 组播路由算法[J]. 计算机工程, 2008, 34(13): 200-203.

[3] Randaccio L S, Atzori L. Group Multicast Routing Problem: A Genetic Algorithms Based Approach[J]. Computer Networks, 2008, 51(14): 3989-4004.

[4] Boettcher S, Percus A G. Extremal Optimization: Methods Derived from Co-evolution[C]//Proc. of the Genetic and Evolutionary Computation Conf.. San Francisco, USA: Morgan Kaufmann, 1999: 825-832.

[5] 赵秀平, 谭冠政. 基于免疫遗传算法的多约束 QoS 组播路由选择方法[J]. 计算机应用, 2008, 28(3): 591-595.

编辑　陈　晖