基于语义的信息时序检测技术设计研究*

郑学伟[1,2]
(1.辽宁广播电视大学 沈阳 110034; 2. 辽宁装备制造职业技术学院 沈阳 110161)

摘 要: 针对信息的交互与获取正日益突破时间与空间的限制,提出了一种基于语义技术的语义域话题关联检测相关性判定模型,模型是基于文本理解和语义分析的判定方法,其核心思想是根据不同话题生成对应的语义结构体,使系统能够实现自动根据语义信息对话题进行相关性判定,仿真实验结果表明文本的误检率还是漏检率都得到了明显的降低,因此,结果证明基于语义的信息时序检测模型能够有效提高对报道中语义空间中主题相关性检测的能力,对于话题的时序检测后期的研究有积极的意义。

关键词: 信息检测;关联检测;语义结构;时序属性
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: D520.6040

Research on design of information detection technique based on semantic

Zheng Xuewei[1,2]
(1. Liaoning Radio and TV University, Shenyang 110034, China;
2. Liaoning Vocational and Technical College of Equipment Manufacturing, Shenyang 110161, China)

Abstract: Information interaction and acquisition are now gradually breaking through the limit of time and space, Text detection technology based on semantic domain has become the new research in this field. This paper proposes topic link detection correlation model based on semantic technology, which involves the basic researches of text comprehension and semantic analysis in relevance determination. It is of great importance for subsequent researches in topic temporal detection. The core of this method is to establish the semantic structures for specific topics, by which the relevance determination can be achieved automatically reacting to the semantic information.

Keywords: topic detection; link detection; semantic structure; temporal feature

1 引 言

作为语义网技术的关键组成部分之一,信息检测(topic detection,TD)主要研究基于文本的语义自动识别、分类以及检测等方面。TD 研究的对象包括以文本描述的确定时间地点发生的事件和其相关外延的话题。相关研究还延伸到事件本身后续的报道中。TD 研究起源于美国国防高级研究计划局(DARPA),最初的重点是当时处于文本研究的重点信息检索和信息抽取技术是否能够有效应用于 TD 问题[1]。目前在文本信息检测领域中对于语义的识别、分析和归纳主要的方式还是人工处理,效率非常低同时误检率也很高。因此,设计出一种基于语义分析面向文本数据进行时序检测识别的应用模型,能够对相近话题及其后续报道合理的组织成为一个有机整体显得尤为重要。

2 基础研究

最初的 TD 研究将信息定义为“话题”。含义由最初事件慢慢扩展至间接或导致发生的延伸事件,以及同这一事件直接或间接相关的其它事件。目前“话题”的定义是:“一个话题由一个种子事件或活动以及与其直接相关的事件或活动组成”[2]。话题的外延不能无限扩展。随着信息检测研究的逐步发展,目前已经将信息切分任务(SST)不作为 TD 研究的基础部分[3],并首次提出了有指导的自适应话题跟踪(ATT)和层次信息检测(HTD)概念。关联检测(LDT)研究是 TD 研究的重要基础工作。研究的主要内容是对随机抽取的两篇报道进行分析检测是否属于同一主题。作为 TD 研究的基础,LDT 研究的辅助作用无法忽视。目前大部分 LDT 研究主要聚焦于文本的特征描述与语义选择,手段主要是采取向空间模型,报道相似性判定由

收稿日期:2016-04
* 基金项目:辽宁省教育厅科技资助(L2014579)、辽宁省现代远程教育学会专项规划课题(2014XH-BXFZ-12)、辽宁广播电视大学规划课题(2014XB02-12)资助项目

余弦夹角进行衡量。权重由文本中的特征描述的概率决定。概率的计算依据语言模型(简称为 LM)描述产生。最终的相关性则依据概率分布采用 KLD 算法综合得出。

2.1　跨语言信息检测(TD)

随着地球村时代的到来,基于跨语言交流的 TD 研究已经变成 TD 研究领域最重要的研究方向,几乎所有关于 TD 领域的研究都要对语料本身的研究做出深入的探索,目前机器翻译(MT)使用的主要技术模型还是 NIST 系统,MT 系统可以在 TD 过程中对不同的语料进行相互之间的转化,将不同类别的源语言通过机器翻译形成多源单一语言(MLS),但是由于各种语言的结构与表达差异很大,机器翻译不能圆满的解决 TD 问题。如何使系统能够不脱离任何一种语言的本源环境进行 TD 分析是目前难以解决的问题,本地语言假设(NLH)认为两篇报道的匹配算法都只能在源语言的环境下才能达到最佳的效果[4]。NLH 的不足是源语言结构的性能依赖于最初通过匹配算法得到语言结构,如果采用基于统计策略解决跨语言的问题,可以采用 Bayesian 算法描述由报道组成的独立特征集合来匹配话题相关度,但是报道间不相关的信息干扰因素太多也会导致性能很难有更大的提高[5]。因此可以在统计策略的基础上通过采用自然语言信息结合来进行改进可以得到比较好的效果,利用语言特征所在的文本位置进而获取其上下文的权重来匹配其在目标语言中对应的涵义,采取以上方法可以明显提高跨语言 TD 系统的性能。

2.2　关联检测

关联检测(LDT)研究是 TD 研究中一个重要的分支,主要是对不同报道是否在本质上属于同一话题进行检测,LDT 研究属于 TD 研究领域底层设计中的一项基础研究,可以在 TD 研究的任何一个领域发挥作用[6]。对一篇新的报道相关性的检测方式是要判断这篇报道中的话题是否与目标报道集合中所有的报道达到要求的相关性,反复的检测过程实质上重复多次的文本数据相关性的匹配运算。因此,基于文本语义的内部层次结构、规则与相关性检测将直接决定 LDT 的效果,这也是 LDT 的核心问题。在关联检测的基础上应用统计策略的 LDT 研究根据报道共有特征的权重进行特征集合的匹配运算。量空间模型(VSM)根据余弦夹角衡量报道之间的相关性并估算文本中的权重,根据一元语言模型(简称 ULM)描述运算得出的概率,利用相对熵对集合中报道进行相关性综合评估[7]。TD 系统时序检测模型的建立不能仅仅依靠内容,对于不同报道间的关联性以及后续的延伸报道必须要考虑其时序的特征,目前研究主要包括以下两个方面:1)话题单元信息权重的衡量,即描述出目标话题单元在报道中的权重;2)实体相关性的线性分析,即通过人工的方式干预实体相关性匹配的结果。未来 TD 研究领域将主要侧重于概率模型与 NLP 相融合进行研究[8]。

3　基于语义的话题关联检测模型

3.1　语义域

在语义域空间中维护语义的一致性主要通过语义片段来实现并由语境分析来提高其全面性,以图 1 为例,句 a 包含"金大中获得和平奖"为内容的片断,就实质内容来看,句 a 句 c 和句 e 相对一致,可以构成语义域(a-c-e),图 1 中的报道内容主要围绕"获奖"解读了 3 个方面的内容,非别是:"事件"、"原因"以及"介绍"。在这其中"事件"是主体,而"原因"和"介绍"是这一主体对应的拓展话题,在图 1 中,"获奖"这一报道在主体的基础上聚集了一系列的包含起因、发展、推导以及因果等语义片段为框架,并引导其围绕主题进行论述,并最终建立起整个文章的框架。再以此为媒介展开论述。因此,主题的挖掘要符合语义结构,从报道中提炼出新的结构-语义域,并根据算法计算出其出现的频率。

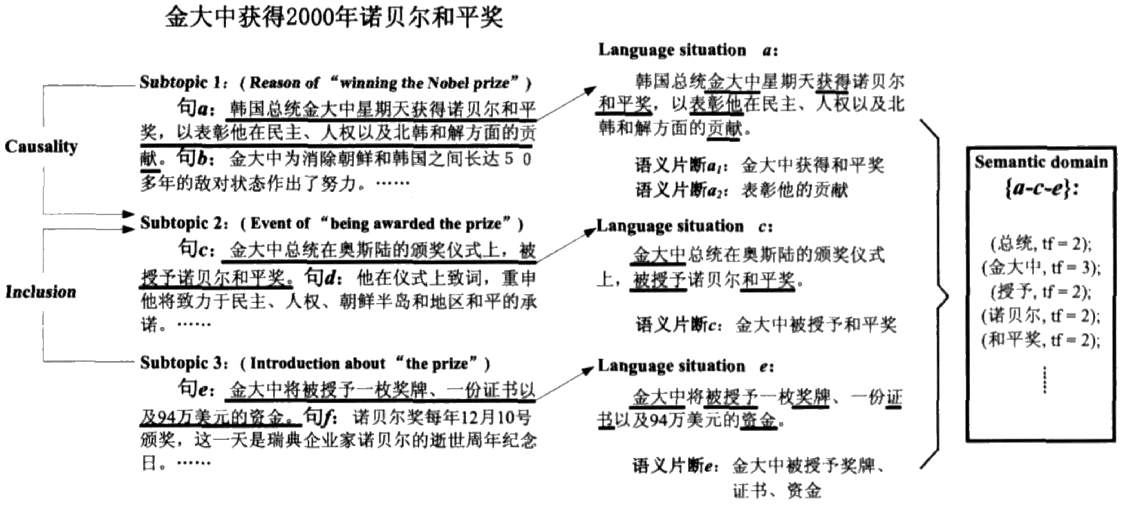

图 1　报道结构化分析样例

语义域同子话题的区别在于，语义域侧重的是内部结构的统一；子话题则依赖话题的定义，可以以一篇报道的主题为集合中心建立与语义片段之间的联系。图 1 中 Subtopic 1“金大中生平在人权、民主与和解方面做出的贡献”，通过语义片断 a1：“金大中获得和平奖”可以与主题建立联系。语境抽取原则的设置将直接影响语义的分析结果，小则导致语义的描述无法完整，过大则影响一致。

图 1 中的句 c 是一个设置的比较合理的语境，整句的语义统一于“金大中被授予和平奖”。语义倾向性的衡量是构造语境并进行语义域凝聚的过程中的关键环节。假设待测报道 D，语义域提炼过程如下：

1）设 D 中集合为 $S=\{s_1,s_2,s_3,\cdots,s_n\}$，建立对应的语义空间；

2）针对 S 中所有句子对 $\{(s_i,_j)\mid s_i,s_j\in S\}$，相关度 $P(s_i\mid s_j)$；

3）训练阈值 θ，将相关度大于 θ 的候选语义域设置为目标语义域；

候选语义域的重组过程是整个策略的核心部分，提炼成功的基础是语义分析传递性是否成立。图 1 中的句 a 包含两个语义片断 a1，a2，其中 a1 使句 a 相关于句 c；a2 所在的子句使句 a 相关于句 d，但句 c 和句 d 的语义并不相关，传递性不成立。语义域的合并与分解是提炼过程不可或缺的环节。如图 2 所示。s_i 与 s_k 之间的关联度由他们与 s_j 之间的关联度间接得到，若阈值 θ 设定的过高或者过低将导致语义域聚集的效果发生错误或者离散化，语义域将借鉴 TFIDF 计算权重较高的集合组成语义空间。

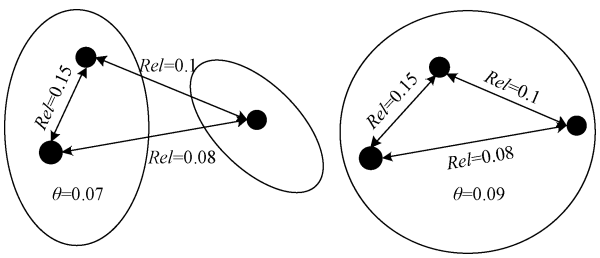

图 2　非传递性语义域凝聚关系

3.2　语义域信息检测模型（semantic domain information detection model，SDIDM）

如果对个别高频特征进行屏蔽可能会导致在语义倾向的问题上忽略语义上的联系。如果词的特征比较相似，检测系统很难发现语义上的差别，为解决这一难题，词汇间的关联性可以通过依存二元词对（以下简称依存对）描述语义的能力来判定，重要性通过其词特征出现的频率来决定。假定有一语义域，存在词 x 与 y 依存关系，则权重计算公式如下：

$$W_{[x,y]}=\frac{tf_{SD}(x)}{l}\cdot\frac{tf_{SD}(y)}{l}\sum_i\log\left(a+\frac{n_i}{n_j}\right)\qquad(1)$$

式中：$tf_{SD}(x)$ 与 $tf_{SD}(y)$ 分别表示 x 与 y 的特征频率 TF_{SD}；l 表示语义域的词特征总数；i 为 (x,y) 在依存树中所处的位置。

测定主题在语义空间各结构中的概率分布，假定存在任意报道对，SDIDM 以语义域为媒介通过关联检测判断报道对主题的一致性。假设待测报道 D 的主题为 T，SDIDM 通过统计 T 的语义空间在 D 各语义域内的概率分布，其公式如下：

$$P(t\mid T)=\sum_{r\in\mathbf{R}_r}\boldsymbol{P}(\boldsymbol{t}\mid\boldsymbol{r})\boldsymbol{P}(\boldsymbol{r}\mid\boldsymbol{D})\qquad(2)$$

式中：t 为主题 T 语义空间中的某一特征；r 为 D 中某一语义域。通过公式(2)可以推断出，如果一篇报道的语义空间在另一篇报道的语义域中分布频繁，可以假定这两篇报道是相关的或者具有延伸关系。一元特征的 $P(t\mid r)$ 根据 TFSD 和 IDF 判定，$P(t\mid r)$ 如式(1)。$P(t\mid D)$ 将语义域 r 作为主题的概率，式(3)对报道 D 中的语义域进行排序，公式如下：

$$P(r\mid D)=\frac{\mid r\mid}{\sum_{s\in r}loc(s)}\cdot\log\frac{size(r)}{coll.size}\qquad(3)$$

$\mid r\mid$ 是语义域 r 中句子的总数，$loc(s)$ 表示句子 s 在 D 中的位置，$size(r)$ 是 r 包含的特征数量；$coll.size$ 是报道 D 包含的特征总数。SDIDM 根据 $size(r)$ 和 $coll.size$ 进行主题相关性匹配选择排序最靠前的语义域作为主题 T 的描述，根据式(2)建立语义域语言模型；最后对报道中的语义相关性进行评估。

4　关联检测实验设计

为测试并验证语义域语言模型的应用效果，采用实验 LDC 对基于文本形式的中文语料进行评测，实验的验证目标为描述语义的误检率与漏检率。实验分别基于一元特征和依存对设计 SDIDM。实验分析 SDIDM 要考虑两个参数：依存对权重计算中的平滑因子 α；主题语义空间的特征个数 N[9]。在公式(1)中，语义的表述能力主要靠对数表达式在依存树中的位置关系决定，语义表达能力的差异主要靠平滑因子 α 来进行调节。α 值越大则予以表达的敏感性越小。取特征个数 N 数量相同，这样可以使相关性指标在标准下进行比较。取 $\alpha=0.2$ 和 $\theta=0.09$ 时主题规模的概率分布情况，横轴为主题包含的特征数 N；图 3 纵轴分别对应某一 N 值的报道数在语料中的(a)误检率与(b)漏检率。为保证检测结果的有效性，参考模型选取 TPIC 模型[10]与 SDIDM 模型进行比较分析。检测率曲线如图 3 所示。

训练过程以 10 为颗粒度逐步调整 N 值，忽略对应语义域中权重相对较低的特征，用权重较高的特征对其进行补充，并基于各报道共有的高权重特征建立话题模型，实验结果表明不管是误检率还是漏检率都得到了明显的降低，影响性能的主要因素仍是主题语义描述的不精确性，

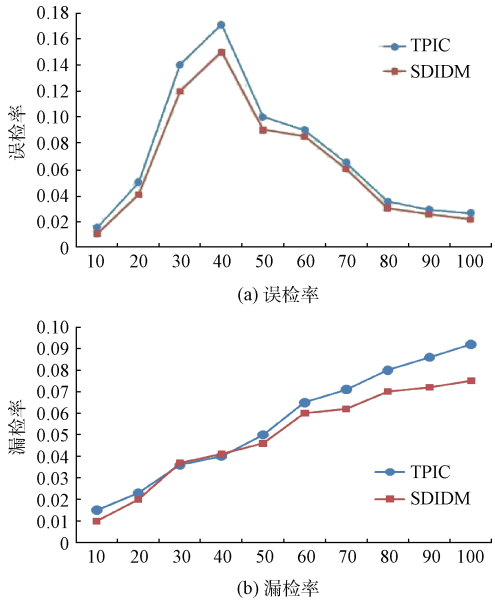

图 3　SDIDM 主题挖掘包含依存对的分布趋势

因此,经实验分析验证,本文设计的基于语义的信息时序检测模型能够有效提高对报道中语义空间中主题相关性检测的能力。

4　结　　论

针对 TD 领域中的研究现状,本文设计出了基于语义的信息时序检测模型,并实现了基于语义的相关性的概率估计。模型中事件的挖掘和描述对于建立更直观准确的话题模型具有重要意义。信息时序检测模型应用效果有效的验证不同语义结构的划分相关性判定的意义。本文尝试研究采用基于报道结构的划分为语言描述提供相关性运算帮助,为后续研究提供了理论基础。

参考文献

[1]　胡敏,罗珣,马韵洁. 基于语义矩阵反馈的多特征融合三维模型检索方法[J]. 电子测量与仪器学报,2012,26(4):325-330.

[2]　胡步发,王金伟. 双模态及语义知识的三维人脸表情识别方法[J]. 仪器仪表学报,2013,34(4):873-880.

[3]　洪宇. 基于语义结构和时序特征的话题检测与跟踪技术研究[D]. 哈尔滨:哈尔滨工业大学,2009.

[4]　曾维. 手势识别系统中手指及指尖检测方法[J]. 国外电子测量技术,2013,32(4):39-42.

[5]　霍玮,李丰,丁兆伟. 一种提高时序安全属性静态检测实用性的方法[J]. 计算机学报,2012,18(2):244-257.

[6]　丁兆云,周斌,贾焰,等. 微博中基于多关系网络的话题层次影响力分析[J]. 计算机研究与发展,2013,50(10):2155-2175.

[7]　曾依灵,许洪波,吴高巍. 一种基于语料特性的聚类算法[J]. 软件学报,2010,21(11):2802-2813.

[8]　金聪,金枢炜. 面向图像语义分类的视觉单词集成学习方法[J]. 电子测量技术,2012,35(8):53-56.

[9]　薛峰,周亚东,高峰. 一种突发性热点话题在线发现与跟踪方法[J]. 西安交通大学学报,2011,45(12):64-69.

[10]　杨攀,桂小林,田丰. 一种高效的用于话题检测的关键词元聚类方法[J]. 西安交通大学学报,2011,46(10):24-28.

作者简介

郑学伟,1979 年出生,计算机工学硕士学位,辽宁广播电视大学(辽宁装备制造职业技术学院)信息中心副主任、副教授,研究方向为知识管理、语义网技术。

E-mail:41969142@qq.com