第16卷 第9期
2006年9月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.9
Sep. 2006

基于改进向量空间模型的话题识别与跟踪

宋 丹[1],王卫东[2],陈 英[2]
(1.大连理工大学 计算机科学与工程系,辽宁 大连 116024;
2.东北电力大学 计算机系,吉林 吉林 132012)

摘 要:话题识别与跟踪旨在发展一系列基于事件的信息组织技术,通过监测以实现对新闻媒体信息流中新话题的自动识别和已知话题的动态跟踪。文中提供一种利用改进的向量空间模型进行识别和跟踪的方法。没有使用传统向量空间模型中单个向量,而是按照语义将特征词划分为4个组(人物、时间、地点、内容)并形成4个向量空间。每个空间进行独立的权重计算和相似度计算。实验证明这些方法是有效的。

关键词:话题识别与跟踪;向量空间模型;时间表达
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2006)09-0062-03

Topic Detection and Tracking with a Developed Vector Space Model

SONG Dan[1], WANG Wei-dong[2], CHEN Ying[2]
(1.Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024, China;
2. Department of Computer Science, Northeast Dianli University, Jilin 132012, China)

Abstract: Topic detection and tracking is an event-based information organization task where online new streams are monitored in order to spot new unreported events and link documents with previously detected events. So present an approach that formalizes temporal expressions and augments spatial terms with ontological information and uses this data in the dictation. In addition, instead using a single term vector as document representation, split the terms into four semantic classes and process, including character, time, space and content, and weigh the classes separately. The approach is motivated by experiment.

Key words: topic detection and tracking; vector space model; temporal expressions

0 引 言

话题识别与跟踪(TDT)作为信息检索的一个新的研究方向,旨在发展一系列基于事件的信息组织技术,通过监测以实现对新闻媒体信息流中新话题的自动识别(first story detection)和对已知话题的动态跟踪(topic tracking, cluster detection)。一个TDT系统的功能与一位信息工作者的工作相似,对于一个新的报道能够将其汇总到已检测到的话题中或者将这篇报道视为一个新的话题。这个识别和跟踪过程很难用传统的信息检索方法实现[1]。

文中提供了一个基于语义分组空间向量模型的TDT方法。该方法中的每个语义组由语义相近的词组成,例如:地点组、时间组、人物组和内容组。将一篇新闻报道用四个独立的空间向量表示(即地点向量、时间向量、人物向量和内容向量)。文中粗略描述利用这个复杂表示方法实现话题识别与跟踪的过程,并与利用单个向量的向量空间模型进行比较。

收稿日期:2005-12-01
作者简介:宋 丹(1980-),女,辽宁锦州人,硕士研究生,研究方向为话题识别与跟踪。

1 前人的工作

话题识别与跟踪的基本思想源于1996年,当时美国国防高级研究计划委员会(DARPA)提出需要一种能自动确定新闻信息流中话题结构的技术[2],这一方向的确立与发展是在话题识别与跟踪(TDT)系列评测会议的推动下进行的。话题识别与跟踪更强调对新信息的发现能力,关心涉及特定话题而且相对广泛的主题类别的信息。近年来,对动态的话题识别和跟踪开始变为热点。

话题识别可以看作是一种按事件的聚类,研究者常采用的算法有:增量k-means聚类、agglomerative聚类、单遍聚类等[3]。有多种不同方法在话题跟踪中被尝试使用,如Rocchio分类方法、决策树方法、基于HMM的语言模型等等[3]。目前常用的话题/报道模型有:语言模型(LM)和向量空间模型,其中向量空间模型是目前最简便高效的文本表示模型之一[4]。

2 改进的向量空间模型

一篇新闻报道最少应该有:什么时间?在什么地点?什么人?发生了什么事?以前的话题识别与跟踪方法都试图将这些方面压缩为一个向量进行表示。而在文中提供的这种改进的向量空间模型中,给上述的4个问题各自

分配一个语义组,进而形成 4 个向量空间,如图 1 所示。

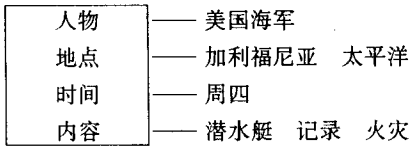

图 1　改进的向量空间模型

一个改进的向量空间模型的例子:"周四,美国海军的潜水艇在加利福尼亚附近的太平洋海域,挑战世界潜水艇潜水最深纪录时起火"

3　相似度计算

通过利用语义分组的表示方法,可以分别对每组进行相似度计算。这样文中的相似度的比较方法就与传统的基于单个空间向量的相似度比较方法有所不同。在这节中,首先给出针对四个向量均适用的通用权值计算方法,接着分别给出时间和地点的相似度计算方法,最后给出笔者提出的的 TDT 算法。

3.1　通用权重及相似度计算方法

典型的新闻报道不同于侦探小说,新闻报道一般会在前几句中讲出所报道的事件。因而,利用特征词出现的等级(即这个特征词所在的句子的次序),来衡量这个特征词的重要性。一个特征词 t 在报道中出现 m 次,则 t 的等级得分计算公式为:

$$rs(t) = \sum_{k=1}^{m} \frac{1}{2^{\ln t_k}} \tag{1}$$

t_k 是特征词 t 的第 k 次出现所处的等级。在等级计分中,一篇报道中的第一个句子的所有词的得分为:$\frac{1}{2^{\ln 1}} = \frac{1}{1} = 1$。

为了确定两个文档之间的交集部分的权重,计算了两个报道的交集部分权重与两个文档的权重之和的比。当然每个特征词自身的有价值性是不同的,因而遵照传统的 IR 的做法,将等级分数乘以倒置文档频率 IDF(inversed document frequency)[5]。例如,$\boldsymbol{X}$,$\boldsymbol{Y}$ 为两个特征词集,则权重的相似度(RWS)为:

$$\mathrm{RWS}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{k=1}^{|\boldsymbol{X}\cap\boldsymbol{Y}|} rs(t_k) * \mathrm{IDF}(t_k)}{\sum_{j=1}^{|\boldsymbol{X}|} rs(t_j) + \sum_{l=1}^{|\boldsymbol{Y}|} rs(t_l)} \tag{2}$$

因此,如果两篇报道是完全相同的,则 $\mathrm{RWS}(\boldsymbol{X},\boldsymbol{Y}) = 1$。如果 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 没有任何相似之处,则 $\mathrm{RWS}(\boldsymbol{X},\boldsymbol{Y}) = 0$。利用公式(2)计算人物向量和内容向量的相似度。

3.2　时间的相似性计算方法

如果在两篇报道中都有"上周一",并不能说明它们之间相似,因为"上周一"是一个相对的概念,是随说话的时间的不同而改变的。为时间表达构造了一个自动转换器,将没有意义的时间表达映射到日历上将其标准化。把标准化后的时间信息用一个全局时间轴上的点来表示。

两篇报道的时间相似度是用时间点对点的匹配来衡量的,考察时间段对应的开始点和结束点。如图 2 所示,对角线上的点代表了时间轴上的两个同步点,阴影区域代表重叠的间隔,它们重合的越多则它们的相似度越大。

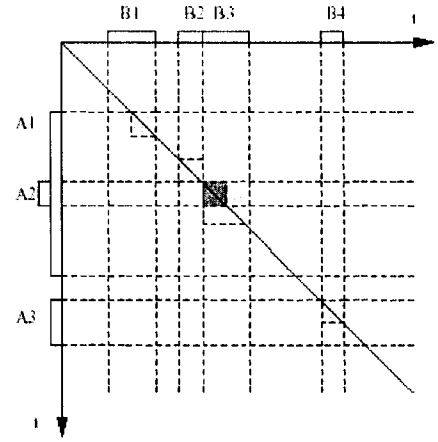

图 2　表示 A 和 B 两个时间段集合的坐标

用 $u_t: T \times T \to IR$ 来计算两个时间段的相似度:

$$u_t([t_i,t_j],[t_k,t_l]) = \frac{2\Delta([t_i,t_j] \cap [t_k,t_l])}{\Delta(t_i,t_j) + \Delta(t_k,t_l)} \tag{3}$$

在实际中,使用"覆盖矩阵"来计算函数 u_t 的值。$T_{i,j}$ 代表 T_i 中的第 j 个时间段。$T_{i,j}$ 的"覆盖"是求 i 行或 j 列中 u 的最大值(用 $v_{i,j}$ 来表示)。整体的"覆盖"用 $v_{i,j}$ 之和与时间间隔数的比来表示。如两个间隔 T_1 与 T_2,T_1 包含 n 个间隔,T_2 包含 m 个间隔,则计算公式为:

$$\mathrm{cover}_t(T_1,T_2) = \frac{\sum_{i=1}^{n} v_{1,i} + \sum_{j=1}^{m} v_{2,j}}{n+m} \tag{4}$$

还要考虑时间特征词的等级得分相似性(见公式(2)),因而两篇报道的时间相似度最终计算公式为:

$$\mathrm{sim}_t(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{cover}_t(\boldsymbol{X}_t,\boldsymbol{Y}_t) * \mathrm{RWS}'(\boldsymbol{X}_t,\boldsymbol{Y}_t) \tag{5}$$

$\boldsymbol{X}_t$ 和 $\boldsymbol{Y}_t$ 为报道 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 各自的时间向量,$\mathrm{RWS}'(\boldsymbol{X}_t,\boldsymbol{Y}_t)$ 是不乘 IDF 的 $\mathrm{RWS}(\boldsymbol{X}_t,\boldsymbol{Y}_t)$。

3.3　地点相似度的计算

文中利用地理知识来计算地点的相似性,这样会比单纯地比较两个地理名词正确得多。例如:当报道安徽的洪水时,特征词:安徽,长江和芜湖市,在表面上看它们没有任何的共同点,但它们在地理上是邻近的,依据地理知识就可以明白它们是相关的。一个简单的例子:把大连市扩充为辽宁省大连市。

图 3 表示了一个简单的地理树。树中每个结点代表一个地点,如果要比较两个地点的相似度,只需知道它们的共同路径与总路径的长度比,因而地点 l_1 和 l_2 相似度 u_s 为:

$$u_s(l_1,l_2) = \frac{(\mathrm{level}(l_1 \cap l_2))}{(\mathrm{level}(l_1) + \mathrm{level}(l_2))} \tag{6}$$

如果两个地点完全相同,则 $u_s(l_1,l_2) = 1$。现在比较法国和德国,结果为 1/(2 + 2) = 1/4,因为共同的路径长度是 1,而它们的各自路径均为 2。同样,中国和法国的结果是 0/(2 + 3) = 0。巴黎和法国的相似度为 2/(2 + 3) = 2/5。

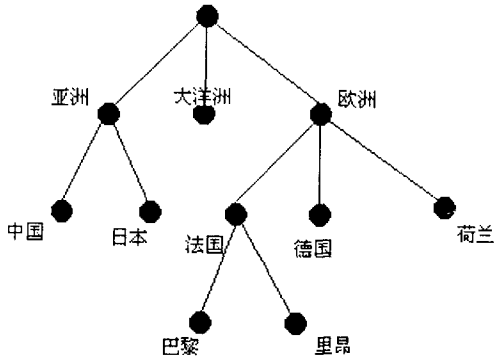

图3 一棵简单的地理树

因为一个地点特征词要与另一则报道的所有地点特征词进行相似度比较,所以使用3.2节中的"覆盖矩阵",对每一行或列只取最大的。两则报道的地点重合性与时间覆盖性的计算公式类似,如下:

$$\mathrm{cover}_s(L_1,L_2)=\frac{\sum_{i=1}^{n}v_{1,i}+\sum_{j=1}^{m}v_{2,j}}{n+m} \tag{7}$$

类似公式(5),使用带IDF的RWS,则报道 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的地点相似度的计算公式为:

$$\mathrm{sim}_s(\boldsymbol{X},\boldsymbol{Y})=\mathrm{cover}_s(\boldsymbol{X}_s,\boldsymbol{Y}_s)*\mathrm{RWS}(\boldsymbol{X}_s,\boldsymbol{Y}_s) \tag{8}$$

3.4 TDT算法

文中提出把4个语义向量的相似度加权求和来比较亲疏关系。如果加权和大于 θ 则认为它属于这个话题,否则就被认为是新报道,并被添加到事件队列。公式(9)中,β_c 反映的是语义组 C 的重要性。

$$\mathrm{sum}=\sum_{c\in C}\beta_c*\mathrm{sim}_c(v_c,e_c) \tag{9}$$

另外,也考虑到当某个向量相似度为0时,要做相应的扣分。如:不能仅仅因为时间、地点和人物有较高相似而确定报道的相似,所以要因为内容向量的相似度为0而在加权求和之后再减分。

文中采用的这种启发式的聚类算法如图4所示。

```
1   found←();
2   for each new document d
3   v←Build-Vector(d);
4   max←0;event←();
5   for each found e
6   dist,←();
7   for each semantic class c
8   add(sim(Vc,Ec),dist);
9   end;
10  if(sum>max)
11  then max←sum;
12  envent←e;
13  f1;
14  end;
15  if(max>θ)
16  then
17  else add(v,found);
18  f1;
19  end;
```

图4 TDT算法流程图

4 实验及其结果

笔者从网上的2004年的4月1日到2004年的12月1日的10000多篇新闻报道中手工挑出5807篇。训练集包括1918篇79个事件的报道,测试集包括3909篇85个事件的报道。利用Connexor[4]的名称实体识别器提取出地点和人物。把话题识别和跟踪系统的性能用准确率、召回率及它们两者的联合 $F1$-measure表示。评估的方法遵循下面的公式:

$$\text{准确率 } P=\frac{\text{系统识别出的相关报道数}}{\text{系统找出的所有报道的总数}}$$

$$\text{召回率 } R=\frac{\text{系统识别出的相关报道数}}{\text{所有相关的报道总数}}$$

$$F1\text{-measure}=F1=\frac{2PR}{P+R}$$

用文中提供的启发式阈值聚类方法进行实验。而且为了能够做一个比较,用余弦系数[5,6]的方法做了另两次实验,分别是用单向量和文中提供的四向量模型。在选择相同阈值的前提下。实验结果如表1所示。

表1 识别和跟踪结果

方法	识别 P	识别 R	识别 Fd	跟踪 P	跟踪 R	跟踪 Ft	$(Fd+Ft)/2$
余弦系数	0.562	0.346	0.428	0.315	0.704	0.435	0.431
余弦系数(四向量)	0.672	0.401	0.514	0.384	0.672	0.489	0.502
启发式聚类	0.684	0.893	0.775	0.692	0.543	0.609	0.692

从表1中不难看出文中提供的这种启发式方法效果比其他两个理想。

5 结束语

话题识别与跟踪(TDT)作为信息检索的一个新的研究方向,到目前为止,话题识别与跟踪领域的大部分研究都是借用信息检索的某些方法,只是通过调整某些参数来使这些方法更适合于处理话题(事件)。但由于话题识别与跟踪研究的某些特殊性,决定了仅仅利用现有信息检索方法来进一步提高TDT系统的性能是很困难的,要想突破必须要借助更多的自然语言理解技术。

文中提供了一种借助自然语言理解技术,对特征词进行语义分组,形成具有4个独立向量空间的改进的向量空间模型,并利用这个模型采用一种启发式的聚类方法进行识别和跟踪。

参考文献:

[1] Yang Y C, Donddington J, Peerce R, et al. X Learning approaches for detecting and tracking news events[J]. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999(14):32-43.

[2] Carbonell A J, Doddington J, Yamron G, et al. Topic detection and tracking pilot study final report[A]. In: proc DARPA Broadcast News Transcription and Understanding Workshop [C]. San Francisco, CA: Morgan Kanfmann Publishers Inc, 1998: 194-218.

(下转第67页)

件的模糊量表示了现象的发生程度,其等级划分为:一定、极可能、强可能、可能、一般、弱一般、极弱、无。每一个等级都对应一个[0,1]内的一个取值区间,如表1所示。

表1　模糊系数对应的量词表

模糊量词	取值区间
一定	[1.00,1.00]
极可能	[0.95,0.99]
强可能	[0.80,0.94]
可能	[0.65,0.79]
一般	[0.45,0.64]
弱一般	[0.30,0.44]
极弱	[0.10,0.29]
无	[0.00,0.00]

②权系数。

在某一规则中,前件是关于 p_i 的集合,对 p_i 在此集合中的量度用 w_i 的大小来表示,也表示了其在规则中的贡献度。所以,可以根据 w_i 的大小来判断前件(现象)集合中前件的重要程度,同时又要要求这些 w_i 的和应为整数1,因此根据经验和实际情况可以将其划分为以下几个等级:极重要、很重要、重要、一般、不重要、可有可无。每一等级的量度根据系统需要进行取值,也可以动态地在不同类型的故障规则中分配值,同时这些值还可以根据将来系统的运行情况进行机器学习,调整它们的量度。默认情况下,使用的量度如表2所示。

表2　默认权系数量度表

等级	权值量度
极重要	6
很重要	5
重要	4
一般	3
不重要	2
可有可无	1

4　故障知识表达实例

按照上述的知识表达方法,下面有一个具体的实例。

在电脑故障中有一个很典型的例子——死机。导致死机的原因有两大类,即硬件原因和软件原因。比如硬件资源冲突、设备不匹配、病毒感染、系统文件丢失等等,当然这些原因不仅可以导致死机,还可以导致其他的故障。在这里按照文中所述的方法,可以表述如下规则(如表3所示)。

表2中列出了电脑死机时的知识表示规则(未列出规则中详细结论),根据当前的模糊量词和权值量度电脑死机极有可能是显卡故障引起的,如显卡松动、显卡时间较长沾满灰尘、显卡金手指部分变脏或脱落等。当然,如果表中的模糊量词和权值量度取值不同可能会得到其他原因引起的死机。所以,按照这种方式就可以比较清楚合理地表达故障知识,将模糊量词和权值量度转换成相应的数值,即可以使用推理机进行诊断。

表3　电脑死机规则表

IF	模糊量词	权值量度	
病毒感染	极弱	不重要	
提示系统文件丢失	无	可有可无	
开机无显	可能	很重要	
花屏	强可能	很重要	
颜色不正常	可能	很重要	
屏幕上有乱码	可能	重要	
超频	无	可有可无	
CPU风扇转动不畅	极弱	不重要	
机内发出"嘟嘟"报警声	一般	不重要	
Then	软件故障引起死机(详细)		
	CPU故障引起死机(详细)		
	显卡故障引起死机(详细)		√
	内存故障引起死机(详细)		

5　结束语

文中分析了电脑故障的分类和特点,提出了电脑故障的抽象知识模型,并且由此给出了基于模糊产生式规则的知识表达方法来表达其领域知识,同时还讨论了其知识规则的不确定性,给出了具体的表达实例,为下一步采用基于模糊匹配和语义距离结合的推理的电脑故障诊断系统打下基础,这也是笔者正在研究的工作。

参考文献:

[1] 吕家国,李桂玲.计算机及网络故障诊断与维护[M].北京:科学出版社,2004.

[2] 新时代工作室.电脑故障排除1000例[M].北京:机械工业出版社,2005.

[3] 蔡自兴,徐光祐.人工智能及其应用(第3版)[M].北京:清华大学出版社,2003.

[4] Luca C, Roberto R. Diagnosis of multiple faults with flow-based functional models: The functional diagnosis with efforts and flows approach[J]. Reliability Engineering & System Safty, 1999, 64(2): 137-150.

[5] 刘　铭,时　昕,姚燕南.基于数据库的电力设备故障诊断模糊专家系统的设计与实现[J].计算机工程,2001(3):75-77.

[6] 刘　旭.一种诊断性专家系统的设计与实现[J].计算机工程,2001(2):90-92.

(上接第64页)

[3] 李保利,俞士汶.话题识别与跟踪研究[J].计算机工程与应用,2003,39(17):6-10.

[4] 林鸿飞,高　天,姚天顺.中文文本的可视化表示[J].东北大学学报,2000,21(5):501-503.

[5] 李晓明,阎宏飞,王继民.搜索引擎[M].北京:科学出版社,2005.

[6] 金　珠,林鸿飞.基于HowNet的话题跟踪及倾向性分析[J].情报学报,2005,24(5):10-22.