话题关联识别中报道信息的动态扩充研究

张晓艳 王 挺
(国防科技大学计算机学院 长沙 410073)

摘 要 话题关联识别用于判断新闻报道对流中每对中的两篇报道是否描述了同一个话题。为解决其中报道篇幅短小、稀疏问题严重及其内容存在漂移等问题,提出了一种动态信息扩充技术,用于改进报道表示模型。该技术用过去最新的话题相关报道来扩充当前报道,动态更新原有模型。此外,还研究了扩充信息的精化问题,通过有选择地加重一些重要特征的权重来减小扩充过程中噪音带来的影响。该方法在TDT4中的中文语料上进行了实验,结果表明动态信息扩充技术能够较大幅度地改进话题关联识别的性能,对多种特征采取的精化技术也对性能改进产生了较大影响。

关键词 话题关联识别,动态信息扩充,报道模型
中图法分类号 TP301 **文献标识码** A

Research on the Dynamic Extending of Story in Story Link Detection

ZHANG Xiao-yan WANG Ting
(Department of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract Story Link Detection is to determine whether two stories are about the same topic. To overcome the limitation of the story length, sparse data and the drifting problem in story content, this paper provided a technology of dynamic information extending to improve the story representation model. It extended the current story with its previous latest topic-related story. The refinement on the information for dynamic extending was also studied. It aims to reduce the influence of the noise introduced when extending by increasing the weights of some important features in the extending story. This method was used for Story Link Detection on the TDT4 Chinese corpus. The experiment results indicate that the technology of dynamic extending and the refinement of extending information can both affect the performance of story link detection systems evidently.

Keywords Topic detection and tracking, Dynamic information extending, Story representation model

1 引言

新闻是人们的主要信息来源,也是自然语言处理研究人员关注的对象。话题发现与追踪(Topic Detection and Tracking, TDT)[1,2]就是以大规模的新闻语料库为研究对象的,该研究通过监控新闻报道所描述的话题,来发现新的用户感兴趣的信息并跟踪下去,最后将涉及某个话题的报道组织起来以某种方式呈现给用户。作为话题发现与追踪的核心技术,话题关联识别[1]的任务是判断新闻报道对流中每对中的两篇报道是否描述了同一个话题。这里的话题指发生在特定时间、地点的一个核心事件或活动,以及所有与之直接相关的事件或活动。该任务被认为是基于话题研究的基础,有着较为重要的意义。

目前应用到话题关联识别中的方法主要分为两类:基于向量空间模型的方法[3,4]和基于概率模型的方法[5,6]。两者各有优缺点,其中前者一直是话题关联识别研究的主流,它首先将文本内容转换成易于数学处理的向量,使得各种相似运算和排序成为可能,在TDT研究中一直都表现较好。但该模型的局限在于其独立性假设,即向量特征间相互独立,模型转换的过程中不考虑关联信息;概率模型理论基础扎实,有较好的发展潜力,但由于报道通常较为简短精练,使模型原本就存在的稀疏问题更加严重,对训练数据较为依赖的平滑技术在有实时性限制的TDT研究中效果也会大打折扣。

针对新闻报道信息含量较少,难以在表示模型中准确描述其话题,这一问题在上述两类方法中都有研究。其中心思想都是对报道的内容进行信息扩充。但已有研究的扩充信息通常来自历史训练语料或第三方知识源。例如,文献[7]利用互信息度量训练语料中命名实体之间的关联度,并利用该关联信息把相关的命名实体扩充到当前处理的新闻报道中;文献[5]通过建立相关模型的方法从训练语料中抽取"主题"相关的报道扩充当前报道。但无论是历史训练语料还是第三方知识源,由于它们和测试报道在时间上的较大差异,很难从中

到稿日期:2008-12-08 返修日期:2009-02-25 本文受国家自然科学基金资助项目(60403050),新世纪优秀人才支持计划(NCET-06-0926)资助。

张晓艳(1981-),女,博士生,CCF会员,主要研究方向为话题发现与追踪,E-mail:zhangxiaoyan08@hotmail.com;王 挺(1970-),男,博士,教授,博士生导师,CCF会员,主要研究方向为自然语言处理、计算机软件等。

获取与测试报道真正"话题"相关的信息，通常是和报道中用词或词分布相近的一些知识或文档，不能从根本上解决信息缺少的问题，反而可能引入大量噪音，使模型中的报道话题发生错误漂移。

经过分析发现：话题相关的信息最可能来自测试过程中已经处理的新闻报道，它们在时间和内容上都与将来的未来测试数据最接近。根据这一特征的启发，从已处理报道中抽取与当前报道话题相关的报道进行扩充，以缓解另一个报道在以后再次出现时模型表示中的稀疏问题，而且这种信息扩充同时也兼顾了新闻报道中话题的动态发展特性，这正是本文方法的中心思想所在。但该技术在加入有用信息的同时也引入了一定的噪音。为此，通过加重其中一些重要信息的比重来相对降低噪音带来的影响。实验表明，无论是动态扩充技术还是对扩充信息的研究都在话题关联识别上取得了较好的改进。

本文第 2 节重点介绍基于动态信息扩充的话题关联识别方法，主要包括表示模型、信息动态扩充的方法；第 3 节描述动态扩充信息中的多种重要特征及其精化策略；第 4 节给出实验结果及详细分析，并与相关研究进行比较；最后做出总结。

2 动态信息扩充

报道通常以尽可能短的篇幅来描述所发生的事情，这使表示模型较难准确描述报道所包含的话题。而且，新闻数据流中的话题存在动态演化特征，进一步加大了模型表示的难度。若想同时解决好上述两个问题，必须在新闻报道中加入有效的信息。扩充信息通常有以下来源：

• 历史训练语料：虽然较易获取，但其中话题相关信息很少或没有，只能对报道进行"主题"相关的信息扩充，而且这类信息对解决话题漂移帮助也不大。

• 第三方知识源：通常指 Hownet、同义词词林等，虽然能在一定程度上缓解新闻报道的稀疏问题，但由于扩充的是重复信息，而非新的话题相关信息，因此不能兼顾解决话题漂移。

事实上，背景测试语料（指已处理的测试数据）才是最有可能包含话题相关，使稀疏问题和漂移问题都得到缓解的信息。本文的扩充信息来源正是背景测试语料。在背景测试语料中选择的用于扩充报道的信息，应该来自与之话题相关的新闻报道。对这类信息有两种选择方法：背景测试语料中所有与之话题相关的报道，或者其中最新的与之相关的报道。文献[8]中的实验表明，使用后者比前者性能要好。原因可能在于前者在扩充进较多重复信息的同时也引入了较多噪音，从而降低有用信息在表示模型中的比重，导致模型间的相似度降低，丢失率上升，这一观点也得到了实验数据的支持。本文实验中的扩充信息都来自背景测试语料中与当前测试报道话题相关的最新报道。

2.1 表示模型

基于向量模型构建的话题关联识别系统是目前性能最好的系统之一。本文也采用向量空间模型表示报道。向量特征为预处理之后的词，但对于同一个词，如果标记为不同词性，则认为是不同特征。报道预处理包括分词、词性标注、停用词过滤，其中分词和词性标注由中科院的汉语词法分析系统 ICTCLAS[1] 完成，所使用的停用词表共包括 507 个停用词，停用词过滤时不考虑词性信息。

特征权重计算采用增量 $tf*idf$ 方法[8]。与传统 $tf*idf$ 方法相比，主要不同之处在于增量 $tf*idf$ 计算是动态的。这是因为在话题关联识别中，测试数据按时间先后排序，随着已经比较的报道对数的增多，用于统计的背景测试语料动态递增，从中获取的一些统计信息也是动态的。因此我们称之为增量 $tf*idf$ 方法，这种做法使统计获得的信息与当前处理报道最接近，能在最大程度上反映真实情况。

选取相似度计算函数的一个重要标准是该函数是否能够区分描述相同话题和描述不同话题的新闻报道对。在众多的相似度计算方法中，余弦函数[4]性能最好、最稳定[9]。向量经过标准化后，余弦函数仅是两个向量的内积，即向量的夹角余弦值。本文的所有实验都采用余弦函数计算相似度。

2.2 动态扩充方法

动态扩充技术的中心思想在于：比较两个报道的话题相关性时，在生成报道表示模型的同时对报道内容进行动态扩充。由于扩充信息来自背景测试语料中最新话题关联报道，而且背景测试语料动态递增，因此同一篇报道若处在不同的比较对中，其扩充的报道也可能不一样，进而生成的模型也不一样。方法的动态性也是本文信息扩充的优势之一。下面首先描述使用动态扩充方法的话题关联识别实现流程：

1）对训练语料进行预处理，包括分词、词性标注、停用词过滤。

2）确定扩充阈值：在训练语料（同样是有序的新闻报道对序列）上实现无信息扩充的话题关联识别方法，找出该方法性能最优时的关联阈值，把该值作为动态话题关联识别测试时的扩充阈值。由于本文关注的是系统的最优性能，而关联阈值的作用在于获取系统当前性能，因此本文没有介绍如何确定动态话题关联识别的关联阈值。

3）顺序处理测试语料中的每个新闻报道对：

• 对两个报道进行预处理，包括分词、词性标注、停用词过滤；

• 如果比较对中的报道有最新的话题相关报道，则用其对该报道进行动态扩充，同时生成新的表示模型用于比较；

• 若两模型相似度大于扩充阈值，则比较对中的报道互为其最新的话题相关报道，由此看出一篇报道的最新相关报道至多为一个；

• 若两模型相似度大于关联阈值，则认为比较对中的两篇新闻报道话题相关。

考虑了两种扩充方法：增量扩充和平均扩充，区别主要在于如何处理扩充报道和被扩充报道中共有特征的权重。首先假设被扩充报道 s_1 的初始表示模型（即无信息扩充的向量表示模型）为 $M_1=\{(f_{1i},w_{1i})\mid i\geqslant 1\}$，扩充报道 s_2 的初始表示模型为 $M_2=\{(f_{2i},w_{2i})\mid i\geqslant 1\}$，扩充后 M_1 为被扩充报道的新表示模型。

1）增量扩充：对 M_2 中的每一个特征 f_{2i}，若 $f_{2i}=f_{1j}$，$f_{1j}\in M_1$，则 w_{1j} 保持不变，否则把 (f_{2i},w_{2i}) 加入到 M_1 中，扩充后

[1] 计算所汉语词法分析系统 ICTCLAS3.0 白皮书，http://www.i3s.ac.cn/Manual/

的 M_1 称为增量模型。

2)平均扩充:对 M_2 中的每一个特征 f_{2i},若 $f_{2i}=f_{1j}$,$f_{1j}\in M_1$,则 $w_{1j}=0.5*w_{1j}+0.5*w_{2i}$,否则把$(f_{2i},w_{2i})$加入到 M_1 中,扩充后的 M_1 称为均值模型。

扩充后的表示模型不仅包括报道本身的内容,还有其最新话题相关报道的内容,在信息含量上更加丰富,而且由于扩充信息来自于话题相关的报道,话题的动态演化特性也得到了兼顾。因此新的表示模型对报道话题的描述能力更强,模型相似度对话题关联性的衡量也更加准确、合理。实验[8]表明,平均扩充方法比增量扩充方法在保持误判率不变的情况下具有较低的丢失率,原因可能在于:平均扩充比增量扩充能更倾向于用两篇报道的平均值来表示报道,更能突出它们的公共部分,而这部分和话题的种子事件通常比较接近,正是表述话题的关键部分。本文后续的扩充信息精化研究都基于平均扩充方法。

3 信息精化

上述动态信息扩充方法用话题相关报道扩充时,在加入话题相关信息的同时,不可避免地也引入了噪音。而且,模型中的特征对报道所含话题的描述能力不一样,甚至起混淆作用,单靠 $tf*idf$ 方法计算权重还不足以对它们进行很好的区分。为降低噪音带来的影响,或者更加凸现一些特征在表示话题时的作用,我们仔细分析特征的报道位置信息和词性信息,对其中一些重要特征进行精化,达到与过滤噪音近似的目的。下面将给出 3 种精化特征及其精化策略。

需要指出的是,该精化策略不仅适用于扩充报道,也适用于被扩充报道。因此,在验证精化策略的有效性时,它们也同样被用于比较对中的两篇报道中。

3.1 精化特征

1)核心特征

通过人工大量阅读新闻报道发现,新闻写作者对标题的选择和使用非常慎重。人们往往仅通过浏览标题就能了解报道内容的大概。对那些没有标题的报道,例如广播、电视新闻等,通常也会首先通过一两句话来概括接下来要播报的内容。这些信息为核心信息,在表示报道话题上有重要作用。

核心信息的抽取方法如下:对于有明确标题的报道直接从中抽取标题作为核心信息;否则抽取出文本内容的第一句进行判断,若该句不含"报告、报道、播报、收听、主持、广播、收看"类词语,把第一句作为核心信息;否则认为该句是与报道该新闻的人或单位有关的信息,与话题无关,继续抽取第二句作为核心信息。最后对抽取出的核心信息进行预处理(分词、词性标注、停用词过滤),获取核心特征集合。

2)名实体特征

话题中心事件及其直接相关事件的主要构成要素包括:人名、地名、组织名、时间词和数词,即命名实体,它们对报道所描述话题起到了一个很好的定位作用。但是由于话题具有时间效应和动态发展的特性,这使描述同一话题的不同报道内的时间词和数词变化比较大,表示方法多样化,因此前三类实体词在表达报道话题上显得更加重要,称这类词为名实体特征。其获取方法较简单:对报道内容进行预处理之后直接判断候选特征的词性即可。

3)依存名词特征

如果说上述核心特征和名实体特征对新闻报道中的话题起到了概括和定位作用,那么名词通常可看作是对报道内容进行具体化和细化。但并不是所有的名词都对话题描述同等重要,那些与名实体相依存的名词在表现或丰富话题内容上比其他名词往往更为重要,这些词为依存名词特征。依存名词和"主题"词比较接近,对这些词进行精化对突出报道中的话题有帮助。依存名词的获取方法如下:

• 获取新闻报道的依存关系对集合:首先对文本预处理,即分词、词性标注(包括命名实体识别);然后在本文的词性列表和句法分析器词性列表之间作映射,主要是命名实体的词性映射。如果分析器中没有相应的词性与之对应则映射到上一级词性上去。例如:Stanford-Parser 中待分析句子中的单词词性没有对应的命名实体词性,就把识别出的命名实体中的人名、地名、组织名的词性都统一映射为名词词性;最后输入斯坦福大学开发的中文依存句法分析器 Stanford-Parser[2],输出依存关系对集合。所有句子分析结果合并后构成整篇报道的依存关系对集合。

• 对集合中的每个依存关系对作判断:如果关系对中一方是名词,另一方是名实体特征,那么该名词是依存名词特征,应该被精化。

3.2 精化策略

在对上述 3 类特征信息进行精化时,需考虑两种简单直观的精化策略:

1)权重加倍精化

计算出候选特征的 $tf*idf$ 权重之后,对该特征进行判断,若该特征属于待精化特征集合,则直接加倍其权重;否则保持权重不变。最后对生成权重向量进行标准化。

2)频次加倍精化

在统计出该特征在报道中出现的次数后,判断候选特征是否包含在精化信息特征集合内,若属于精化信息,则加倍特征频次;否则保持原频次不变。然后计算特征的 $tf*idf$ 权重并标准化。

4 实验及分析

4.1 实验语料及评价方法

本文实验采用 TDT4[3] 中文语料中的 12334 个报道比较对,其中前 2334 个用于训练获得扩充阈值,后 10000 个用于测试本文所描述多个方法的有效性。

为准确评价系统性能,本文使用 TDT2003 的评测方法及其实现软件[10]。该软件基于系统的丢失率和误报率,构造了性能的主要评价指标:错误识别代价,其计算公式如下:

$$C_{det}=C_{miss}*P_{miss}*P_{target}+C_{fa}*P_{fa}*P_{non\text{-}target} \quad (1)$$

其中,C_{det} 是错误识别代价;C_{miss},C_{fa},P_{target} 和 $P_{non\text{-}target}$ 都是预定义值,其中前两个是丢失和误报一个话题相关报道对的代价,后两个是两个报道是否话题相关先验概率,取值分别为 1,0.1,0.02 和 0.98;P_{miss} 与 P_{fa} 分别是系统识别结果中对话

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://projects.ldc.upenn.edu/TDT4/

题相关判断的丢失率和误报率。

为了使得到的性能指标具有可比性，通常将上述代价做归一化处理得到$(C_{det})_{norm}$，归一化公式如下：

$$(C_{det})_{norm} = C_{det} / \min(C_{miss} * P_{target}, C_{fa} * P_{non\text{-}target}) \quad (2)$$

归一化处理后的识别代价最小值仍然为0，但如果取值为1，则意味着该系统的性能不强于总是给出肯定判定或者否定判定的系统。

以上描述都是基于一个话题。评测包含多个话题的关联识别系统有话题加权和报道加权两种方法，通常采用前者为主要评价指标，本文展示的所有话题关联识别实验结果都基于话题加权。

错误识别代价取决于一个好的识别方法和阈值选择方法。如果阈值选择不好，即使系统有很低的错误识别代价，其当前错误识别代价也可能很高，本文重点介绍识别方法，主要关注系统的最小检测代价，即最优性能。

4.2 实验结果

为了验证动态扩充方法以及信息精化技术的有效性，实现了以下4类话题关联识别系统：

1)基准方法：用向量模型表示报道，用余弦相似度计算模型相似度，即话题相关度。

2)在基准方法基础上使用动态信息扩充方法。

3)在基准方法基础上分别对核心特征、名实体特征和依存名词特征进行精化，每种特征精化分别使用两种精化策略，共3组系统。

4)在基准方法基础上使用动态信息扩充方法外，还分别对核心特征、名实体特征和依存名词特征进行精化，每种特征精化分别使用两种精化策略，共3组系统。

首先，分析两种精化策略的有效性。图1是上述使用了精化技术的识别系统的性能比较，共有6组。其中前3组是第3类系统，后3组是第4类系统，都依次对核心特征、名实体特征、依存名词特征进行精化。每组中的前一个系统使用频次加倍精化，后一个系统使用权重加倍精化。对每个系统，在图中依次列出其错误识别代价、丢失率和误判率。总体来说，对核心特征、名实体特征、依存名词特征3类信息使用权重加倍的精化策略较使用频次加倍好，识别代价更低，主要由于其在降低丢失率上更有优势。接下来仅分析使用权重加倍策略时的相关实验结果，如表1所列。

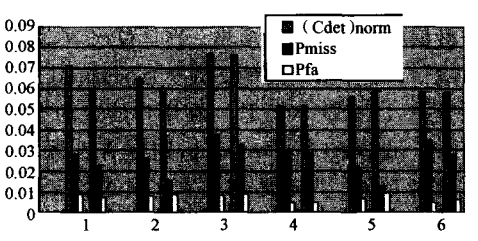

图1 对精化信息使用不同策略下的性能对比图

表1 话题关联系统实验结果

	动态信息扩充	核心特征精化	名实体特征精化	依存名词特征精化	$(C_{det})_{norm}$	P_{miss}	P_{fa}
基准方法					0.0840	0.0514	0.0076
基准方法+动态信息扩充	√				0.0588	0.0345	0.005
基准方法		√			0.0591	0.0224	0.0075
基准方法+信息精化			√		0.0602	0.0156	0.0091
				√	0.0762	0.0329	0.0088
基准方法	√	√			0.0518	0.029	0.0046
+动态信息扩充	√		√		0.0583	0.0123	0.0094
+信息精化	√			√	0.0576	0.028	0.006

从表中可以看到：

1)动态信息扩充技术对话题关联识别有非常明显的改进作用。单独使用时使识别代价从基准方法的0.0840降到0.0588，改进了30%，比最有效的核心特征精化技术(单独使用的识别代价为0.0591)都要好。

2)3类信息的精化也分别在不同程度上使表示模型更准确地描述了报道所含的话题，它们的有效性按照依存名词特征、名实体特征、核心特征的顺序依次上升，分别使基准方法的性能改进9.29%，28.33%，29.64%。

3)从基本评价指标上看：

·在丢失率上，动态信息扩充技术的改进(降低32.88%)稍差于对3类信息中任一类信息的精化(最小改进幅度为36%)；3类信息中名实体特征则对改进丢失率贡献最大，在基准方法和动态信息扩充方法下都使丢失率保持在1.5个百分点甚至更少，可见名实体仍是改进报道表示模型、使之对话题描述更加充分准确的最重要的一类词。

·在误判率上，在动态扩充技术单独使用时就大幅降到0.5%，而3类信息精化则对误判率改进甚微，甚至导致了误判率的上升。3类信息中以核心特征对误判率的改进最有效：使基准方法的误判率由0.76%降为0.75%，使动态信息扩充下的误判率由0.5%降为0.46%。其他两类信息精化则使误判率上升。经过分析，认为原因可能在于：精化信息中的噪音会提高表示模型的相似度，如果大于关联阈值则表现为使系统的误判率上升。此外，精化策略虽然能缓解"新闻比较对中两篇报道是关于同一个话题的不同侧面描述"这一现象，然而对"新闻比较对中两篇报道是关于同类中的不同话题"却帮助不大，反而会受其影响而更加严重。

此外，还对上述3类信息的组合精化进行了研究，发现：对核心特征进行精化是使性能取得较大幅度改进的一个必要措施，但是并不是一味强调增加所有重要的信息就可以无限度地改进系统性能，两两组合或3个组合时的识别代价均高于单独使用核心特征精化时的代价。

4.3 结果分析

由于新闻报道在篇幅上的限制以及内容上存在的话题漂移问题，使得仅依靠报道的内容不能准确描述报道中的话题，对报道进行信息扩充是解决该问题的一个有效的方法，本文正是从这一点出发提出了一种新的动态信息扩充方法并对信息进行精化。与相关研究[5,7]相比，本文使用了更合适的扩充信息来源，并对扩充信息进行了话题区别度方面的研究。相关方法实现后，在本文数据上的测试结果稍差于上述实验结果，说明了本文方法的优越性。

使用信息扩充技术时应从以下几点考虑：

1)扩充信息来源：本文从背景测试数据中获取真正话题相关的扩充信息，比同义词、主题相关信息、共现信息等更真实有效。文献[5]从训练语料中抽取"主题"相关信息，不仅信息不够准确，对训练数据量要求较大，而且很难抽取到相关信

(下转第241页)

进行的，如模拟网络瞬间遭到大量的恶意数据包攻击，若本文设计的调控系统工作在真实的网络环境中，性能表现会比实验时要好一些。

结束语 本文基于内分泌系统构造的调控系统，仅用两种激素模拟网络的可用性和安全性，事实上，正如内分泌系统本身的复杂性一样，网络的性能评价也是很复杂的。这就需要：一方面对内分泌系统进行更深入的研究，从而提取更优的调节机制用于设计基于内分泌系统的调控算法；另一方面对网络性能评估指标进行深入的研究，设计出更科学合理的评价方案。这些都是本文下一步研究的重点。

参 考 文 献

[1] 刘克龙，蒙杨，卿斯汉. 一种新型的防火墙系统[J]. 计算机学报，2000，23(3)：231-236

[2] 张磊，卿斯汉. 一个基于 Agent 的防火墙系统的设计与实现[J]. 软件学报，2000，11(5)：642-645

[3] Fulp E W. Optimization of network firewall policies using ordered sets and directed a cyclical graphs，1265[R]. Winston-Salem，USA：Wake Forest University，2004

[4] Chen Wenhui，Wang Weiping，Li Zhepeng. Dynamic update of firewall policy based on MFDT，2375 [R]. Winston-Salem，USA：Wake Forest University，2006

[5] 张万会，王复周. 神经、免疫及内分泌系统间的关系[J]. 生理科学进展，1993(03)：261-268

[6] 陈得宝，赵春霞. 基于内分泌调节机制的粒子群算法[J]. 控制理论与应用，2007，24(6)：1005-1010

[7] 刘宝，丁永生，王君红. 一种基于内分泌超短反馈机制的智能控制器[J]. 计算机仿真，2008，25(1)：188-191

[8] 王祎，陈为栋，顾幸生. 基于内分泌激素调节机制的免疫算法的 Flowshop 调度问题[J]. 系统仿真学报，2008，20(13)：3425-3430

[9] 逄曙光. 内分泌与代谢病的免疫学发病机制研究[D]. 济南：山东大学，2008

(上接第 203 页)

息。文献[7]把过去共现过的命名实体都扩充进模型，门槛过低，导致引入大量噪音，淹没了真正的话题相关信息，很难再对扩充后的报道话题关联性进行判断。

2)扩充信息种类：通过实验[11]发现，表示模型转换过程中若有信息丢失，系统性能会受损。例如："印度古吉拉特邦发生地震"，其中"地震"不是实体词却与该话题关系密切。文献[7]的方法只使用命名实体表示报道，是方法性能较差的原因之一。文献[5]及本文所提方法都使用整个相关报道进行扩充，后者还对扩充信息做了进一步分析。

3)扩充信息量：扩充信息并不是越多越好，相反扩充信息越多，噪音就越多，尤其是相关信息较少的情况下，加大扩充信息量会带来系统性能的急剧下降[7]。本文方法使用最新话题相关报道扩充的性能就优于使用所有话题相关报道进行扩充[8]，但对信息累积精化却带来了性能损失。在实现文献[5]中方法时也对扩充报道个数进行了实验，实验表明随着扩充报道个数的增多，系统性能逐渐下降。

结束语 针对报道表示中存在的稀疏问题和话题动态演化问题，本文提出一种信息动态扩充方法，并对模型中的核心特征、名实体特征、依存名词特征 3 类信息进行精化，以进一步改进话题表示模型。该方法用于改进话题关联识别研究中的报道表示。实验表明，无论是动态扩充方法还是 3 种特征精化都能够较好地改进系统性能，尤其是扩充技术和核心特征精化对降低误判率和丢失率都有较大的影响，是改进识别效果的两个有效途径。同时发现，对重要信息的定位及组合精化策略都还需要进一步研究。在知识表示方面也要认识到：虽然使用了比词法更进一步的句法知识，但仍然停留在较浅的知识层次，应该进一步挖掘。

参 考 文 献

[1] James A，et al. Introduction to Topic Detection and Tracking in Topic Detection and Tracking：Event-based Information Organization[M]. Kluwer Academic Publishers，2002：1-16

[2] Wayne，Charles L. Topic Detection and Tracking (TDT)：Overview & Perspective[C]// Proceedings of the Broadcast News Transcription and Understanding Workshop. Lansdowne，Virginia，1998

[3] Margaret C，Ao F，Giridhar K，et al. UMass at TDT 2004[C]// Proceedings of the 7th Topic Detection and Tracking (TDT2004). Gaithersbury，2004

[4] Francine C，Ayman F，Thorsten F. Multiple Similarity Measures and Source-Pair Information in Story Link Detection[C]// HLT-NAACL 2004. Boston，2004：313-320

[5] Victor L，James A，Edward D，et al. Relevance models for topic detection and tracking[C]// Proceedings of Human Language Technology Conference (HLT). California，2002

[6] Ramesh N. Semantic language models for topic detection and tracking[C]// Proceedings of the HLT-NAACL 2003 student research workshop. Edmonton，2003

[7] Chirag S，Bruce C W，David J. Representing documents with named entities for story link detection (SLD)[C]// CIKM 2006. Virginia，2006

[8] Zhang Xiaoyan，Wang Ting，Chen Huowang. Stroy Link Detection based on Dynamic Information Extending[C]// Proceedings of the International Conference on The Third International Joint Conference on Natural Language Processing (IJCNLP2008). Hyderabad，2008

[9] Thorsten B，Francine C，Ioannis T. Topic-Based Document Segmentation with Probabilistic Latent Semantic analysis[C]// Proceedings of the International conference on Information and Knowledge Management (CIKM). McLean，2002

[10] The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan [OL]. http：// www. nist. gov/speech/tests/tdt/tdt2003/evalplan. htm

[11] 张晓艳，王挺，陈火旺. 基于 SVM 的多向量文本表示模型话题关联识别研究[C]// 第七届中文信息处理国际会议. 武汉，2007