时间信息在话题检测中的应用研究*)

赵 华[1] 赵铁军[1] 赵 霞[2]

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)[1] (泗水县高峪中心中学 济宁 273212)[2]

摘 要 为了克服话题检测中使用静态阈值的缺点,我们提出了基于时间信息的动态阈值模型。在该模型中,探索了一种比值法来选择与某个特定报道最相似的话题。实验结果表明,动态阈值模型很好地改善了话题检测系统的性能。

关键词 话题检测,动态阈值,比值法

Using Temporal Information in Topic Detection

ZHAO Hua[1] ZHAO Tie-Jun[1] ZHAO Xia[2]

(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)[1] (GaoYu Middle School, Jining 273212)[2]

Abstract In order to overcome the shortcoming of the static threshold in the topic detection research, we propose a dynamic threshold model incorporating temporal information as a major component. In this model, we explore a ratio method to select the optimal topic. Experimental results indicate that the model proposed in this paper is very successful.

Keywords Topic detection, Dynamic threshold, Ratio method

1 引言

话题检测是的话题检测与跟踪(Topic Detection and Tracking,简称 TDT)评测中的一项评测任务,它是指将新闻数据流中的报道归入不同的话题,并在必要时建立新话题的技术[1]。它本质上类似于无指导的聚类研究,但是聚类是基于全局信息实现的,而话题检测是以增量方式实现的。话题检测可以分为两个阶段:检测新话题的出现;将后续报道加入相关的话题。图 1 表示了一个话题检测系统的基本思想。

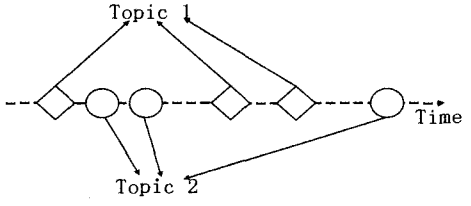

图 1 话题检测的基本思想

话题检测研究包括以下几个方面的关键技术:报道与话题模型,相似度计算,阈值策略以及检测算法等。文中首先介绍了我们在基本模型中对上述几个问题的解决方法,然后提出了一种基于时间信息的动态阈值模型,并探索使用一种比值法来选择最相似的报道。该模型克服了传统的静态阈值的缺点,取得了很好的效果。

本文第 2 节简单介绍了在话题检测研究中常用的方法,并分析了动态阈值模型的思想来源;第 3 节介绍了我们用于话题检测的基本模型;第 4 节详细描述了基于时间信息的动态阈值模型;第 5 节是实验及结果分析;最后给出了本文得出的结论。

2 相关工作

目前大多数话题检测算法都是基于某种聚类思想实现,比如:Single-Pass 聚类方法[2,3],K-NN[4],K-means[5] 以及层次聚类算法[2]。话题通常用质心来表示,话题特征项的权值重通常用 TFIDF 计算。

在话题检测研究中,通常都是预先设定阈值,并且一旦设定在整个检测过程中就不再改变,这种做法是不合理的。首先,由于话题的时间集中性[6],随着时间的推移,和该话题相关的报道将越来越少,所以应该在检测过程中逐渐加大阈值;其次,不同的话题应该采取不同的阈值。有些研究者提出了基于报道在信息流中的位置的动态阈值策略[6],但是有时和某个话题相关的报道可能会很多,所以上述的动态阈值的方法会使得很多和话题相关的报道不能被检测到。基于以上的分析,我们提出了基于话题的持续时间的动态阈值模型。

3 话题检测基本模型

这节将详细介绍我们用于话题检测的基本模型,将其作为我们后续的改进方法的基础。

3.1 预处理及报道模型

在基本模型中,预处理操作包括去停用词以及词形还原。

我们用向量空间模型来表示报道。假设 S 是一个经过预处理的报道,$term_1$,$term_2$,…,$trem_k$ 是出现在 S 中的 k 个不同的词,那么 S 可以表示成:$S=(term_1,w_1;term_2,w_2;term_k,w_k)$,$w_i$ 是 $term_i$ 在 S 中的权值,由公式(1)中所示的 TFIDF 公式计算得到。

$$w_i=tf_i\times\log(N/n_i+0.01) \quad (1)$$

其中,tf_i 是 $term_i$ 在 S 中的词频,N 是所有已输入报道的总数,n_i 是这 N 个报道中含有 $term_i$ 的报道的个数。

3.2 基于 VSM 的话题模型

在基本模型中,话题由单一质心表示,质心用向量空间模型(VSM)表示。为了将话题表示成质心,须经过抽取特征项和计算特征项权值两步。

*)基金项目:国家自然科学重点基金项目(60435020),国家 863 高科技项目基金资助项目(2004AA117010-08)。赵 华 博士生,主要研究方向为自然语言处理;赵铁军 博士,教授,博导,研究方向自然语言处理、机器翻译、人工智能;赵 霞 教师。

我们使用文档频次(Document Frequency, DF)作为抽取特征项的标准,抽取过程如下:

(1)统计当前时刻话题中所有不同词在话题中的文档频次,并将这些词按照其文档频次的高低顺序排列。话题中包含的词的集合以及词在话题中的 DF 值随着相关报道的加入而不断变化,它们可以由公式(2)以及公式(3)计算得到:

$$T(t)=\bigcup_{s_i\in T}\{word \mid word\in S_i, word_k\neq word_j\} \quad (2)$$

$$DF(word,t)=\sum_{S_i\in T} Appear(word,S_i) \quad (3)$$

其中 $T(t)$ 和 $DF(word,t)$ 分别表示在 t 时刻话题 T 中包含的词的集合以及词 $word$ 在 T 中的文档频次,S_i($1\leqslant i\leqslant StoryNumber$)是 t 时刻 T 中包含的报道,$1\leqslant k,j\leqslant|T(t)|$,$StoryNumber$ 是话题 T 在 t 时刻包含的报道的个数;当 $word$ 在 S_i 中出现时,$Appear(word,S_i)=1$,否则 $Appear(word,S_i)=0$。

(2)从上述排列中按照文档频次从高到低的顺序抽取特定数目的词作为话题的特征项。

话题的特征项的权值是通过简单平均的方法计算得到的,如公式(4)所示:

$$Weight(\delta,t,T)=\sum_{S_i\in T} W(\delta,S_i)/StoryNumber \quad (4)$$

其中 δ 表示话题 T 的一个特征项,$Weight(\delta,t,T)$ 表示 δ 在 t 时刻在 T 中的权值,$W(\delta,S_i)$ 是 δ 在 S_i 中的权值,由公式(1)计算得到。

3.3 报道和话题的相似度

在基本模型中,使用 Cosine 函数计算报道和话题之间的相似度。假设 $w_{s1},w_{s2},\cdots,w_{sn}$ 和 $w_{t1},w_{t2},\cdots,w_{tn}$ 分别表示特征 $\delta_1,\delta_2,\cdots,\delta_n$ 项在报道 S 以及话题 T 中的权值,那么用 Cosine 计算的 S 和 T 之间的似度如下所示:

$$Cos(S,T)=\sum_{k=1}^{n} w_{sk}\times w_{tk}/\sqrt{\sum_{k=1}^{n} w_{sk}^2\times\sum_{k=1}^{N} w_{tk}^2} \quad (5)$$

3.4 基于 Single-Pass 聚类方法的话题检测算法

我们基于 Single-Pass 聚类策略算法实现话题检测。该算法按报道输入的先后顺序依次处理数据流中的报道,直到所有的报道处理完毕,具体过程如下:

(1)对报道进行预处理,建立报道的向量模型;

(2)如果报道是数据流中的第一个报道,则成立一个以该报道为种子的话题,并建立话题的向量模型;

(3)如果报道不是数据流中的第一个报道,则计算报道和已存在话题的相似度,记录最高相似度以及取得最高相似度的话题 T。如果最高相似度超过了预设的阈值,则表示报道和话题 T 相关,将报道加入话题 T,并更新话题 T 的向量模型;否则成立一个以该报道为种子的新话题并建立该话题的向量模型;

(4)重复上述过程直到数据流中的所有报道都处理完毕。

4 基于时间信息的动态阈值模型

基于第 2 节的分析,提出了一个基于时间信息的动态阈值模型,如下所示:

$$Threshold(T,t)=\theta+\alpha^{*}(Time(S)-Time(T)) \quad (6)$$

其中 $Threshold(T,t)$ 表示 t 时刻话题 T 的阈值;θ 是一个常数,表示话题刚建立时的阈值;α 是一个可调参数,表示时间信息在动态阈值中所占的比例。$Time(S)$ 和 $Time(T)$ 分别表示报道时间和话题的建立时间,话题的建立时间是指话题的种子报道的时间。实验中,依据我们所使用评测语料中所标记的时间,报到时间和话题的建立时间都是以天为单位。

在基于动态阈值模型的话题检测算法中,每个话题的阈值都可能不相同。为了让报道和不同的话题的相似度具有可比性,提出了比值法,用来选择与某个报道最相似的话题,其基本思想如下:

(1)首先计算报道和话题的相似度,然后计算相似度与该话题的阈值的比值;

(2)得到最大比值(大于 1)的话题作为与报道最相似的话题;

(3)如果所有的比值都小于 1,那么新成立一个话题。

5 实验和结果分析

本文采用 LDC 提供的 TDT Pilot 语料作为评测语料,该语料包含 15,863 个英文报道,所有报道组成一个文本文件,报道时间从 1994 年 7 月 1 日到 1995 年 6 月 30 日。语料中定义了 25 个话题,覆盖了语料中的报道涉及的部分话题。

5.1 评价标准

我们采用 TDT 评测中所使用的归一化检测开销 $(C_{Det})_{Norm}$ 来评测我们的话题检测系统,公式如下:

$$(C_{Det})_{norm}=\frac{C_{miss}P_{miss}P_{t\,arg\,et}+C_{fa}\cdot P_{fa}\cdot P_{\neg t\,arg\,et}}{\min(C_{miss}\cdot P_{t\,arg\,et},C_{fa}\cdot P_{\neg t\,arg\,et})} \quad (7)$$

其中,P_{Miss} 为系统的漏报率;P_{FA} 为系统的误报率;P_{target} 为在信息流中看到一个新话题的概率,而 $P_{\neg target}$ 是在信息流中看到一个老话题的概率,$P_{\neg target}=1-P_{target}$;$C_{Miss}$ 为漏报一个新事件的代价;C_{FA} 为误报一次的代价。

实验中,C_{Miss}, C_{FA} 和 P_{target} 的值分别为 1.0,0.1,0.02[1]。

5.2 实验结果分析

我们设计了两个话题检测系统,分别称为 SYSTEM-1 和 SYSTEM-2,描述如下:

(1)SYSTEM-1:该系统也可称为 baseline,基于基本模型实现,其中阈值设为 0.4,话题质心的特征项个数为 100。其评测结果为:$P_{Miss}=0.2766$,$P_{FA}=0.0120$,$(C_{Det})_{Norm}=0.3253$。

(2)SYSTEM-2:该系统与 SYSTEM-1 的唯一区别是在该系统中我们采用了基于时间信息的动态阈值模型,而不是静态阈值,其中 $\theta=0.4$。我们在 α 取不同值时对 SYSTEM-2 的性能进行了验证,结果如表 1 所示。

表 1 SYSTEM-2 实验结果

α	P_{Miss}	P_{FA}	$(C_{Det})_{Norm}$
0.001	0.2893	0.0021	0.2997
0.002	0.2299	0.0017	0.2381
0.003	0.2469	0.0014	0.2539
0.004	0.2451	0.0011	0.2505
0.005	0.3056	0.0009	0.3101

从表 1 可以看出本文提出的基于时间信息的动态阈值模型取得了很好的效果。但是,随着 α 值得增大,话题的阈值将变得越来越大,结果可能使得许多相关报道不能被检测到。所以在使用本文提出的动态阈值模型时,α 应该取比较小的值。

结论 本文首先介绍了我们的话题检测系统中使用的基础模型,然后提出了基于时间信息的动态阈值模型。从实验结果可以看出该动态阈值模型取得了很好的效果。从本文的

实验可以看出，时间是话题的重要特征，其应该在话题检测研究得到更加充分的应用。

参 考 文 献

1 The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan. Available at: http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm, April 2003

2 Allan J, Carbonell J. Topic Detection and Tracking Pilot Study: Final Report. In: Proceeding of the DARPA Broadcast News Transcriptions and Understanding Workshop, February, 1998. 194～218

3 Yang Yiming, Carbonell J, Brown R, Pierce T, Archibald B T, Liu Xin. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32～43

4 Yang Yiming, Ault T, Pierce T, Lattimer C. Improving text categorization methods for event tracking. In: Proc. ACM SIGIR, 2000. 65～72

5 Lavrenko V, Allan J, DeGuzman E, LaFlamme D, Pollard V, Thomas S. Relevance models for topic detection and tracking. In: Proc. Human Language Technology Conference (HLT), 2002

6 Allan J, Papka R, Lavrenko V. On-line New Event Detection and Tracking. In: Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Australia, August 1998

7 Allan J, Lavrenko V, Jin H. First story detection in TDT is hard. In: Proc. 9th Conference on Information Knowledge Management CIKM, McClean, VA USA, 2000. 374～381

(上接第 203 页)

公式的含义是：计算各特征词类别 FC_j 中的特征词 W_i 到该类别中心对应的特征词 Z_j 的距离，并求这些距离之和，得到各类的适应度。所有类的适应度之和加 1 并求倒数，得到染色体 Ind 的适应度。

4.5 遗传算子

选择策略对算法性能的影响起到举足轻重的作用。在种群进化过程中，采取精英保留策略，首先保留遗传过程中的精英个体。然后采用轮盘赌法，由适应度函数对应的概率分布确定把当前群体中的第 i 个体 Ind_i 按选择概率 $Ps(Ind_i)$ 抽出，并进行交叉和变异，以提高群体的平均适应度。$Ps(Ind_i)$ 由公式(8)进行计算。

$$Ps(Ind_i)=\frac{Fitness(Ind_i)}{\sum_{j=1}^{G}Fitness(Ind_j)} \tag{11}$$

其中 G 为群体大小，$Fitness(Ind_i)$ 为第 i 号个体的适应度。

遗传交叉算子有多种，可以选择单点交叉和多点交叉，这里我们根据实际情况选择单点交叉。交叉方法为：从当前群体中按轮盘赌法选择两条染色体，随机选取交叉点位置，将两条染色体从交叉点处分为左右两半段，按概率 P_c 依次将两条染色体的右半段互换并重新连接，得到两条新的染色体。

在遗传算法中，变异算子以一个较小的概率 P_m 随机地改变染色体上的某些位串。变异算子使用的几率不大，没有必要对染色体上的每个基因位都考察是否变异，只需随机地考察其中的某些位。变异过程为：随机地确定基因的变异位置，以事先确定的变异概率 P_m 对这些位置的基因进行变异。

与种群初始化方法相似，在进行染色体变异时，我们同样根据 K-means 算法的特点，随机选取特征词的分布概率作为变异点基因值。

4.6 算法停止标准

在实际系统中，我们采取如下停止标准：

(1)固定最大遗传代数 $GNUM$。当算法进行 $GNUM$ 代遗传后停止。最大遗传代数 $GNUM$ 依赖于模型复杂度；

(2)根据遗传收敛的程度，我们给出另一种停止标准，即群体的平均适应度连续多代遗传后仍无明显变化，遗传算法停止。

5 实验结果与分析

针对本文提出的基于混合遗传算法的文本特征词聚类问题，我们进行了两项实验。实验中各参数：种群规模 $G=100$，最大进化代数 $GNUM=100$ 代，交叉概率 $P_c=0.86$，变异概率 $P_m=0.02$，精英个体数 $Elite=4$，聚类数 $K=10$。

实验一(算法稳定性比较)：分别采用 K-means 聚类算法、小生境遗传算法 NGA、小生境混合遗传算法 NHGA(K-means+NGA)对 10 篇测试文档中 635 个特征词进行聚类，每种算法测试 50 次。测试结果见表 1。

从实验结果可知 K-means 算法是一种快速的聚类算法，但算法稳定性较差，这正是 K-means 算法对初始聚类中心依赖性的体现；NGA 算法在聚类稳定性上优于 K-means 算法，但是遗传进化代数较大，运算时间相对较长；NHGA 是一种高效而稳定的遗传算法，这正是由于 NHGA 算法结合 NGA 算法的稳定性和 K-means 算法的高效性所获得的优势。

表 1 算法稳定性比较

	K-means	NGA	NHGA
平均进化代数/迭代次数	28	39	15
获得最优解的次数	36	44	49

实验二(聚类准确率比较)：我们从国家语委现代汉语语料库的 10 种类型文档中各抽取 3000 篇文档作为训练语料，再从训练语料中每类文档抽取 10 篇作为测试语料。测试语料经过特征抽取后得到 2156 个特征词，使用训练语料对这 2156 个特征词进行统计。将这些统计数据作为聚类的原始数据，分别采用上述三种算法进行文本特征词聚类。聚类结果见表 2。

表 2 聚类准确率比较

	K-means	NGA	NHGA
聚类准确率	74%	79%	91%

从实验结果可以看出，该实验与实验一中的结果较为一致。实验表明 K-means 算法聚类准确度较低，NHGA 则是一种准确度较高的聚类算法。

结论 本文通过给出一种将小生境遗传算法和 K-means 算法相结合的小生境混合遗传算法，克服了传统 K-means 算法对初始聚类中心的依赖，充分发挥了小生境遗传算法的全局优化能力和 K-means 算法的快速局部寻优能力，有效地均衡了算法对聚类空间的探索和开发能力，实验证明该算法是一种高效可行的文本特征词聚类方法。

由于 K-means 算法必须人为确定 K 值，本文所提出的 NHGA 算法同样存在这种问题，下一步我们将进一步分析并使用遗传算法对 K 值的选择进行优化处理。

参 考 文 献

1 姜宁，史忠植. 文本聚类中的贝叶斯后验模型选择方法[J]. 计算机研究与发展，2002，39(5)：580～587

2 Clausi D A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation [J]. Pattern Recognition, 2002, 35: 1959～1972

3 焦翠珍，戴文华. 基于混合并行遗传算法的多目标约束优化技术研究[J]. 沈阳农业大学学报，2006，37(1)：125～127

4 Glover F, Kelley J, Laguna M. Genetic algorithms and tabu search: a hybrids for optimization [J]. Computers & Operations Research, 1995, 22(1): 111～134