doi:10.3969/j.issn.1000-436x.2013.08.001

基于依存连接权 VSM 的子话题检测与跟踪方法

周学广，高飞，孙艳
(海军工程大学 信息安全系，湖北 武汉 430033)

摘 要：针对在新闻话题中报道突发、热点相似且子话题层次丰富的现象，依据增量 TF-IDF 值构造特征维，生成全局向量；然后在时间窗内生成特征连接权的局部邻接图，利用依存句法进行分析降维；最后采用领域词典加权，时间阈值衰减；从而构造出利用依存连接权 VSM 进行关联分析的子话题检测与跟踪（sTDT）计算方法。实验表明，利用依存关联分析使文本表示由线性变为平面结构，能够有效地提取描述子话题；在人工标注的测试语料下，其最小 DET 代价比经典方法至少降低 2.2%。

关键词：话题检测与跟踪；依存连接权；关联词对；报道关系检测；向量空间模型
中图分类号：TP391 **文献标识码**：A **文章编号**：1000-436X(2013)08-0001-09

Sub-topic detection and tracking based on dependency connection weights for vector space model

ZHOU Xue-guang, GAO Fei, SUN Yan
(Department of Information Security, Navy University of Engineering, Wuhan 430033, China)

Abstract: Aiming at the phenomenon that there are abrupt reports, similar topics and abundant levels of subtopics in the news, a novel method based on relationship analysis using dependent sentence pattern was proposed for sub-topic detection and tracking (sTDT), which constructed feature dimensions to generate the global vectors according to the increment of TF-IDF, and then created the partial adjoin map based on the connection weights within the time window and decreased the dimensions through dependent sentence pattern. Finally, a novel method for sTDT computing was built with adjoins dictionary weights and time threshold attenuation. Experiments show that the proposed method transferrs the text from linear to plane structure, and extracts the subtopics effectively, of which the minimum DET cost is reduced by at least 2.2 percent than that of classical methods.

Key words: topic detection and tracking; dependency connection weights; associating words group; report relation detection; vector space model

1 引言

话题检测与跟踪（TDT, topic detection and tracking）是为了解决信息时代数据过载问题而提出的技术，其核心是在连续的数据流中检测知识并分析和预测数据趋势。文献[1]综述了 TDT 的基本研究工作，主要包括话题检测（TD, topic detection）以及话题跟踪（TT, topic tracking）两方面的内容。前者主要检测和组织未知的话题模型，后者是在 TD 的基础上或根据相关的先验知识对已知话题进行追踪和分析。

由于互联网的无监督特性导致网络环境中自由信息泛滥，通过 TD 技术，可以检测出蓄意制造不良报道集的源头；然后采用 TT 技术对网络舆论的内容进行分析，将具有不良性质的报道聚集形成话题，跟踪和分析此类话题的动向，一旦该话题的发展超过预期警戒阈值，就可以督促有关人员采取行动加以约束。因此，TDT 技术在信息安全、行业调研、军事信息安全等诸多领域应用广泛。

自 1996 年以来，在 TDT 系列测评会议的推动下，来自各方面的学者把 TDT 任务分为面向新闻报道切分、面向已知话题跟踪、面向未知话题检测、

收稿日期：2012-05-15；**修回日期**：2012-11-19
基金项目：海军工程大学科学研究基金资助项目(HGDYDJJ10008)
Foundation Item: The National Natural Science Foundation of Naval University of Engineering (HGDYDJJ10008)

报道相关性检测以及首报道检测 5 个子任务。TDT 采用最广泛的概率模型之一是向量空间模型(VSM)。作为最早使用自然语言处理技术解决 TDT 问题的学者之一，Allan James 等人采用 VSM 对话题和报道进行描述，并赋予命名实体（NE, named entities）更高的权重，以解决 TDT 中的事件检测任务[2,3]。PONTE 等人[4]采用 VSM 模型和特征上下文扩展技术执行关联检测任务，不仅有助于解决 VSM 模型存在的数据稀疏问题，还能削弱特征的歧义性。Tu 等人提出一套新话题检测的指标体系，从新颖性和时效性角度对话题检测与跟踪进行了研究[5]。ZENG 等人采用隐马尔科夫模型(HMM, hidden Markov model)来表示文本，以克服复杂度随着词汇增加而增大问题[6,7]。ZHAO 等人使用聚类算法检测社区网络话题，通过引入链接分析提高检测准确率[8]。

上述研究工作把重点放在话题的检测上，没有对话题的层次结构进行精确划分，基本不能精确地反映话题的层次以及趋势。

为解决上述问题，本文从以下 2 个方面开展研究，一是改进向量空间模型，引入词语特征之间的连接权值，通过使用依存树分析构造有向节点，在外部引入领域内的命名实体词典并放大相应权值，从而构建依存连接权 VSM 模型(DCW_VSM, dependency connection weights for VSM)；二是借鉴报道聚类——报道型检测体系的高效计算方法计算 DCW_VSM，构建子话题检测与跟踪(sTDT)方法。最后通过相关实验验证本文提出方法的有效性。

2 基于依存连接权的 VSM 模型

2.1 领域词典和连接权计算

报道文档需要相应的文档表示模型，通常使用语言模型和 VSM，前者利用词元的组成概率进行类别判断，后者将文本表示成向量形式计算类别的相似度。由于一个话题中的新闻报道通常具有高相似性，因此，需要寻求更细微的表示模型。

根据相关的研究成果，在 TDT 中引入命名实体能够显著地提高检测的效果，但在中文处理环境中，单字之间没有显著的分隔关系，中文文档的分词以及人名地名识别问题都是待解决的难题，一些领域内专业命名实体的识别效率始终不高。本文引入领域内的命名实体专业词典，目的是提高报道模型表示的准度；通过建立人工命名实体专业词典，有效区别用户所期望关注的兴趣点。

Florian Holz 以单词“abu ghraib”为例，检测随着话题的演变其表述的内涵也在发生变化[9]。受其启发，本文在话题的演变与中心漂移的过程中，引入连接权变化参数 CW(connection weights)，CW 的计算方法如下

$$CW_i(t_a,t_b)=\eta_i(t_a,t_b)C_0(t_a,t_b)$$
$$=\frac{1}{K_N}(\mathrm{rank}(C_i(t_a))-\mathrm{rank}(C_i(t_b)))C_0(t_a-t_b) \quad (1)$$

式(1)的相关参数参见图 1 中算法描述。

CW 算法如图 1 所示。

输入：语料 C

输出：话题连接权变化参数 CW

1) 以天为单位分割时间窗（TimeWindow），按照时间信息分割语料 C，得到各个时间切片的语料 C_i。

2) 对于每个 C_i，使用 TD-IDF 方法计算每个特征词 t 在此时的重要度指标 $C_i(t)$。

3) 按照 t 的大小排序得到特征词在时刻 i 的序列，截取其中前 K_N 项特征，并记录下特征 t 在此时序列中的位置，记为 $\mathrm{rank}(C_i(t))$。

4) 对于截取出的任意 2 个特征 t_a 和 t_b($0\leqslant a<N, 0\leqslant b<N$)，设置一个文本距离窗 TextWindow，记录下 t_a 和 t_b 在此范围内出现的共现频数为 $C_0(t_a, t_b)$。

5) 参照式(1)计算此时特征词 t_a 和 t_b 的连接权变化参数 $CW_i(t_a,t_b)$

（通常情况下，TextWindow 以文本一句话结束为标志）

图 1 连接权变化参数 CW 计算算法

2.2 依存分析

在计算连接权的共现频数 $C_0(t_a, t_b)$时，可能会发生词语组合爆炸现象，因此，需要对其进行语法分析。语法分析通常使用短语结构和依存分析等方法，短语结构方法更多依赖于人工经验得出的规则集合，具有一定的局限性。依存分析方法是法国语言学家 L.Tesniere 在文献[10]中首先提出的，核心是用连接表示词与词之间的关系，这种关系表现为一个词有向地支配或者受支配于一个词，这种支配关系是和语义相关联的，非常适合关联语义修正。因此，本文采用依存文法分析对词语之间的连接权以及同现关系进行修正，通过 ctbparser 开源工具包获取词语之间的依存关系[11]。

依存分析实际上是一种统计分析方法，根据句法树库的不同，存在着不同的标注体系。其中，中文宾州树(CTB, penn Chinese treebank)是出现较早且受人们研究较多的中文结构语法库。在进行依存分析连接权修正时，利用 CTB 统计得到的结构句法信息对话题文本的来源进行结构分析，选取有效搭配对进行连接权修正以提取主要关联。有效搭配

对指的是在一定的层次距离 D 之间有依存关系的有效词组成的搭配词对[12]，这里的有效词定义为动词、名词和形容词。

利用依存分析可以将扁平的句子结构转化为带有层次的树形结构，如图 2 所示。图中粗线连接的两词可以被看作句子的主要关联内容。经过依存分析之后，出现一个不具有父节点的中心词，例如图 2 中的“话题”，相当于依存树的根节点。

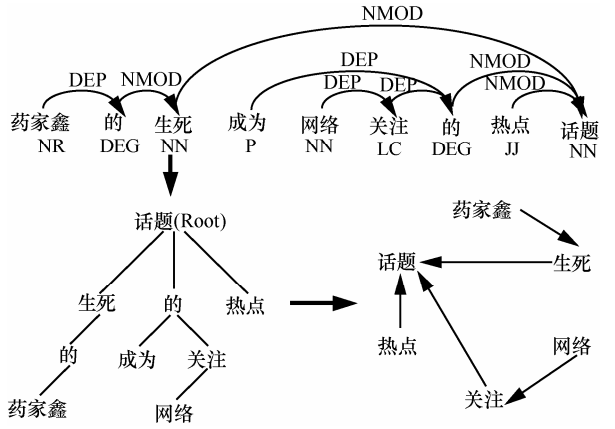


图 2　依存关联分析

对于新闻报道进行依存关联分析的重点是依存树有效搭配词对提取，即提取一个三元组(W_i, W_j, d_{ij})，其中，d_{ij} 表示词对之间的依存强度。对于依存分析生成的依存句法树，经过去冗余的剪枝、合并以及嫁接操作，以实现对有效关联特征词对的提取。

2.2.1　剪枝操作

为了充分挖掘句子的主干，或删除量词、助词及介词标签，可以通过剪枝操作完成。

在词语节点中剪枝操作包括两方面内容：忽略分词节点剪枝与删除分词节点剪枝。前者对于剪枝操作的作用域仅仅局限于当前节点中，其子节点并不发生作用；后者则对当前节点以及所有子节点进行裁剪。

如图 3 所示，左侧表示对于节点 B 的忽略操作，即直接将其子节点 D 依靠在父节点 A 上，当前节点不加入关联词对的生成分析，该项操作通常应用于助词及介词标签如 DEC、DEG、P 等；右侧表示对节点 F 的删除操作，即直接去除所有的孩子节点 G。在依存树中，删除剪枝操作常用于量词与程度词标签如 DT、M 等。

2.2.2　合并操作

为了能够较好地表达原句含义的中心词纳入特征空间，需要进行合并操作。

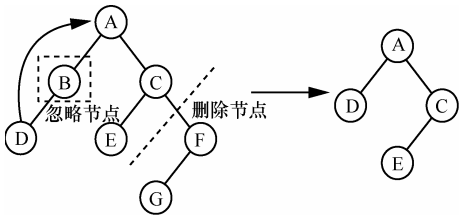


图 3　依存树的剪枝操作

依存树的压缩合并是将处于不同层级的子节点归并到同一父节点中，即孙子节点直接作为孩子节点与父节点产生词语关联关系。压缩合并操作与层级相关，可以用 $N(n_s, n_e)$表示，其中，N 表示待压缩合并的父节点，n_s 为归并操作起始层级，n_e 为归并操作终止层级，显然 $n_e-n_s\geq 2$。

图 4 中的合并操作可以表示为 C(1,3)，压缩之后将节点 G 并入节点 C 中。对依存树进行归并操作改变了原始的依赖关系，但是在话题跟踪应用中，对依赖关联的要求并不苛刻。

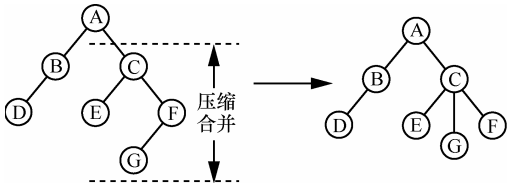


图 4　依存树的压缩合并操作

2.2.3　嫁接操作

通过移动嫁接操作可以检测并列的名词主体。嫁接操作包括移动嫁接操作以及复制嫁接操作，按照作用域的不同分为节点嫁接操作以及子树嫁接操作。复制嫁接操作中，对原节点或子树不产生影响，嫁接过程通过拷贝节点进行操作；在移动嫁接中，原节点或子树会发生变化，相当于在复制嫁接基础上对原节点进行剪枝处理。

图 5 对节点 D 进行子树嫁接操作，对节点 F 进行节点嫁接操作，左右两侧的依存树分别为复制与移动嫁接得到的结果。在量词节点、时间词节点或并列关系的标签和具有被动句法的标签如 CC、BA 中，可以利用嫁接操作完成检测并列名词主体工作。

2.3　依存连接权 VSM 模型的文档表示方法

向量空间模型 VSM 中常使用词袋法(BOW)，将报道文档进行特征向量表示并利用停用词表过滤。该方法具有计算效率高、实现简单便捷的特点，适合对文档总体信息的表示。本文在 VSM 的基础上，引入词语特征之间的连接搭配关系 CW，建立一种新型依存连接权模型 DCW_VSM。

在 DCW_VSM 模型中，将文本的表示分为本体向量(OV, origined vector)与子体连接图(SCG, sub-connection graph)2 个部分。

本体向量 OV 采用词袋法进行表示，使用增量的 TF-IDF 方法进行过滤，用以描述文本总体特征，在 OV 降维时，使用经过时间片合并的完整文本库 C，从而保证文档有相同的特征维，可见本体向量是一个全局向量，该全局向量可根据应用按照一定的时间策略更新（一季度或半年更新一次）。

子体连接图（SCG）定义为一个特征节点交互的无向图，在进行节点选择的过程中使用 TF-IDF 方法，降维时采用时间窗文本库 C_i，并计算节点之间的连接权变化参数 CW（见 2.1 节）。可见 SCG 选择的节点向量为一个局部向量，节点特征维随时间而改变。此外，为了突出命名实体特征在子体向量图中的作用，对命名实体节点的 CW 进行加权处理，即 $CW_{ne}=CFIT\_ne\times CW_{ne}$(其中，$CFIT\_ne$ 是对命名特征的加权系数，本文实验中设为 3)，整体的文档表示模型 DCW_VSM 如图 6 所示。

图 6 中左侧虚线框表示输入的文档，中间为处理过程，最右侧的虚线框即通过 DCW_VSM 表示的文档逻辑视图（documentation logical view），每个特征词在图中用一个小方框表示，其中灰色方框表示一个命名实体关键词。中间层中的 BOW 与 LOW 分别经过特征提取及依存树分析后降维，得到 OV 和 SCG，使用命名实体词典进行权值修正放大，从而得到最终的文档表示模型。

3　子话题检测与跟踪方法

3.1　话题层次结构

TDT 任务面向的对象主要是新闻信息，并把其作为连续的数据流进行处理，归并相近的信息，提取未知新信息，从而提高知识检测的时间效率。为了精确地反映话题的表述内容，下面给出子话题层次及检测跟踪结构，如图 7 所示。

在话题层与报道层之间搭建新的子话题层，定义子话题为由一个事件引起的且与该事件具有直接因果关系报道的集合。子话题与话题之间的关系通常十

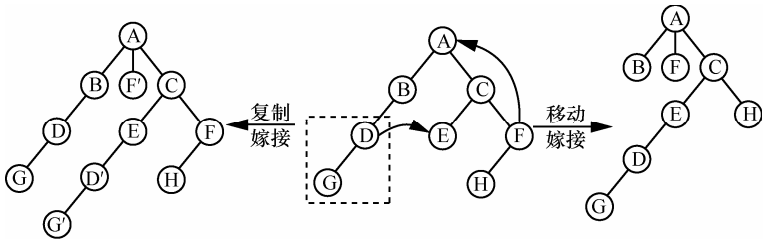


图 5　依存树的嫁接操作

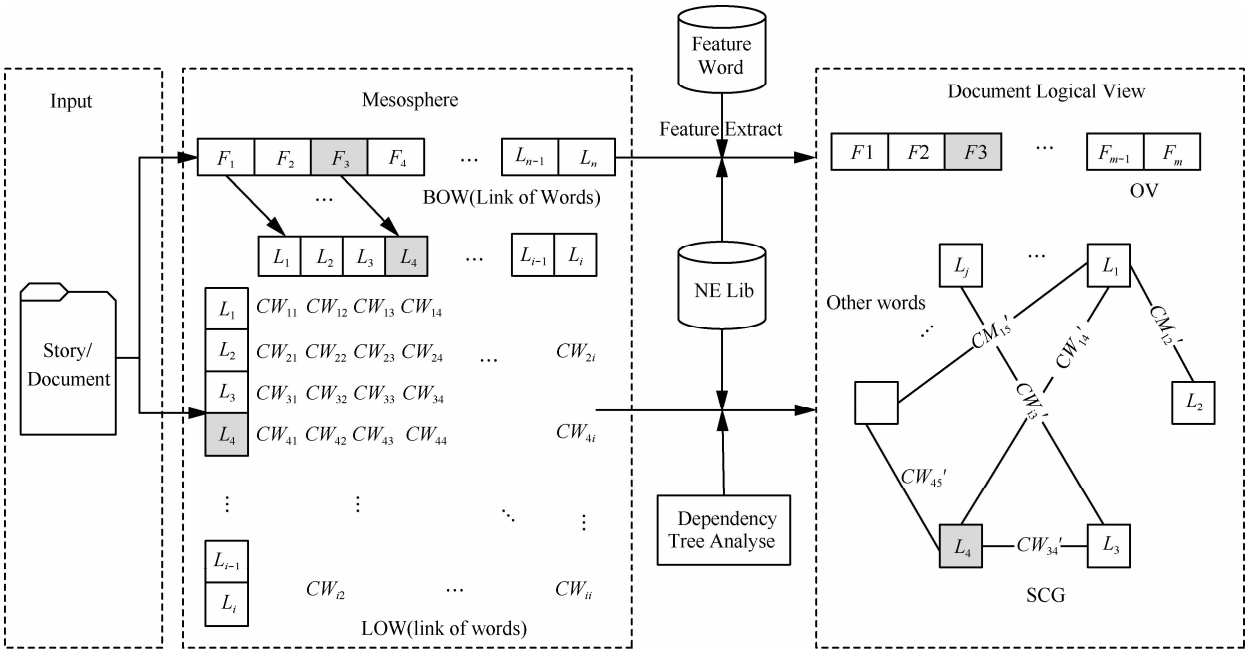


图 6　DCW_VSM 模型的文档表示方法

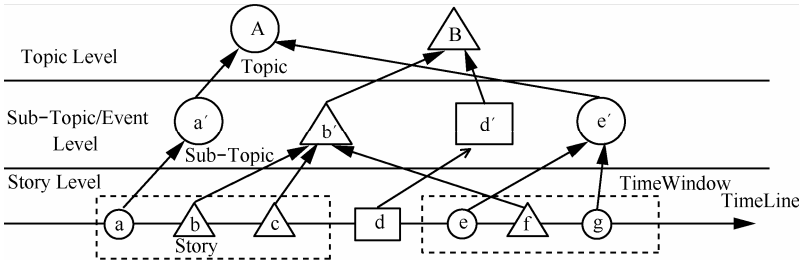


图 7　子话题层次及检测跟踪结构

分微妙：对于两篇报道 x 与 y，如果 x 属于一个种子事件，若 y 的出现与 x 有直接的因果联系，那么 x 与 y 可以并入同一个子话题中；若 y 的出现与 x 没有直接的因果关系，即，若在没有报道 x 的情况下 y 也有可能出现，那么 x 与 y 不能并入同一个子话题，只能作为同一话题层次中的相关报道。例如：在关于“药家鑫”案件的报道中，出现了“重庆药家鑫”、“泰州药家鑫”等相关案件的报道，这些案件的出现没有因果关系，只能属于一个话题而不能属于同一个子话题。另一方面，出现了“律师激情杀人辩护”、“反思社会教育”等报道，这些报道若没有“药家鑫”事件便不会出现，因此把这类报道归并为同一子话题层次中，从而达到精确划分话题层次的目的。

3.2　子话题检测

话题检测体系包括 3 个分支：“报道–报道型(SS)”、“报道–聚类型(SC)”以及融合后的“报道–聚类–报道型(SCS)”[13]，SCS 在不损失准确率的情况下能显著提高检测算法的效率。因此，借鉴 SCS 方法，利用 DCW_VSM 的文档表示模型，构建子话题的检测与跟踪方法。

系统对初始语料进行聚类得到已知子话题模型 Sub-Topic i，对于一篇输入的报道 Story x，将 S_x 分别与子话题模型中的各个报道代表进行相似度计算，令其值分别为 $S_{ik}(x)$，其中，k 表示话题 i 中报道的下标。则报道 x 距子话题 i 的平均相似度为 $\overline{S_i(x)}=\dfrac{\sum S_{ik}(x)}{|k|}$，如果小于一个判别阈值 TRSOD_Dtg，那么为该报道开辟一个新的话题空间并初始化话题模型 Sub-Topic n；否则，比较 S_x 在各个已知话题模型 T_i 中的平均相似度大小，选取平均相似度最大的子话题模型作为 S_x 的最终检测结果，同时调用话题更新策略。

3.3　子体连接图相似度计算

在 SCS 体系中，报道与话题的相似度实质上可以归结为报道之间的相似度，在本文的模型中定义了本体向量和子体连接图，前者用以计算文本的整体相似性，后者用于比较颗粒度更细的子话题。前者直接使用 Cosine 距离公式计算本体相似度，设 v_i, v_j 是两篇报道，$v_i=(w_{i1}, w_{i2}, \cdots, w_{im})$，$v_j=(w_{j1}, w_{j2}, \cdots, w_{jm})$，则本体相似度定义如下

$$SimOV(v_i, v_j)=\frac{\sum_{k=0}^{m} w_{ik}w_{jk}}{\sqrt{\sum_{k=0}^{m} w_{ik}^2}\sqrt{\sum_{k=0}^{m} w_{jk}^2}} \tag{2}$$

而对于后者由于使用了非全局特征，在各个报道中子体连接图的节点存在差异，由此本文采用归一化的方法。

对于两篇报道 S_x 与 S_y，在子体连接图中的节点分别为 N_x 与 N_y，将其分为 3 部分：两篇报道共有的节点 C、x 特有的节点 P_x 以及 y 特有的节点 P_y。令由节点 C 组成的连接图矩阵为 ***AC***，把 ***AC*** 称为下归一矩阵(DNA, downward norm-array)。令由所有节点共同组成的连接图矩阵为 ***ACP***，把 ***ACP*** 称为向上归一矩阵(UNA, upward norm-array)，如图 8 所示。

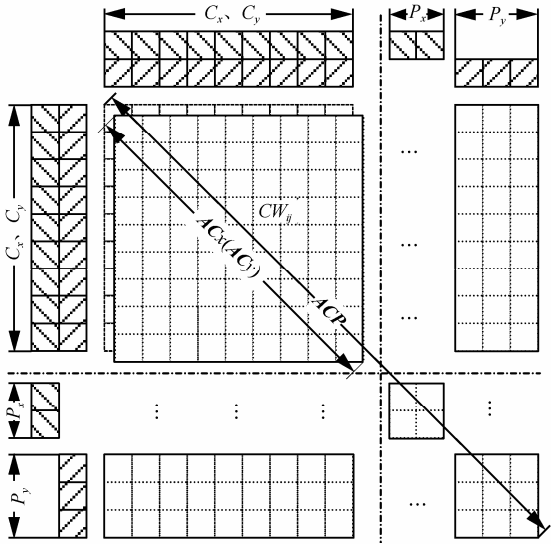


图 8　SCG 相似度计算

图 8 中 C_x 与 C_y 即为两篇报道共有的节点，***AC*** 为下归一矩阵，***ACP*** 为上归一矩阵。观察并分析影响相似度指标的因素有：DNA 与 UNA 的重合程度以及两篇报道在 DNA 的距离。

定义报道 S_x 与 S_y 的下归一矩阵的相似度为 $SimDNA$（S_x，S_y），计算方法如下

$$SimDNA(S_x,S_y)=\frac{\sum_{0}^{|C|}w_x w_y}{\left(\sum_{0}^{|C|}w_x^2\right)^{\frac{1}{2}}+\left(\sum_{0}^{|C|}w_y^2\right)^{\frac{1}{2}}} \tag{3}$$

定义 DNA 与 UNA 的重合程度为一个相似度系数 K，且令 K 的大小同节点相似系数 m 与连接权相似系数 n 相关。

$$K=m\times n=\frac{2|C|}{|Px|+|Py|+2|C|}\times\frac{|AC|}{|ACP|} \tag{4}$$

其中，m 中的向量取模是求向量中的元素数目，n 中对矩阵取模是求矩阵中的元素内容总和。可以看出当 DNA 中的节点所占的比例越大时，节点相似度越大，报道 S_x 与 S_y 的相似度越大，当 DNA 中的连接权值所占的比重越大时，连接权相似度系数越大，报道 S_x 与 S_y 的相似度越大。

由此，可以得到两篇报道 SCG 的子体相似度计算公式

$$SimSCG(S_x,S_y)=K\times SimDNA(S_x,S_y) \tag{5}$$

在相似度计算的过程中本体向量起宏观判别作用，反映报道间的主体差异；子体连接图相似度计算起微观判别作用，反映了子话题的词典命名实体中论述范畴随时间变化的差异，因此，可以得到最终的相似度计算式

$$Sim(S_x,S_y)=\alpha\cdot SimOV(S_x,S_y)+\beta\cdot SimSCG(S_x,S_y) \tag{6}$$

式(6)中由于 K 值的存在，使得 SimSCG 的量级影响较小，因此在设置 β 时适当放大 SimSCG 的影响，文中 α 与 β 的值分别设为 1 与 25。当 α =0 时，相似度公式只考虑连接权的相似程度；当 β =0 时，相似度公式退化为传统的 VSM 模型。

3.4 子话题更新策略

SCS 检测体系中，采用文本代表描述已知子话题模型，在新报道归并到某话题 T_i 时，需要对其中的文本代表进行更新。设定相似度的更新阈值为 $TRSOD\_Fls$($TRSOD\_Fls>TRSOD\_Dtg$)，计算报道 S_x 与话题 T_i 的平均相似度为 $Sim(S_x, T_i)$，当 $Sim(S_x, T_i)>TRSOD\_Fls$ 时将报道加入子话题中，并对文本代表进行更新；否则将报道加入子话题中不进行更新。

在对文本代表进行更新时，首先计算子话题的本体向量中心，使用 Cosine 距离公式计算各个报道到本体向量中心之间的距离，最后选取其中距离最小的前 K_{NUM} 项报道，以此作为子话题模型更新后的文本代表。

4 实验与分析

4.1 实验环境

实验所需的依存分析采用 ctbparser 开源工具包，分词和词性标注采用中文宾州树库标准，在 Visual studio 2010 上采用 C++语言实现。

通常用于话题检测与跟踪的测试语料是 LDC 为 TDT 提供的 5 期语料，分别是 TDT 预研语料、TDT2、TDT3、TDT4 和 TDT5，并从 TDT2 开始支持中文，但是这些语料只对话题进行了人工标注，不能很好地反映子话题的演变趋势。为判断子话题的检测与跟踪算法是否符合人们的主观想法，本文通过互联网收集并下载了 2010 年 10 月 23 日至 2011 年 4 月 19 日有关于“药家鑫”案件的报道，经过去重处理后，包含 621 篇文档共计 287 818 单词。且在收集语料的过程中，人工对报道进行分类，最终形成“药家鑫”案件进展、社会评论以及律师辩护 3 个子话题，试验过程中即以 3 个子话题作为标准测试集对子话题检测与跟踪算法进行测试。

4.2 性能分析

4.2.1 VSM 与 DCW_VSM 性能比较

为了准确检测和跟踪子话题内容，DCW_VSM 必须有更高的灵敏度。图 9 中显示的是 2010 年 11 月 30 日内报道的相似度矩阵，其中，相似度高的两篇文档对应的点呈现出更深的颜色。

图 9(a)反映的是在 VSM 模型下报道的相似度，图 9(b)反映的是在 DCW_VSM 模型下报道的相似度。在下载的语料中，子话题的相似度很高，如图 9(a)中表现出更高的相似度（大部分高达 0.6 左右），而在图 9(b)中相似程度有明显的下降。可以看出，图 9 中的白色虚线代表了相似度的下降情况，即分属于不同子话题的两篇文本由于涵盖了相似的特征词而造成错误的相似值；相反，图 9 中的黑色虚线代表了在关联语义层面上具有真正高相似度的两篇文档。可见本文提出的 DCW_VSM 算法能够较好地增强子话题的检测灵敏度。

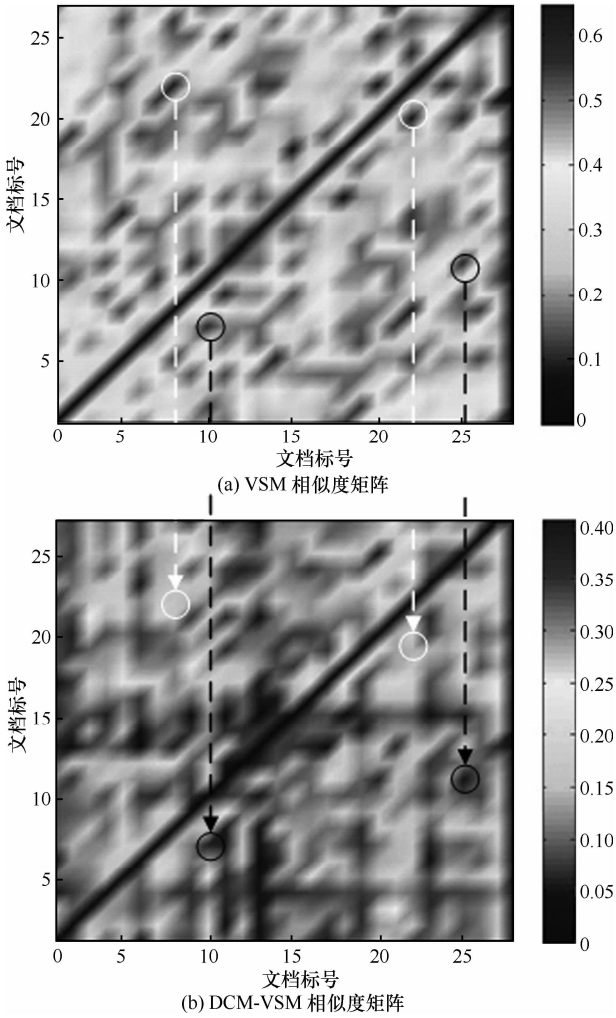


图 9　2 种计算模型的相似度矩阵对比

4.2.2　DCW_VSM 细划子话题分析

在 DCW_VSM 模型中，报道间的相似度是对图 9 的α、β的加权叠加，通过 Single Pass 聚类方法，调整 TRSOD_Dtg，使本文算法产生 3 个子话题如图 10 所示，图中横轴表示的是自 2010 年 10 月 23 日开始的天数，纵轴表示某日产生报道数的对数值，图中黑色虚线表示话题语料包含的报道数随时间的变化趋势，实线分别表示子话题随时间的变化趋势。

可以看出，律师辩护子话题只出现在话题的后期，与原话题有显著区别，反映了原话题中一个时间段的内容，而社会评论和案件进展子话题随原话题的变化而起伏出现。人工分析各子话题的报道内容，律师辩护子话题主要包含的是“激情杀人”的文本，属于律师辩护子话题，社会评论子话题主要包含的是“教育”的文本，案件进展子话题主要包含的是“杀人”的文本，分别可归入社会评论及案件进展子话题中，可见本文算法能够较好地挖掘出子话题的相关内容。

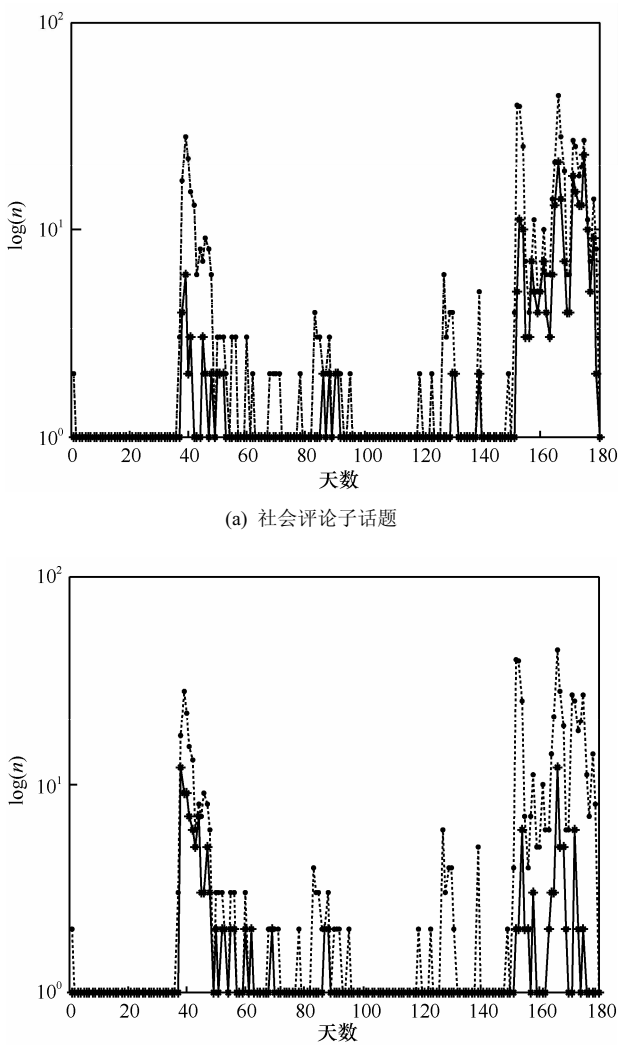
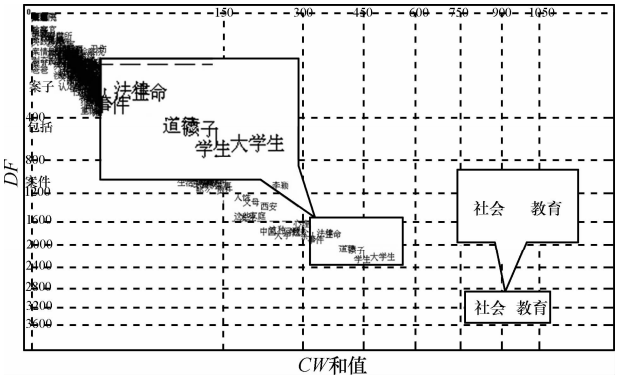


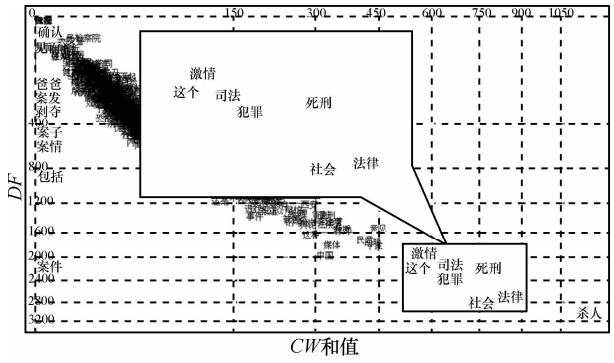
图 10　3 个子话题随时间的变化趋势

在聚类过程中，任何人不知道正确的聚类结果。因此，图 11 中 3 个类别不经过分析便可能会

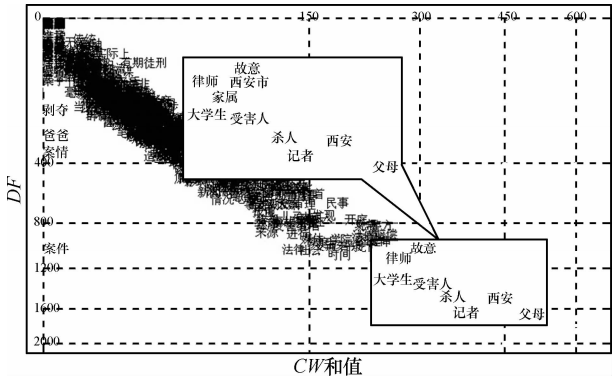
产生令人迷惑或毫无意义的结果。DCW_VSM 的优势之一就是能够使用连接权及依存树来分析该聚类结果表述的主要内容。



(a) 学生社会教育



(b) 激情杀人/判决死刑



(c) 西安大学生杀人

图 11　各子话题的特征词分布情况

图 11 中左上角为坐标原点，每个词条都对应一个坐标，纵轴表示在子话题中出现的词频数 DF，横轴表示该词在当前子话题中连接权 CW 的和值大小，通过统计 CW 以及 DF 即可以将词条在平面上打散开来，位于图中右下方的词条具有较高的词频与连接程度。观察图 11(a)词语主要集中在“学生道德”及“社会教育”方面，图 11(b)主要集中在“激情杀人”、“社会法律”方面，图 11(c)主要集中在“受害人”、“记者”方面，通过依存方式建立的有向连接图，即可以得到各个子话题的描述内容分别为“学生社会教育”、“激情杀人/判决死刑”和“西安大学生杀人”，这与人工设定的 3 个子话题相当接近，并且能够更详尽地表述子话题内容。

4.3　准度测试指标与结果分析

NIST 为 TDT 建立了完整的测评体系[1]，利用系统损耗性能 C_{Det} 曲线来直观地反映系统在误报率与漏报率的表现，在应用过程中使用归一化的性能损耗指标$(C_{\mathrm{Det}})_{\mathrm{Norms}}$来表示，具体如下

$$(C_{\mathrm{Det}})_{\mathrm{Norms}} = \frac{C_{\mathrm{Miss}} \times P_{\mathrm{Miss}} \times P_{\mathrm{target}} + C_{FA} \times P_{FA} \times P_{-\mathrm{target}}}{\min(C_{\mathrm{Miss}} \times P_{\mathrm{target}}, C_{FA} \times P_{-\mathrm{target}})} \quad (7)$$

其中，P_{Miss} 为系统的漏报率，P_{FA} 为系统的误报率；C_{Miss} 为漏报一个新话题的代价，C_{FA} 为误报一次的代价；P_{target} 为在信息流中出现一个新话题的概率，$P_{-\mathrm{target}}$ 为信息流中出现一个老话题的概率，且有 $P_{\mathrm{target}}+P_{-\mathrm{target}}=1$。根据 TDT 的测评标准，设定 C_{Miss}、C_{FA} 及 P_{target} 的值分别为 1.0、0.1、0.02。

各测试指标如表 1 所示，其中，BOW 表示使用词袋法进行子话题的模型表示，LOW 表示使用了依存关联分析之后的连接权表示模型，NE 表示命名实体库。可以看出使用本文方法依存关联分析之后的话题表示模型（LOW+NE）较一般词袋法表示模型（BOW、BOW+NE）在相同的阈值条件下具有更低的归一化错误代价，在平均条件下$(C_{\mathrm{Det}})_{\mathrm{Norms}}$下降了 2.2%。图 12 为错误归一化代价曲线。

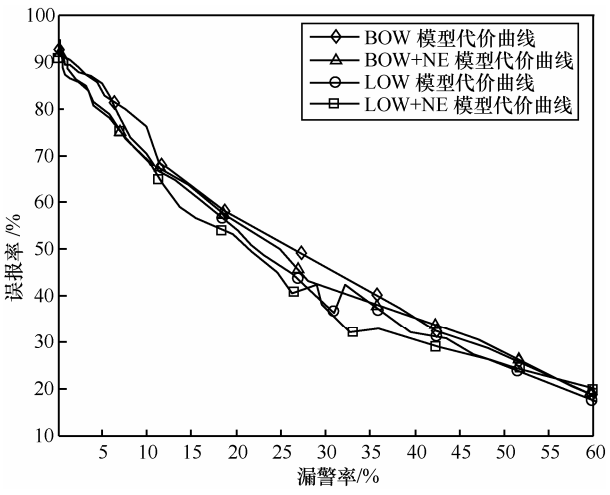


图 12　错误归一化代价曲线

表 1　　**不同话题表示模型下的 P_{FA}、P_{Miss} 以及 C_{Det} 值**

阈值	BOW			BOW+NE			LOW			LOW+NE		
	P_{FA}	P_{Miss}	C_{Det}	P_{FA}	P_{Miss}	C_{Det}	P_{FA}	P_{Miss}	C_{Det}	P_{FA}	P_{Miss}	C_{Det}
0.36	0.395	0.319	2.253	0.36	0.327	2.089	0.003	0.935	0.949	0	0.931	0.931
0.34	0.465	0.273	2.551	0.482	0.262	2.626	0.015	0.904	0.975	0.012	0.904	0.961
0.32	0.734	0.135	3.731	0.734	0.135	3.731	0.032	0.873	1.031	0.02	0.862	0.962
0.3	0.868	0.119	4.374	0.874	0.115	4.399	0.053	0.827	1.085	0.041	0.815	1.016
0.28	0.892	0.112	4.481	0.898	0.108	4.506	0.099	0.762	1.249	0.076	0.735	1.107
0.26	0.939	0.088	4.688	0.939	0.088	4.688	0.146	0.638	1.355	0.146	0.635	1.351
0.24	0.962	0.065	4.779	0.962	0.065	4.779	0.383	0.373	2.25	0.249	0.5	1.718
0.22	0.98	0.05	4.85	0.98	0.046	4.846	0.482	0.285	2.649	0.471	0.304	2.611
0.2	0.997	0.027	4.913	0.997	0.027	4.913	0.687	0.112	3.478	0.681	0.115	3.454
0.18	0.997	0.015	4.901	0.997	0.015	4.901	0.792	0.104	3.987	0.804	0.1	4.04
0.16	0.997	0.004	4.89	0.997	0.004	4.89	0.939	0.104	4.703	0.936	0.112	4.696

可以看出，本文方法（LOW+NE 曲线)更加靠近坐标原点，误报率以及漏报率同时较传统方法(BOW 曲线)低，拥有更小的错误归一化代价。另外还看到这样一个现象：引入命名实体库 NE 的表示模型较不引入 NE 的对应模型性能有所改善，如图 12 中 BOW+NE 曲线较 BOW 曲线稍好；而 LOW+NE 较 LOW 曲线表现稍好。因此得到结论：NE 的引入反映了子话题所表现的侧重点，对提高子话题跟踪准确率有帮助。实验表明，与传统的话题表示模型 VSM 相比，使用 DCW_VSM 模型进行 sTDT 能够有效地降低错误归一化代价，达到更加有效地提取子话题的目的。

5　结束语

本文提出了一种子话题检测与跟踪方法，使用关联词邻接图方式拓展向量空间模型，引入命名实体词典，利用依存树进行分析，提取主要表述要点，减少计算量。通过实验可以验证，本文方法能够有效提取识别相关子话题，降低性能损耗指标。

本文后续的研究将在 CCV_VSM 模型的基础上，加入有监督指导的分类，将子话题分为事件性报道以及评论性报道 2 类，从而能够更加有效地反映子话题内容。另一方面，是在公开测试集（例如本课题组参与的 NLP&CC2012 公开评测任务[13]）上开展 sTDT 研究，以期将本文方法进一步提炼和推广。

参考文献：

[1] 洪宇，张宇，刘挺等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6):71-87.
HONG Y, ZHANG Y, LIU T, *et al*. Topic detection and tracking review[J]. J of Chinese Information Processing, 2007, 21(6):71-87.

[2] ALLAN J, JIN H, RAJMAN M, *et al*. Topic-based novelty detection[A]. Proceedings of the Johns Hopkins Summer Workshop[C]. CLSP, Baltimore, 1999.

[3] ALLAN J, LAVRENKO V, JIN H. First story detection in TDT is hard[A]. Proceedings of 9th Conference on Information Knowledge Management[C]. Washington, DC, 2000. 374-381.

[4] PONTE J, CROFT W B. Text segmentation by topic[A]. Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries[C]. Europe: ECDL, 1997. 113-125.

[5] TU Y N, SENG J L. Indices of novelty for emerging topic detection[J]. Information Processing and Management, 2012, 48:303-325.

[6] ZENG J P, ZHANG S Y. Variable space hidden markov model for topic detection and analysis[J]. Knowledge-Based Systems, 2007, 20: 607-613.

[7] ZENG J P, ZHANG S Y. Incorporating topic transition in topic detection and tracking algorithms[J]. Expert Systems with Application, 2009, 36:227-232.

[8] ZHAO Z Y, FENG S Z, WANG Q, *et al*. topic oriented community detection through social objects and link analysis in social networks[J]. Knowledge-Based Systems, 2012, 26: 164-173.

[9] FLORIAN H, SVEN T. Towards automatic detection and tracking of topic change[A]. Proc of Cicling 2010[C]. Iasi, Romania, LNCS 2010. 327-339.

[10] TESNIERE L, ÉLÉMENTS D. Syntaxe Structurale[M]. Paris: Klincksieck, 1959.

[11] QIAN X, ZHANG Q, *et al*. 2D Trie for fast parsing[A]. Proc of Coling[C]. Beijing, 2010. 904-912.

[12] XIA F. The part-of-speech tagging guidelines for the penn chinese treebank(3.0)[EB/OL]. http://www.cis.upenn.edu/~chinese/posguide.3rd.ch.pdf.

[13] http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html[EB/OL]. 2012.

作者简介：

周学广（1966-），男，江苏高邮人，博士，海军工程大学教授、博士生导师，主要研究方向为信息安全、网络安全以及密码学。

高飞（1988-），男，江西上饶人，海军工程大学硕士生，主要研究方向为网络安全。

孙艳（1983-），女，湖南浏阳人，海军工程大学博士生，主要研究方向为信息内容安全、网络安全。