

文章编号:1001-9081(2011)01-0093-04

doi:10.3724/SP.J.1087.2011.00093

基于语义相似度的论坛话题追踪方法

席耀一,林琛,李弼程,周杰,许旭阳

(信息工程大学 信息工程学院, 郑州 450002)

(Brian3333@163.com)

摘要:现有的话题追踪方法大多面向新闻数据,将其应用于论坛时效果不够理想。结合论坛的特点,提出一种基于语义相似度的论坛话题追踪方法。该方法首先通过构建话题和帖子的关键词表建立其文本表示模型,然后利用知网计算两个关键词表的语义相似度并以此作为帖子与话题的相关程度,最后根据相关程度实现论坛话题追踪。该方法较好地避免了向量空间模型的缺陷。实验表明,该方法能比较有效地解决面向论坛的话题追踪问题。

关键词:话题追踪;论坛;关键词;语义相似度;向量空间模型

中图分类号: TP391 **文献标志码:**A

Method for BBS topic tracking based on semantic similarity

XI Yao-yi, LIN Chen, LI Bi-cheng, ZHOU Jie, XU Xu-yang

(Institute of Information Engineering, Information Engineering University, Zhengzhou Henan 450002, China)

Abstract: To study the BBS topic tracking, the paper discovered that most of the traditional methods of topic tracking deal with news reports, and they are not suitable when they are applied to BBS. The paper utilized the characteristics of BBS and presented a topic tracking method for BBS data based on semantic similarity. This method firstly constructed keywords tables of topic and post as their representation models, and then computed the two tables' semantic similarity with the help of HowNet which is served as correlation degree between post and topic. Finally, this method used the correlation degree to realize BBS-oriented topic tracking. This method effectively avoids the disadvantage of Vector Space Model (VSM). The experimental results show that this method can solve the problem of BBS-oriented topic tracking effectively.

Key words: topic tracking; BBS; key word; semantic similarity; Vector Space Model (VSM)

0 引言

随着互联网的发展,传统新闻媒体与论坛间的议题互动日益频繁^[1]。往往当新闻媒体刚刚开始对某一话题进行报道时,网民就会在论坛上对该话题展开热烈讨论并发表大量与之相关的帖子。因此,有效地追踪论坛上与某话题相关的帖子可以帮助政府部门及时掌握网民关于某一话题的看法,或帮助企业及时了解网民关于某一产品的评价等。因此,研究面向论坛的话题追踪具有较大意义。

解决传统话题追踪任务(Traditional Topic Tracking, TTT)的代表性方法可以分为两类:一类是使用分类策略,利用先验相关报道训练构建话题模型,进行有指导的学习,如 K-最近邻(K-Nearest Neighbour, KNN)、决策树(Decision tree, D-tree)等;另一类是基于信息检索的方法,包括向量检索和概率检索。无论哪一种方法,都要解决以下关键问题:话题和报道(帖子)的文本表示模型、话题与报道(帖子)间的相似度计算模型,而相似度计算模型又与文本表示模型相关。

传统研究在构建文本表示模型时,最常使用的是向量空间模型(Vector Space Model, VSM)。虽然利用VSM表示新闻报道时能取得不错的效果,但是在表示帖子时效果并不好。因为它存在如下缺陷:1)忽视了特征自身携带的语义信息,将特征纯粹地看成符号,机械地比较特征之间形式上的相似性;2)无法充分体现特征词地位的不平等。由于向量空间模

型中的向量维数较高,重要的特征词经常会被区分能力较弱的特征词所淹没。

针对VSM的第一种缺陷,Juha Makkonen^[2]与焦健等人^[3]分别提出了解决办法。前者通过构建语义类别来表示话题,并给出了计算语义类别间相似度的方法;后者则利用知网的知识系统解决了近义动词的情况,但是没有解决名词、命名实体等的近似情况。针对VSM的第二种缺陷,Wei Zheng等人^[4]提出利用文本摘要技术构建关键词依存轮廓文件来表示话题,并通过计算轮廓间的相似度进行话题追踪。任晓东等人^[5]提出使用命名实体组成的话题类中心来代替VSM。这种方法将命名实体这类区分能力比较强的特征筛选了出来,较好地解决了VSM的第二个缺陷。但是在计算话题和报道的相似度时,只利用了类中心里相同的词,并没有考虑不同词之间的语义相似度。林鸿飞等人^[6]提出了一种基于语义框架的话题追踪方法。该方法对新闻报道进行了一定的语义分析,将新闻报道分解为四个问题:什么人、什么时间、什么地点、发生了什么事,并分别建立了对应的槽,槽中的值为描述相应问题所用的关键词。该方法的优点在于将区分能力强的关键词置于语义框架下,不再像VSM那样简单地罗列特征项,较好地解决了VSM的第二个缺陷。但在计算话题和报道的框架相似度时,该方法只利用了框架中相同的词,并没有考虑不同词之间的语义相似度。

在使用VSM表示论坛发帖数据时,VSM的缺陷比较突

收稿日期:2010-06-07 ;修回日期:2010-07-16。 基金项目:国家863计划项目(2007AA01Z439)。

作者简介:席耀一(1987-),男,河南洛阳人,硕士研究生,主要研究方向:网络新闻检测与追踪;林琛(1981-),女,山东威海人,博士研究生,主要研究方向:网络数据挖掘;李弼程(1970-),男,湖南衡阳人,教授,博士生导师,主要研究方向:智能信息处理;周杰(1984-),男,湖北武汉人,硕士研究生,主要研究方向:文本情感分析;许旭阳(1985-),男,河南商丘人,硕士研究生,主要研究方向:中文信息抽取。

出。因为论坛发帖数据不同于新闻报道,新闻报道数据具有“长文本”特性,而论坛发帖数据具有不同的特性。1)大量帖子属于短文本。短文本中词的个数少,描述话题的特征非常稀疏。而且网络发帖本身有很大的随意性^[7],作者对同一个事物可能有多种描述,这些描述语义上是完全近似的,但是在形式上可能有很大不同。如果只机械地比较特征之间形式上的相似性而忽略其语义信息,很容易造成大量帖子漏检。2)VSM 中会包含大量与话题本身无关的词语,因此很容易造成大量无关帖子被追踪到,影响系统的准确率。

因此,本文提出了一种基于语义相似度的论坛话题追踪方法。该方法首先通过构建话题关键词表和帖子关键词表作为话题和帖子的文本表示模型,然后计算两个关键词表的语义相似度并以此作为帖子与话题的相关程度,实现话题追踪。通过构建关键词表,可以较好地解决 VSM 的第二种缺陷,将无关以及不重要的特征去除掉;通过词汇语义相似度计算,可以较好地解决 VSM 的第一种缺陷,能够充分利用帖子为数不多的特征的语义信息。实验结果表明该方法能够较有效地解决论坛话题追踪问题。

1 话题和帖子的文本表示模型

本文通过构建关键词表来表示话题和帖子。选取的关键词应该是关于该话题的重要特征描述,既能反映话题主要内容,又能把不同话题区分开来。

1.1 话题的文本表示模型

NIST(美国国家标准与技术研究院)规定用 N 篇 ($N = 1, 2, 4$) 初始报道构建话题的初始模型。对于话题的初始相关报道,直接使用常用的 TF-IDF 加权方法来选取关键词,效果并不理想。本文采用文献[8]中提出的 ATF × PDF(Average Term Frequency-Proportional Document Frequency) 方法,每个词语 i 的权重 w_i 计算式如下:

$$w_i = \frac{\sum_{j=1}^N tf_{ji}'}{N} e^{n_j/N} \quad (1)$$

其中: N 为文档集包含的文档数, n_i 为文档集中包含词 i 的文档数, n 为第 j 个文档的词表大小, tf_{ji}' 表示词语 i 在文档 j 中的归一化词频,由式(2)计算。

$$tf_{ji}' = tf_{ji} / \sqrt{\sum_{i=1}^n tf_{ji}^2} \quad (2)$$

tf_{ji} 为词语 i 在文档 j 中的词频。由于文档集合中各个文档的长度不等,文档越长词语在文档中出现的次数概率越大,因此通过归一化来降低文档长度对词频的影响。

由于虚词没有实际意义,而实词中名词、动词、命名实体即可很好地表示话题的主要内容,因此本文只从文档的名词、动词、命名实体中选取关键词。

1.2 帖子的文本表示模型

关键词一般会在文本的标题及段首等位置出现,且可能不止一次出现。另外,邱立坤等在文献[9]中将论坛话题分为三类:事件性话题、褒贬性话题和讨论性话题。无论是哪类话题,网民都有可能从各个角度来描述话题,因此使用的词语也会不同。但是其描述的中心都会围绕该话题中的中心词语,而中心词语大多是命名实体且属于关键词。

考虑到以上两点,结合帖子的短文本特性,本文提出如下关键词提取方法。

- 1) 提取帖子中的所有特征,主要是名词和动词。这样可

以将短文本为数不多的特征保留下来,避免关键词的遗漏。

2) 对步骤 1) 中得到的所有特征进行过滤,过滤掉同时满足下述三个条件的特征:①非命名实体;②未出现在标题和段落前三分之一处;③词频为 1。

通过过滤,可以在保留大量关键特征的同时过滤掉大量与主题无关的特征,既提高了算法的效率,又能保证算法的性能。

2 话题与帖子之间的相似度计算模型

传统话题追踪方法多采用 VSM 表示话题,因此采用的相似度计算模型也多是余弦相似度模型。本文在计算相似度时主要考虑了话题与帖子之间的语义相似度。由于话题与帖子是用关键词表示的,因此首先研究词语之间的语义相似度计算,然后计算两个关键词表的语义相似度,并以此作为话题与帖子之间的相似度。

2.1 词语的语义相似度计算

刘群等人^[10]提出利用知网计算词语语义相似度,但是对于知网未收录的词语,例如大部分命名实体等并没有给出计算其语义相似度的方法。因此,本文既借鉴文献[10]中的方法计算知网收录词语的语义相似度,又给出了知网未收录词语的语义相似度计算方法。

2.1.1 知网收录词语的语义相似度计算

知网是一个以汉语和英语的词汇所代表的概念为描述对象的常识知识库,它对于词汇的语义描述具有明显的结构化特征。在知网中,词汇语义的描述被定义为义项(概念),每一个词可以表达为几个义项。义项又是由一种知识表示语言来描述的,这种知识表示语言所用的词汇称做义原。义原是从所有汉语词汇中提炼出的可以用来描述其他词汇的不可分割的基本元素。

刘群等人所提出的计算词汇语义相似度的方法主要是利用知网的结构性特征,将词汇的语义相似度计算层层分解。由于词汇的语义要用义项(概念)来描述,而义项又由义原来描述,因此可以将两个词汇的语义相似度计算首先分解为义项(概念)相似度计算,然后再将义项(概念)相似度计算分解为义原相似度计算,实现问题的简化。

设定两个词汇 w_1 和 w_2 ,如果 w_1 有 n 个义项, w_2 有 m 个义项,那么规定 w_1 与 w_2 之间的相似度是各个义项的相似度最大值,即

$$\text{Sim}(w_1, w_2) = \max_{i=1,2,\dots,n, j=1,2,\dots,m} \text{SimWS}(s_{1i}, s_{2j}) \quad (3)$$

其中: $\text{Sim}(w_1, w_2)$ 表示两个词汇的相似度, $\text{SimWS}(s_{1i}, s_{2j})$ 表示两义项的相似度, s_{1i}, s_{2j} 分别表示相应词汇的义项。

$$\text{SimWS}(s_1, s_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i \text{SimWP}_j(p_1, p_2) \quad (4)$$

其中: $\text{SimWS}(s_1, s_2)$ 表示两义项的相似度, $\text{SimWP}_j(p_1, p_2)$ 表示两义原的相似度; $\beta_i (1 \leq i \leq 4)$ 是可调节的参数,且有 $\sum_{i=1}^4 \beta_i = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。

$$\text{SimWP}(p_1, p_2) = \frac{a}{d + a} \quad (5)$$

其中: $\text{SimWP}(p_1, p_2)$ 表示两个义原的相似度, p_1 和 p_2 表示两个义原; d 是 p_1 和 p_2 在义原层次体系中的路径长度,是一个正整数,这里作为两个义原的距离; a 是一个可调节的参数,其含义是相似度为 0.5 时的路径长度。

2.1.2 知网未收录词语的语义相似度计算

对于知网中未收录的词语,本文给出了如下方法计算词语间语义相似度。

1) 对于两个人名。

如果其中一个仅有姓氏,则认为两者的相似度很大。例如“周久耕”和“周”等,“周正龙”和“老周”等。

如果其中一个省略姓氏,则认为两者的相似度很大。例如“范跑跑”和“跑跑”等。

如果其中一个为姓氏加职务,或姓氏加称呼,则认为两者的相似度很大。例如“周久耕”和“周局长”,“范跑跑”和“范先生”等。

2) 对于两个地名。

知网中一般只收录市一级及其以上行政区域的名称,对于县区级及其以下的行政区域,知网并没有收录。对此,本文建立一张未收录地名表,该表包括可能在某一话题中出现但未被知网收录的地名。例如“南京”和“江宁”,其中“江宁”未在知网中收录。但江宁是南京市的一个区,两者之间有较大的相似度。

另外,对于收录的行政区域,一般也只收录名称,并不附加诸如市、县、区、村等后缀名。对此,本文自动对地名的后缀名进行识别,完成该类命名实体的语义相似度计算。如知网中收录了“南京”,但未收录“南京市”,本文认为这两者的相似度为1。

3) 对于两个组织机构名。

如果如果组织机构名中存在全称和简称或简称和简称关系,则认为两者的相似度较大,如“三鹿”和“三鹿集团”等。

4) 其他情况。

如果两个词相同,认为其语义相似度为1。如果不同,首先利用以上规则进行判定,若不满足以上规则,认为其相似度为0。

2.2 话题关键词表与帖子关键词表间的相似度计算

本文以 $\text{sim}(\text{TB}_{\text{topic}}, \text{TB}_{\text{post}})$ 表示帖子关键词表与话题关键词表的相似度。其中, TB_{topic} 表示话题关键词表, TB_{post} 表示帖子关键词表。

在 TB_{topic} 已经构建的前提下,由于帖子的长短不一造成 TB_{post} 的长度 N 或者比 TB_{topic} 的长度 M 大或者比 M 小。为此,本文借鉴信息融合中的顺序加权思想,采用以下方法来计算 $\text{sim}(\text{TB}_{\text{topic}}, \text{TB}_{\text{post}})$ 。

情形①:当 $M \leq N$ 时,按照式(6),(7)计算两个词表的相似度,

$$\text{sim}(\text{TB}_{\text{topic}}, \text{TB}_{\text{post}}) = \sum_{i=1}^M \text{MaxSim}_i \cdot \text{Weight}_i \quad (6)$$

$$\text{MaxSim}_i = \max \{ \text{sim}(w_i, c_1), \text{sim}(w_i, c_2), \dots, \text{sim}(w_i, c_N) \} \quad (7)$$

其中: w_i 与 c_j 分别表示 TB_{topic} 与 TB_{post} 中的关键词; $\text{sim}(w_i, c_j)$ 表示两个词的语义相似度; Weight_i 表示 TB_{topic} 中词 w_i 的权重。通过加权,可以充分体现关键词的重要程度。

情形①适合于较长帖子的判断,该方法并没有采取归一化,因为归一化会削弱重要关键词的区分能力。另外,情形①将所有关键词均纳入计算,可以充分利用关键词表的所有特征。

情形②:当 $M > N$ 时,对于 TB_{post} 中的每一个词,分别计算它与 TB_{topic} 中的词的语义相似度并取最大值 MaxSim_j ,同时用 TB_{topic} 中对应的词的权重 Weight_i 进行加权, MaxSim_j 计算

如下:

$$\text{MaxSim}_j = \max \{ \text{sim}(w_1, c_j), \text{sim}(w_2, c_j), \dots, \text{sim}(w_M, c_j) \}, j \in [1, N] \quad (8)$$

然后统计满足 $\text{MaxSim}_j \geq 0.4$ 的个数 Num 。最后按式(9)计算两个词表的相似度:

$$\text{sim}(\text{TB}_{\text{topic}}, \text{TB}_{\text{post}}) = \begin{cases} \frac{\sum_{k=1}^{\text{Num}} \text{MaxSim}_k \cdot \text{Weight}_i}{\text{Num}}, & \text{Num} \neq 0, \text{MaxSim}_k \geq 0.4 \\ 0, & \text{Num} = 0 \end{cases} \quad (9)$$

情形②适合于较短帖子的判断,该类帖子特征非常稀疏,通常只有一两个关键词与关键词表中的词语相似。因此该方法只将 MaxSim_j 大于0.4的词语加入词表相似度的计算并采取了归一化措施(本文认为当两个词语的语义相似度大于0.4时,这两个词语的语义很接近)。这样能够将真正与关键词相似的特征提取出来,避免了无关特征削弱关键词对话题区分的作用。

3 基于语义相似度的论坛话题追踪

本文以帖子关键词表与话题关键词表的相似度作为帖子与话题的相关程度。即

$$\mu_{\text{topic}, \text{post}} = \text{sim}(\text{TB}_{\text{topic}}, \text{TB}_{\text{post}}) \quad (10)$$

其中: $\mu_{\text{topic}, \text{post}}$ 表示帖子与话题的相关程度。当帖子与话题的相关程度大于一定阈值时,认为帖子与话题相关,从而实现话题追踪。具体算法流程如下。

- 1) 利用初始相关报道构建话题关键词表,长度为 M 。
- 2) 对于每一篇到来的帖子构建帖子关键词表,长度为 N 。

3) 若 $M \leq N$,则利用式(6)计算两个关键词表的相似度得到帖子与话题的相关程度 $\mu_{\text{topic}, \text{post}}$ 。同时给定阈值 A ,若 $\mu_{\text{topic}, \text{post}} \geq A$,判断该帖子为相关帖,反之则不相关。

若 $M > N$,则利用式(9)计算两个关键词表的相似度得到帖子与话题的相关程度 $\mu_{\text{topic}, \text{post}}$ 。同时给定阈值 B ,若 $\mu_{\text{topic}, \text{post}} \geq B$,判断该帖子为相关帖,反之则不相关。

4 实验结果及分析

本文选取社会上影响较大的四个话题进行追踪,如表1所示。实验数据如下:

- 1) 与各话题相关的初始相关新闻报道;
- 2) 从各大论坛、贴吧(包括强国论坛,网易论坛,百度贴吧等)上下载的帖子,有与各话题相关的,也有不相关的,不相关的帖子涉及军事、体育、经济、娱乐等各个领域。

本文采用传统的关于追踪系统评价的评测标准,即使用召回率和准确率以及两者综合指标(*F1-Measure*)来评价该追踪系统的性能。具体如下:

$$\text{召回率 } R = \frac{a}{a + c}$$

$$\text{准确率 } P = \frac{a}{a + b}$$

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

其中: a 为追踪到的相关帖子数, b 为追踪到的不相关帖子数, c 为未追踪到的相关帖子数。

本文使用基于 VSM 的传统话题追踪方法与前面所提出的方法进行对比,结果如表 1 所示。

表 1 本文和 BaseLine 实验结果

话题	相关帖子数	不相关帖子数	基于 VSM 的 TTT 方法			本文的方法		
			召回率/%	准确率/%	F1-Measure/%	召回率/%	准确率/%	F1-Measure/%
范跑跑事件	90	600	84.44	73.79	78.76	88.89	98.77	93.57
周久耕事件	95	600	93.68	90.82	92.23	89.47	95.51	92.39
三鹿奶粉事件	95	600	96.84	32.86	49.07	86.32	87.23	86.77
周老虎事件	99	600	96.84	85.19	90.64	90.91	90.91	90.91

本实验中,由于部分话题所包含的帖子中描述话题的特征与新闻报道比较一致,因此造成对于部分话题,基于 VSM 的 TTT 方法召回率比较高,但是与本文的方法相比仍稍差。尤其在准确率上,基于 VSM 的 TTT 方法性能不是很稳定,对数据的依赖性比较强。经过分析实验结果发现,使用基于 VSM 的 TTT 方法追踪到的无关贴确实是因为话题模型中的大量无关特征造成的。另外,对比两者的综合评价指标 *F1-Measure*,本文的方法也优于基于 VSM 的 TTT 方法。

5 结语

本文利用论坛帖子的特点研究了论坛话题追踪,较好地解决了帖子短文本特性对话题追踪的影响。本文方法对话题关键词表的准确度要求较高,当话题重心发生漂移时须更新相关报道并重新构建关键词表,否则可能造成后续帖子的漏检。此外,本文对分词效果要求较高,需进行必要的未登录词识别。未来的工作主要集中在以下几点:首先,在现有方法的基础上加入对话题重心漂移的研究;其次,通过获取话题关键词,实现对话题在论坛上的发展演变趋势的分析。

参考文献:

- [1] 陈映. BBS 与主流报纸的议程互动[D]. 广州: 暨南大学, 2005.
- [2] MAKKONEN J. Semantic classes in topic detection and tracking [D]. Helsinki: Department of Computer Science, University of Hel-

(上接第 88 页)

- [2] PARSOPoulos K E, VRAHATIS M N. Particle swarm optimization method for constrained optimization problems[C]// Proceedings of the 2nd Euro-International Symposium on Computational Intelligence. Slovakia: IOS, 2002: 214–220.
- [3] VENTER G, HAFTKA R T. Constrained particle swarm optimization using a bi-objective formulation[J]. Structural and Multidisciplinary Optimization, 2010, 40(1/6):65–76.
- [4] CHOOTINAN P, CHEN A. Constraint handling in genetic algorithms using a gradient-based repair method[J]. Computers and Operations Research, 2006, 33(8):2263–2281.
- [5] ZAHARA E, KAO Y T. Hybrid Nelder-Mead simplex search and particle swarm optimization for constrained engineering design problems[J]. Expert Systems with Applications, 2009, 36(2): 3880–3886.
- [6] DEB K. An efficient constraint handling method for genetic algorithms[J]. Computer Methods in Applied Mechanics and Engineering, 2000, 186(2/4):311–318.
- [7] LIU HUI, CAI ZIXING, WANG YONG. Hybridizing particle swarm optimization with differential evolution for constrained numerical and engineering optimization[J]. Applied Soft Computing, 2010, 10(2):629–640.
- [8] HE QIE, WANG LING. A hybrid particle swarm optimization with a feasibility-based rule for constrained optimization[J]. Applied Mathematics and Computation, 2007, 186(2):1407–1422.
- [9] KEANE A J. Experiences with optimizers in structural design[EB/OL].[2010-02-10]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.50.7516>.
- [10] KANNAN B K, KRAMER S N. An augmented lagrange multiplier based method for mixed integer discrete continuous optimization and its applications to mechanical design[J]. Journal of Mechanical Design, 1994, 116(2): 405–411.
- [11] RAO S S. Engineering optimization[M]. 4th ed. Hoboken: John Wiley and Sons, 2009.
- [12] ARORA J S. Introduction to Optimum design[M]. 2nd ed. San Diego: Elsevier Academic Press, 2004.
- [13] HE QIE, WANG LING. An effective co-evolutionary particle swarm optimization for constrained engineering design problems[J]. Engineering Applications of Artificial Intelligence, 2007, 20(1): 89–99.
- [14] HUANG FUZHUO, WANG LING, HE QIE. An effective co-evolutionary differential evolution for constrained optimization[J]. Applied Mathematics and Computation, 2007, 186(1):340–356.