融合扩展信息瓶颈理论的话题关联检测方法研究

杨玉珍 [1,2] 刘培玉 [1,2] 费绍栋 [1,3] 张成功 [1,2]

摘 要 话题关联检测的关键任务在于判断给定报道对是否属于同一话题. 现有判断方法往往忽略种子事件与其直接相关事件之间的层次关系. 为此, 通过分析报道内部语义分布规律及篇章结构, 并依据语义分布规则, 利用语义分布规律改进信息瓶颈 (Information bottleneck, IB) 算法, 用于子话题逻辑语义单元的划分, 并利用这些逻辑语义单元表示报道, 进行话题关联检测. 实验证明该方法有较快的收敛速度, 并在一定程度上提高了系统性能.

关键词 关联检测, 逻辑语义单元, 信息瓶颈, 单元特征

引用格式 杨玉珍, 刘培玉, 费绍栋, 张成功. 融合扩展信息瓶颈理论的话题关联检测方法研究. 自动化学报, 2014, **40**(3): 471−479

DOI 10.3724/SP.J.1004.2014.00471

A Topic Link Detection Method Based on Improved Information Bottleneck Theory

YANG Yu-Zhen[1,2] LIU Pei-Yu[1,2] FEI Shao-Dong[1,3] ZHANG Cheng-Gong[1,2]

Abstract Topic link detection aims to detect whether two given stories talk about the same topic, whose key task is how to represent the story utilizing a proper model. In the previous works, the hierarchical relationship between seed events and its directly related events is ignored. Thus, this paper analyzes the regular pattern of semantic distribution and the structure of a story, and proposes a method to divide a story into several sections of sub-topic features based on the regular pattern of semantic distribution and improved information bottleneck (IB) theory. Then, the story represented by the attributes is utilized to do topic link detection. Experimental result shows that this method has a fast convergent rate, and can improve the performance of the system.

Key words Link detection, logical semantic unit, information bottleneck (IB), unit features

Citation Yang Yu-Zhen, Liu Pei-Yu, Fei Shao-Dong, Zhang Cheng-Gong. A topic link detection method based on improved information bottleneck theory. *Acta Automatica Sinica*, 2014, **40**(3): 471−479

话题检测与追踪 (Topic detection and tracking, TDT) 的任务在于不断从时序报道流中识别新的事件, 并能够动态追踪该事件的发展[1]. 这一过程往往通过判断报道流中的新报道与以往报道的相关性来确定: 如果新报道与先验报道直接相关, 则可用来追踪先验报道的发展状况, 若不相关, 则可将该报道判定为新事件. 可见相关性判断贯穿于整个 TDT 过程. 相关性判断在 TDT 中称之为关联检测 (Link detection task, LDT), 定义为: 判断给定报道集合 $D = \{d_1, d_2, \cdots, d_n\}$ 中任一报道对 $\langle d_i, d_j \rangle$ 是否直接相关. 遗憾的是现有 LDT 方法往往将话题视为平面新闻的集合, 不仅难以捕捉事件随时间迁移而产生的动态变化, 而且忽视了种子事件及其直接相关事件之间可能存在的层次关系, 如图 1 所示.

图 1 中报道对 $\langle A, B \rangle$ 源于同一个种子事件, 为相关报道对, 如果将其视为平面新闻集合, 通过两篇报道共同包含的关键词很难确定该报道对相关.

不难发现报道 A 包含: 1) “寻求法律手段解决黄岩岛争端”、2) “强调黄岩岛领土主权”、3) “向美国请求支持” 三个存在因果关系层次结构的子话题. 如果以 A 中的 2) 与 B 关联, 易计算 $\langle A, B \rangle$ 为相关话题.

可见, 关联检测的过程不是将话题视为简单的平面集合,而应将报道的篇章结构及其关键特征的语义分布作为关联检测的依据.

收稿日期 2012-09-28 录用日期 2013-03-11
Manuscript received September 28, 2012; accepted March 11, 2013
国家自然科学基金 (60873247), 山东省自然基金 (ZR2012FM038), 山东省科技发展计划 (2012GGB01194) 资助
Supported by National Natural Science Foundation of China (60873247), Natural Foundation of Shandong Province (ZR2012FM038), and Science and Technology Development Plan of Shandong Province (2012GGB01194)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 山东师范大学信息科学与工程学院 济南 250014 2. 山东省分布式计算机软件新技术重点实验室 济南 250014 3. 山东财经大学图书馆 济南 250014
1. School of Information Science and Engineering, Shandong Normal University, Jinan 250014 2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250014 3. Library of Shandong University of Finance and Economics, Jinan 250014

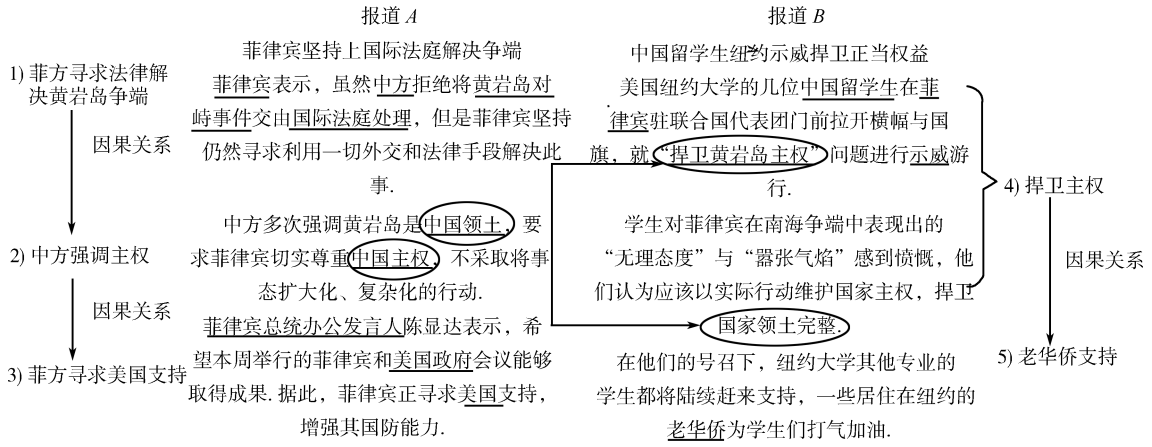

图 1 相关报道示例

Fig. 1 The example of related stories

1 研究现状

目前有关 LDT 的研究多利用统计方法计算报道对的共同特征的数目或权重, 以此作为衡量其相关性的依据. 其中, Kumaran 等[2] 将话题表示为向量空间模型, 估计话题特征在文本中的概率分布, 利用余弦夹角计算报道与话题之间的相似度, 用于衡量报道间的相关性. Allan 等[3] 将报道对 $\langle d_i, d_j \rangle$ 中的一篇报道 d_i 看成一个话题, 采用一元语言模型 (Unigram language model, ULM) 描述报道 d_j 产生话题 d_i 的概率, 并对调报道对中的角色再次估计, 利用相对熵估计该报道对的相关性. Naptali 等[4] 则认为名词的信息含量较高, 于是采用 LSA (Latent semantic analysis) 挖掘上下文中名词间的潜在信息, 利用名词类置信度对给定窗口内出现过的名词打分, 构建话题依赖语言模型用于话题检测. 石晶等[5] 采用 LDA (Latent Dirichlet allocation) 模型完成文本篇章的分割, 并在此基础上确定片断主题, 用于主题分析.

基于统计的方法操作简单, 但基于统计的方法将报道视为词袋的集合, 且词与词之间相互独立, 这种方法忽略了特征与特征之间关系, 反映的是报道间的相似性而非相关性. 如: 报道对 “香港撞船事故死亡人数升至 38 人” 与 “沅江撞船事故死亡人数升至 12 人”, 均包含 “撞船”、“事故”、“死亡人数” 等高频特征, 它们属于相似话题, 却不是相关话题.

为了捕捉特征之间的关系, 语言学知识被引入到统计方法中. 其中, Nallapati 等[6] 将命名实体和词性标注分类, 并利用话题特征产生于不同类别中的概率估计其权重, 该方法有利于提取话题中的关键特征, 但是却存在难以覆盖所有语法现象的缺陷. Chemudugunta 等[7] 将层次性语义概念与统计话题模型统一到一个概率模型下, 用于话题检测. 洪宇等[8] 将语义域的概念引入语言模型, 用于话题关联检测, 很好地解决了报道间的关联问题, 但该方法依赖于依存句法分析, 增加了系统的复杂度. Wang 等[9] 将话题表示为时间、地点、事件、时间和地点特征、发布时间五个维度, 从五个维度计算话题间的相关性, 该方法在一定程度上提高了 LDT 的准确率, 但没有解决报道中话题间的层次关系问题.

为了推动话题检测与追踪技术的发展, 近期研究者们将话题的演化特征[10] 及结构特征[11] 融入话题模型构建中. 为了捕捉有效时间段内相关话题, Garrido 等[12] 认为远距离监督可以有效地从文本集合中抽取话题关联关系, 结合日期标签可以直接确定这些信息. 该算法有利于抽取话题中的关联关系, 但却依赖于时间标记, 而实际上网络中的信息时间标记时常存在不确定现象. 为此, Chambers[13] 结合丰富的语言特征提出两种时间戳自动标注模型. Lakshmi 等[14] 认为具有相同内容的子话题有益于描述话题的局部信息, 基于此, 他们构建了一个用于话题关联检测的凝聚模型, 遗憾的是该模型却没有构建子话题之间的关联关系. Zhang 等[15] 利用层次聚类划分话题, 能够描述话题间的层次关系, 并构建了以话题为根结点、局部事件为叶子结点的树状话题模型. Nomoto[16] 构建了一个两层相似模型, 该模型的第一层次为报道层, 第二层次为通过反馈收集的话题层, 最后利用该层次模型进行话题关联检测, 与 Zhang 等[15] 的层次模型相比, 虽然该模型划分粒度相对较粗, 但这两种模型却有着异曲同工的效果. 该类模型能够静态描述话题内部结构之间的关系, 但却很难依据话题动态演化而更新模型. Zhang 等[17] 通过挖掘报道中涉及的各个事件的关键词元, 并利用关键词元进行事件检测与关系发现. 虽然 Zhang 等意识到话题中各事件之间的相互关系, 却没有对关键词元进行划分.

为了提高 LDT 的精度, 捕捉话题动态演化过程中的特征, 在上述研究的基础上, 本文从更加细微的粒度描述话题与子话题之间的关系, 最终构建一个基于逻辑语义单元的话题表示模型. 该模型中首先将报道划分为不同的逻辑语义单元, 并将逻辑语义单元表示为主题特征与语义特征两个维度, 利用上述两个维度描述话题语义空间在各逻辑语义单元中的分布, 最后采用相对熵评估报道对间的相关性.

2　逻辑语义单元定义

逻辑语义单元是语义趋于一致的语义片断的集合, 且逻辑语义单元分布于子话题中. 因此, 报道 d_i 可表示为 $d_i = \{t_1, t_2, \cdots, t_n\}$, 其中 $t_i = \{w_1, w_2, \cdots, w_n\}$ 为一个逻辑语义单元, w_i 为特征的语义分布, 且 $t_i \neq t_j$.

由于逻辑语义单元粒度较细, 其特征相对稀疏, 信息含量不足, 容易将不相关话题划分为相关话题. 因此, 本文将逻辑语义单元特征划分为主题特征 $E = \{e_1, e_2, \cdots, e_n\}$ 及单元特征 $W = \{w_1, w_2, \cdots, w_n\}$ 两个维度, 其中主题特征用于描述逻辑语义单元与话题之间的关系, 单元特征用于区分报道中各逻辑语义单元, 最终可将逻辑语义单元定义为: $T_i = \langle E_i, W_i \rangle$.

逻辑语义单元划分过程中, 信息损失量越小划分越精确. 因此, 本文将这一过程视为信息压缩的过程, 依据逻辑语义单元定义首先扩展传统的信息瓶颈 (Information bottleneck, IB) 方法, 并利用扩展的 IB 方法划分逻辑语义单元.

3　融合扩展 IB 理论的逻辑语义单元划分

3.1　主题特征与单元特征提取

3.1.1　主题特征提取

依据定义, 主题特征是逻辑语义单元间相关性判断的第一个层次, 应包含较多的语义信息, 且在整篇报道和逻辑语义单元中均有着较高的权重. 因此, 本文综合利用报道中的特征以及逻辑语义单元中的信息共同计算其权重, 计算方法如下:

$$tf_E(t) = \frac{tf_L}{l} \times \frac{tf_D}{N} \tag{1}$$

其中, tf_L 为特征项在逻辑语义单元中出现的次数, l 为逻辑语义单元的长度, tf_D 为特征项 t 在报道中出现的次数, N 为报道的长度. 由于名词及命名实体具有较高的语义描述性, 实验中对命名实体及名词进行 2 倍加权.

3.1.2　单元特征提取

单元特征为逻辑语义单元的标志, 该特征应尽可能地描述逻辑语义单元. 本文采用 $tf\text{-}idf$ 方法计算单元中的特征权重, 并按照权重大小排序, 选取权重较高的特征作为单元特征.

$$w_{ij} = \frac{tf_{ij} \times \log \dfrac{N}{n_i}}{\sqrt{\displaystyle\sum_{t \in S} [tf_{ij} \times \log \frac{N}{n_i}]^2}} \tag{2}$$

其中, tf_{ij} 表示特征项 t_i 在逻辑语义单元 S_j 中出现的频次; n_i 为包含特征项 t_i 所有逻辑语义单元数.

3.2　扩展的 IB 方法

3.2.1　IB 方法简介

IB 理论源于香农的率失真理论, 由 Tishby 等[18] 率先提出, 并应用于文本聚类中. 朱真峰等[19]、沈华伟等[20] 及 Du 等[21] 又进一步在该算法上进行改进并分别用于文本聚类、社交网络挖掘及文本倾向性分析等领域.

IB 理论的主要思想如下: 给定两个随机变量 X, Y 的联合分布, 从中寻找一个可行的压缩 T, 使得最小化 $I(X;Y)$ 的同时, 最大化 $I(T;Y)$. 具体表示如下:

$$L[p(t|x)] = I(T, X) - \beta I(T, Y) \tag{3}$$

式中, β 为拉格朗日因子, 用以约束 T 中包含较多 Y 的信息, 且 $0 < \beta < \infty$.

IB 方法通过计算条件概率 $p(y|x)$ 与 $p(y|t)$ 之间的 Kullbac-Leibler (KL) 散度来衡量 x 与压缩 t 的失真率[18]:

$$D_{KL}[p(y|x)||p(y|t)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|c)} \tag{4}$$

初始状态时, X 无任何压缩, 即每个 $x \in X$ 为一个类别. 根据局部信息损失最小化原则, 选择合并后 $I(T;Y)$ 为一个类别. 可将 $\langle t_i, t_j \rangle$ 的合并代价定义为

$$\delta I(t_i, t_j) = I(T_{\text{before}} Y) - I(T_{\text{after}} Y) \tag{5}$$

最终可通过一些简单变换合并代价转化为容易计算的 Jensen-Shannon (JS) 距离, 具体计算依据文献 [18] 中的方法.

3.2.2　融合多维度特征的扩展 IB 方法

由于 IB 理论仅涉及特征空间一个维度, 而逻辑语义单元却包含主题特征及单元特征两个维度, 因此, 本文对 IB 方法进行扩展.

为了便于说明, 令 T_i 表示某个待合并逻辑语义单元, W_i 表示逻辑语义单元的特征词空间, E_i 表示

逻辑语义单元的主题特征空间, T_i^* 表示聚类过程的逻辑语义单元, W_i^* 为聚类过程中逻辑语义单元的特征词空间, E_i^* 为聚类过程中逻辑语义单元的主题特征空间.

令 $I(T_i;W_i)$ 为单元特征与逻辑语义单元之间的关系, $I(T_i;E_i)$ 为主题特征与逻辑语义单元之间的关系, $I(T_i^*;W_i^*)$ 描述了聚类过程中单元特征与逻辑语义单元之间的关系, $I(T_i^*;E_i^*)$ 描述了聚类过程中主题特征与逻辑语义单元之间的关系.

经过上述定义, 引入单元特征及主题特征后可将 IB 方法扩展为:

$$\alpha[I(T_i;E_i)-I(T_i^*;E_i^*)]+\gamma[I(T_i;W_i)-I(T_i^*;W_i^*)] \tag{6}$$

其中, $\alpha>0$, $\gamma>0$, 表示指导强度, 且 $\alpha+\gamma=1$.

为了便于计算, 本文将式 (6) 转化为更为容易计算的 Kullback-Leibler 散度:

$$\alpha D_{\rm KL}(k(T_i,E_i)||k^*(T_i,E_i))+\gamma D_{\rm KL}(f(T_i,W_i)||f^*(T_i,W_i)) \tag{7}$$

下面证明式 (6) 与式 (7) 的等价关系.

定义 1. 令 $f(T_i,W_i)$ 为 T_i 和 W_i 的联合概率分布, 则有:

$$f(t_i,w_i)=p(t_i,w_i) \tag{8}$$

定义 2. 令 $f^*(T_i,W_i)$ 表示 T_i 和 W_i 在逻辑语义单元聚类过程中 (T_i^*,W_i^*) 中的联合分布概率, 有:

$$f^*(T_i,W_i) = p(t_i^*,w_i^*)p(t_i|t_i^*)p(w_i|w_i^*)= p(t_i^*,w_i^*)\frac{p(t_i)}{p(t_i^*)}\frac{p(w_i)}{p(w_i^*)} \tag{9}$$

定义 3. 令 $k(T_i,E_i)$ 为 T_i 和 E_i 的联合概率分布, 有:

$$k(t_i,e_i)=p(t_i,e_i) \tag{10}$$

定义 4. 令 $k^*(T_i,E_i)$ 表示 T_i 和 E_i 在逻辑语义单元聚类过程中 (T_i^*,E_i^*) 的联合概率分布, 则有:

$$k^*(T_i,E_i)=p(t_i^*,e_i^*)\frac{p(t_i)}{p(t_i^*)}\frac{p(e_i)}{p(e_i^*)} \tag{11}$$

引理 1.

$$I(T_i;W_i)-I(T_i^*;W_i^*)= D_{\rm KL}(f(T_i,W_i)||f^*(T_i,W_i)) \tag{12}$$

证明. 式 (13) 成立, 同样方法可证明:

$$I(T_i;E_i)-I(T_i^*;E_i^*)=D_{\rm KL}(f(T_i,E_i)||f^*(T_i,E_i))$$

因此, 式 (6) 等价于式 (7). □

3.3 基于扩展 IB 理论的逻辑语义单元划分算法

将扩展的 IB 理论用于逻辑语义单元划分的具体过程如下.

输入. 给定话题报道集合 T, 参数 λ.

输出. 给定话题报道逻辑语义单元 T_i.

Begin:

1) 将 T 中每篇报道, 切分为段落集合 $P=\{p_1,p_2,\cdots,p_n\}$;

2) 为每个段落依据式 (1) 和式 (2) 建立主题特征空间及单元特征空间;

3) 针对每篇报道中任意段落对 $\langle p_i,p_j\rangle$, 依据式 (7) 计算每个段落对之间的损失度 d_{ij};

While true

找出 d_{ij} 最小的 i,j;

合并 p_i, p_j 为 p^*;

从 P 中删除 p_i,p_j, 并把 p^* 加入到 P;

更新与 p^* 相关的 d_{ij};

If $\forall p\in P$, 且 $d_{ij}\geq\lambda$

Break.

$$\begin{aligned}
&I(T_i;W_i)-I(T_i^*;W_i^*)=\sum_{t_i^*\in T_i^*}\sum_{w_i^*\in W^*}\sum_{t_i\in t_i^*}\sum_{w_i\in w_i^*}p(t_i,w_i)\log\frac{p(t_i,w_i)}{p(t_i)p(w_i)}-\\
&\sum_{t_i^*\in T_i^*}\sum_{w_i^*\in W^*}\Big(\sum_{t_i\in t_i^*}\sum_{w_i\in w_i^*}p(t_i,w_i)\Big)\log\frac{p(t_i^*,w_i^*)}{p(t_i^*)p(w_i^*)}=\\
&\sum_{t_i^*\in T_i^*}\sum_{w_i^*\in W^*}\sum_{t_i\in t_i^*}\sum_{w_i\in w_i^*}p(t_i,w_i)\log\frac{p(t_i,w_i)}{p(t_i^*,w_i^*)\dfrac{p(t_i)}{p(t_i^*)}\dfrac{p(w_i)}{p(w_i^*)}}=\\
&\sum_{t_i^*\in T_i^*}\sum_{w_i^*\in W^*}\sum_{t_i\in t_i^*}\sum_{w_i\in w_i^*}f(t_i,w_i)\log\frac{f(t_i,w_i)}{f^*(t_i,w_i)}=\\
&D_{\rm KL}(f(T_i,W_i)\|f^*(T_i,W_i))
\end{aligned} \tag{13}$$

4　基于逻辑语义单元的话题关联检测

尽管报道被划分为逻辑语义单元的集合, 但只有能够精确描述报道主题的逻辑语义单元才有意义. 因此, 需要选择出最可能精确描述报道主题的逻辑语义单元.

直观上, 位置越靠前的句子越能精确描述报道的主题, 位置靠后的句子信息含量较小, 句子的重要程度与其在整个报道中的位置呈反比. 此外, 逻辑语义单元越长, 信息含量越多. 因此, 本文依据下式对逻辑语义单元进行排序:

$$Weigh(T_i) = \frac{|T_i|}{\sum\limits_{s\in T_i} loc(s)} \log \frac{size(E_i)+size(W_i)}{docsize} \tag{14}$$

其中, $|T_i|$ 表示逻辑语义单元中句子的总数, $loc(s)$ 表示逻辑语义单元 T_i 中句子的位置, $size(E_i)$、$size(W_i)$、$docsize$ 分别表示逻辑语义单元主题特征数目、单元数目以及报道的主题特征数目.

本文选择 Kullback-Leibler 散度来评估报道对间的相关性, 判断过程如下: 若判断报道对$\langle D_1, D_2\rangle$ 的相关性, 其中 D_1 被划分为三个逻辑语义单元 $T_1 = t_1, t_2, t_3$; D_2 被划分为两个逻辑语义单元 $T_2 = t_4, t_5$; 且每个逻辑语义单元 t_i 中包含特征词 w_i 及主题特征 e_i. 逻辑语义单元相关性可利用单元特征相关性及主题特征相关性来判断. 单元特征的相关性计算如下:

$$D_{\mathrm{KL}}(t_1||t_4) = \sum_w p(w_i|t_1)\log\frac{p(w_i)}{p(w_i|t_4)} \tag{15}$$

其中, $p(w_i|t_1)$ 与 $p(w_i|t_4)$ 分别为特征 w_i 在逻辑语义单元 t_1 和 t_4 中的概率分布.

主题特征相关性计算方法如下:

$$D_{\mathrm{KL}}(t_1\|t_4) = \sum_e p(e_i|t_1)\log\frac{p(e_i)}{p(e_i|t_4)} \tag{16}$$

其中, $p(e_i|t_1)$ 与 $p(e_i|t_4)$ 分别为特征 e_i 在逻辑语义单元 t_1 和 t_4 中的概率分布.

5　实验与结果分析

5.1　实验语料

本文实验采用语言标准协会提供的 TDT4 中文文本语料进行评测, 该语料共包含 26 066 个报道对, 其中 3 075 对相关报道, 其余为不相关. 实验采用开放测试, 首先从上述语料中抽取 1 400 对相关报道、8 000 对不相关报道, 共同组成训练集, 剩余作为测试集.

5.2　评测指标

实验以检测错误代价作为评价指标, 检测错误代价是由美国国家标准与技术研究院针对 TDT 发布的评测指南, 它分别从漏报率及误报率进行评价. 计算方法如下:

$$(C_{\mathrm{Det}})_{\mathrm{norm}} = \frac{C_{\mathrm{Miss}}P_{\mathrm{Miss}}P_{\mathrm{target}} + C_{\mathrm{FA}}P_{\mathrm{FA}}P_{\mathrm{non-target}}}{\min(C_{\mathrm{Miss}}\cdot P_{\mathrm{target}}, C_{\mathrm{FA}}\cdot P_{\mathrm{non-target}})} \tag{17}$$

其中, C_{Det} 为检测错误代价, P_{Miss} 与 P_{FA} 为系统漏报率及误报率; 漏报是指系统未能识别出新的话题, 误报是指系统将旧话题的相关报道误判为话题; C_{Miss} 表示漏报代价系数, 这里取 $C_{\mathrm{Miss}} = 1$; C_{FA} 代表误报代价系统, 本文取 $C_{\mathrm{FA}} = 1$; P_{target} 和 $P_{\mathrm{non-target}}$ 为先验目标概率 (通常取 $P_{\mathrm{target}} = 0.02$, $P_{\mathrm{non-target}} = 1 - P_{\mathrm{target}}$).

5.3　实验流程

本文设置了三组实验, 分别用于验证下述问题.

1) 实验 1: 设置四个系统用于检测本文提出的逻辑语义单元层次模型的可行性.

System-1: 该系统以整篇报道作为话题关联检测的单位, 主要采用向量空间模型表示报道, 利用余弦距离计算话题间的相关性;

System-2: 该系统以整篇报道作为话题关联检测的单位, 采用一元语言模型表示报道, 具体采用文献 [3] 中的方法;

System-3: 该系统以逻辑语义单元作为话题关联检测的单位, 且逻辑语义单元采用一元语言模型表示;

System-4: 该系统以逻辑语义单元作为话题关联检测的单位, 并将逻辑语义单元表示为主题特征及单元特征两个维度的语言模型.

语义单元合并损失函数采用式 (6) 计算, 报道对关联检测采用式 (15) 及 (16) 计算. 需要说明的是 Sytem-3 仅从特征的维度描述逻辑语义单元, 利用式 (10) 计算损失度时仅考虑式中的第 2 项, 在关联检测过程中也仅采用式 (14) 进行计算.

2) 实验 2: 用于验证本文提出的基于逻辑语义单元划分的关联检测模型是否优于现有模型. 本文分别与文献 [8] 中的 SDLM (Semantic domain language model) 模型、文献 [9] 中提出的 EMW (Event words model) 模型以及文献 [16] 中提出的 TTSM (Two-tier similarity model) 模型进行对比. 其中, 对于文献 [9] 中的 EMW 模型, 本文采用了该模型中的 $\mathrm{EMW}_{\mathrm{withoutPD}}$ 的形式; 文献 [16] 中的 TTSM 模型, 本文采用该模型中的 1st+2nd 模型; 最后为文本提出逻辑语义单元层次模型, 这里仍然

用 System-4 表示.

3) 实验 3: 用于验证本文采用扩展信息瓶颈方法凝聚逻辑语义单元是否可行. 由于凝聚策略可分为基于划分的方法及基于凝聚的方法两类, 针对这两种类型设置实验, 其中基于划分的策略, 本文采用目前表现较好的 LDA 划分策略; 基于凝聚策略的方法则分别采用余弦距离 (cosine, COS)、KL 距离、JS 距离及欧几里德距离 (Euclidean distance, EUC) 进行比较. 本文方法这里仍表示为 IIB (Improved information bottleneck method).

5.4 实验结果与分析

1) System-3 及 System-4 的参数估计

System-3 及 System-4 中涉及合并代价参数和指导强度参数. 合并代价参数指导着逻辑语义单元聚类的过程, 其值越大, 报道集合所包含的逻辑语义单元越少, 同时逻辑语义单元中包含的报道数目也随之增加. 随着该值的持续增高, 逻辑语义单元内无关报道数目也随之增加. 如图 2 所示.

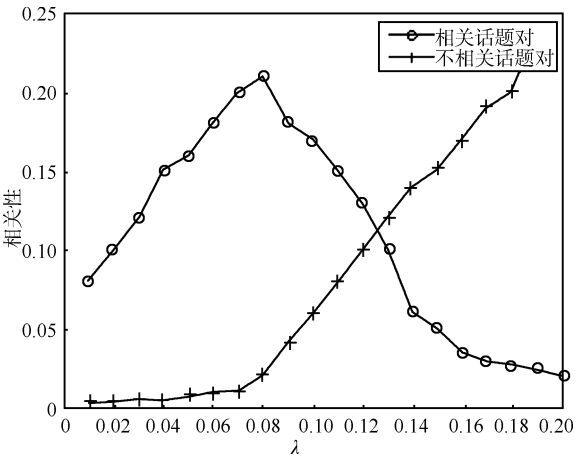

图 2 段落对相关度随参数 λ 变化趋势

Fig. 2 The change trend of relevance between paragraph pair with parameter λ

从图 3 中可以看出, 当 λ 值为 (0.07, 0.09) 时, 无论是相关段数目还是逻辑语义数目均呈上升趋势, 直至 $\lambda = 0.09$ 时, 相关段落对的平均值处于一个衰减点. 随着 λ 不断增加, 逻辑语义单元数目不断减少, 不相关段落对的数目也不断增加. 同时, 在 $\lambda < 0.07$ 时的一段间隔内, 虽然逻辑语义单元数目及相关段落对数均呈上升趋势, 但是由于此时合并代价相对较小, 逻辑语义单元中所包含的段落对较少, 存在较多独立段落, 难以描述整个语义空间. 整体来看, λ 的值服从正态分布, 考虑到相关度随合并代价的变化趋势, 并参照语料分布情况, λ 值为 0.07 时, 系统处于最优状态.

本文除了涉及合并代价参数外, 还涉及到一个指导强度参数 α. 该参数主要用来平衡新闻主题及语义单元内的特征对相关度的影响. 图 4 描述了随着参数 α 的变化, 逻辑语义单元的个数及相关段落对数的分布情况.

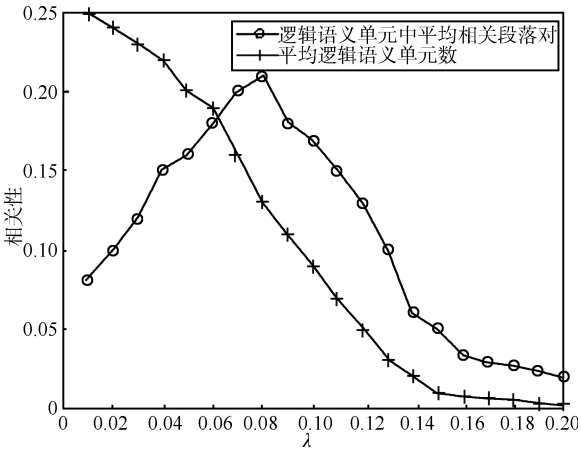

图 3 逻辑语义单元数目及相关度随 λ 变化趋势

Fig. 3 The change trend of relevance between paragraph pair and the number of logical semantic unit with parameter λ

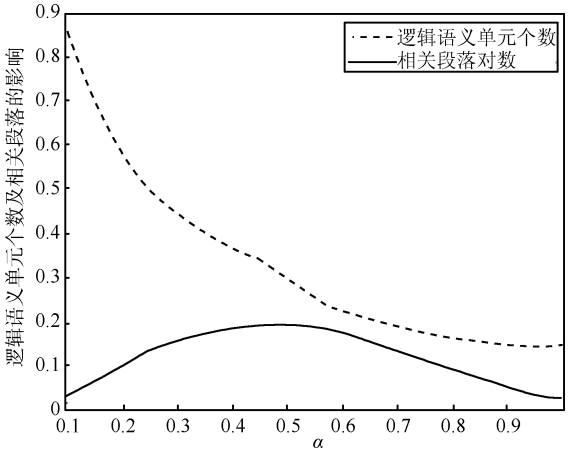

图 4 逻辑语义单元数及相关段落对随 α 的变化趋势

Fig. 4 The change trend of relevance between paragraph pair and the number of logical semantic unit with parameter α

当 $\alpha > 0.5$ 时, 在逻辑语义单元合并过程中, 新闻主题的损失代价对逻辑语义单元个数影响越大, 则单元特征对逻辑语义单元个数影响较大. 如图 4 所示.

新闻主题往往为全篇共有特征, 与标题有着强联系, 当把它的影响扩大到一定程度时, 逻辑语义单元个数将会变小, 所含不相关段落对个数会增加. 相反, 扩大单元特征的影响到一定程度时, 逻辑语义单元粒度将会缩小、数目增加, 出现多个独立段落作为逻辑语义单元, 难以描述语义空间. 本文依据特征空

间的分布情况, 估计 α 值为 0.4.

2) 测试结果对比

a) 实验 1

各系统在测试集上的评测结果如表 1 所示, 易见 System-4、Sytem-3 的评测结果明显优于前两个系统, 并且 System-4 优于 System-3, 这一结果说明将报道表示为层次模型的可行性. System-4 优于 System-3 的原因在于: System-4 面向话题各逻辑语义单元建立话题模型, 并将逻辑语义单元表示为主题特征及单元特征两个维度, 这种方法不仅能够描述话题的核心语义, 并且削弱了大概率主题掩盖小概率主题带来的干扰.

b) 实验 2

表 2 描述了本文提出的关联检测模型与其他各类模型的比较结果, 容易看出 System-4 与 SDLM 明显优于 TTSM 及 EMW. 与 SDLM 方法相比, System-4 提高不大, 但是本文方法却避免了复杂的句法分析和计算, 减小了系统的开销. 从这个角度来看, 本文方法优于 Baseline 的方法.

c) 实验 3

表 3 描述了不同凝聚策略 LDT 评测结果. 结果显示, EUC、COS、KL、JS、IIB 几种不同凝聚策略, IIB 方法最优. 这是因为 EUC、COS、KL、JS 几类基于距离的凝聚策略较适用于中心分布数据, 而对于话题尤其是逻辑语义单元而言, 特征的分布是一种离散状态, 而不是中心分布, 因此效果相对较差. 而 IIB 方法是从局部最优依次向全局最优搜索的方法, 并且采用主题与特征的联合分布表示逻辑语义单元, 有利于描述特征的分布情形. LDA 是一种基于统计的概率生成性模型, 它首先假设每篇文档由一些潜在的话题分布组成, 每个话题又是一组特征的多项分布, 因此 LDA 模型是一种较为实用的话题表示模型. 遗憾的是该模型没有考虑到话题间的层次关系, 这也是本文方法略优于 LDA 的主要原因.

d) 程序收敛性验证

本文中逻辑语义单元聚类过程采用的是一种迭代方法, 因此需要通过实验验证本文方法的收敛性, 实验结果如图 5 所示.

由图 5 可以看出, 本文所提方法经过若干次迭代是可以收敛的, 并且有着较快收敛速度. 当迭代次数超过 6 次以后, 系统性能趋于平稳. 为了充分保证系统性能, 本文实验取迭代次数为 10 次.

表 1　LDT 评测结果对比

Table 1　Comparison of LDT results

System	λ	News subjects	Features	Min norm (Cost)
System-1	-	-	55	0.58021
System-2	-	-	68	0.5983
System-3	0.38	-	63	0.2354
System-4	0.07	11	48	0.18406

表 2　各模型 LDT 评测对比

Table 2　Comparison of LDT results of different models

System	λ	News subjects	Features	Min norm (Cost)
SDLM	0.09	-	35	0.18453
TTSM	-	-	-	0.2033
EMW	-	-	-	0.2554
System-4	0.07	11	48	0.18406

表 3　不同凝聚策略 LDT 评测结果

Table 3　Comparison of LDT results of different cohesion strategies

Clusting	λ	News subjects	Features	Min norm (Cost)
COS	0.37	9	41	0.2301
EUC	0.45	11	51	0.5109
KL	0.09	8	49	0.2487
JS	0.07	9	55	0.2108
LDA	-	-	45	0.1857
IIB	0.07	11	48	0.18406

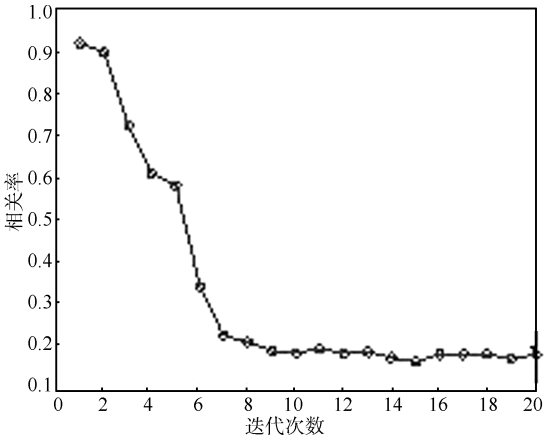

图 5 不同迭代次数下的漏报率

Fig. 5 The change trend of relevance between paragraph pair and the number of logical semantic unit with parameter λ

6 总结

本文考虑报道的篇章结构及语义分布, 利用信息瓶颈方法将报道划分为逻辑语义单元的集合, 针对逻辑语义单元建模, 将逻辑语义单元描述为主题特征及单元特征两部分, 最后利用相对熵计算报道对的相关性. 实验证明这种方法避免了语义建模过程中复杂的句法分析, 并能有效提高相关话题检测的精度.

本文在研究过程中发现, 复杂的句法分析的确能够提高系统精度, 但提升幅度不大. 相对而言, 话题建模过程中引入句法知识反而大幅度增加系统开销. 基于统计的方法相对简单且难以描述话题内部各特征之间的联系. 因此, 在今后的研究中应当寻找一种以统计方法为基础, 合理引入浅层语义知识的途径用以话题建模.

References

1 The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan [Online], available: http://www.itl.nist.gov, September 20, 2012

2 Kumaran G, Allan J. Text classification and named entities for new event detection. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2004. 297−304

3 Allan J, Carbonell J, Doddington G, Yamron J, Yang Y M. Topic detection and tracking pilot study final report. In: Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop. Lansdowne, Virginia: 1998. 194−218 Morgankaufmampubl: shers,

4 Naptali W, Tsuchiya M, Nakagawa S. Topic-dependent language model with voting on noun history. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2010, **9**(2): 1−31

5 Shi Jing, Fan Meng, Li Wan-Long. Topic analysis based on LDA Model. *Acta Automatica Sinica*, 2009, **35**(12): 1586−1592
(石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析. 自动化学报, 2009, **35**(12): 1586−1592)

6 Nallapati R, Feng A, Peng F C, Allan J. Event threading within news topics. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM). New York, USA: ACM, 2004. 446−453

7 Chemudugunta C, Smyth P, Steyvers M. Combining concept hierarchies and statistical topic models. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2008. 1469−1470

8 Hong Yu, Zhang Yu, Fan Ji-Li, Liu Ting, Li Sheng. Chinese topic link detection based on semantic domain language model. *Journal of Software*, 2008, **19**(9): 2265−2275
(洪宇, 张宇, 范基礼, 刘挺, 李生. 基于语义域语言模型的中文话题关联检测. 软件学报, 2008, **19**(9): 2265−2275)

9 Wang L T, Fang L. Story link detection based on event words. In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer-Verlag, 2011, **6609**: 202−211

10 Hu Yan-Li, Bai Liang, Zhang Wei-Ming. Modeling and analyzing topic evolution. *Acta Automatica*, 2012, **38**(10): 1690−1697
(胡艳丽, 白亮, 张维明. 一种话题演化建模与分析方法. 自动化学报, 2012, **38**(10): 1690−1697)

11 Zhu T, Wang B, Wu B, Zhu C X. Topic correlation and individual influence analysis in online forums. *Expert Systems with Applications*, 2012, **39**(4): 4222−4232

12 Garrido G, Peñas A, Cabaleiro B, Rodrigo Á. Temporally anchored relation extraction. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 107−116

13 Chambers N. Labeling documents with timestamps: learning from their time expressions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 98−106

14 Lakshmi K, Mukherjee S. Using cohesion-model for story link detection system. *International Journal of Computer Science and Network Security*, 2007, **7**(3): 59−66

15 Zhang K, Zi J, Wu L G. New event detection based on indexing-tree and named entity. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007. 215−222

16 Nomoto T. Two-tier similarity model for story link detection. In: Proceedings of the 19th ACM International Conference Information and Knowledge Management. New York, USA: ACM, 2010. 789−798

17 Zhang Kuo, Li Juan-Zi, Wu Gang, Wang Ke-Hong. Term-committee-based event identification within topics. *Journal of Computer Research and Development*, 2009, **46**(2): 245−252
(张阔, 李涓子, 吴刚, 王克宏. 基于关键词元的话题内事件检测. 计算机研究与发展, 2009, **46**(2): 245−252)

18 Tishby N, Pereira F, Bialek W. The information bottleneck method. In: Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing. Illinois, USA: IEEE, 1999. 368−377

19 Zhu Zhen-Feng, Ye Dong-Yang, Li Gang. Iterative sIB algorithm based on mutation. *Journal of Computer Research and Development*, 2007, **44**(11): 1832−1838
(朱真峰, 叶东阳, Li Gang. 基于变异的迭代 sIB 算法. 计算机研究与发展, 2007, **44**(11): 1832−1838)

20 Shen Hua-Wei, Cheng Xue-Qi, Chen Hai-Qiang, Liu Yue. Information bottleneck based community detection in network. *Chinese Journal of Computers*, 2008, **31**(4): 677−686
(沈华伟, 程学旗, 陈海强, 刘悦. 基于信息瓶颈的社区发现. 计算机学报, 2008, **31**(4): 677−686)

21 Du W F, Tan S B, Cheng X Q, Yun X C. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In: Proceedings of the 3rd ACM International Conference Web Search and Data Mining. New York, USA: ACM, 2010. 111−120

杨玉珍 山东师范大学信息科学与工程学院博士研究生. 主要研究方向为倾向性分析, 话题检测与追踪. 本文通信作者. E-mail: zscyyz@126.com
(**YANG Yu-Zhen** Ph. D. candidate at the School of Information Science and Engineering, Shandong Normal University. Her research interest covers sentiment analysis and topic detection and tracking. Corresponding author of this paper.)

刘培玉 山东师范大学信息科学与工程学院教授. 主要研究方向为话题检测与追踪, 倾向性分析, 网络信息安全.
E-mail: liupy@sdnu.edu.cn
(**LIU Pei-Yu** Professor at the School of Information Science and Engineering, Shandong Normal University. His research interest covers sentiment analysis, topic detection and tracking, and network and information security.)

费绍栋 山东财经大学图书馆讲师. 主要研究方向为话题检测与追踪, 中文倾向性分析. E-mail: shiyi1984@msn.com
(**FEI Shao-Dong** Lecturer at the Library of Shandong University of Finance and Economic. His research interest covers sentiment analysis and topic detection and tracking.)

张成功 山东师范大学信息科学与工程学院硕士研究生. 主要研究方向为中文倾向性分, 话题检测与追踪.
E-mail: zcg870108@163.com
(**ZHANG Cheng-Gong** Master student at the School of Information Science and Engineering, Shandong Normal University. His research interest covers sentiment analysis and topic detection and tracking.)