基于SVM的中文报道关系识别方法研究

王 强[1],张永奎[2]
WANG Qiang[1],ZHANG Yong-kui[2]

1.山西大学 计算机与信息技术学院,太原 030006
2.计算智能与中文信息处理省部共建教育部重点实验室,太原 030006
1.Department of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2.Key Lab of Ministry of Education for Computation Intelligence and Chinese Information Processing,Taiyuan 030006,China
E-mail:wangqiang25719@163.com

WANG Qiang,ZHANG Yong-kui.Research on Chinese story link detection based on SVM.Computer Engineering and Applications,2008,44(33):141-143.

Abstract: Via analyzing the characteristic of news in the Web,construct the feature vector using features from five entity categories:persons,time,location,organizations,and content.Using story time and entity relatedness for temporal or place vector when calculating their similarity and cosine similarity for others.All the features together with the entity relatedness are integrated by Support Vector Machine(SVM).Experimental results show that this method can improve system performance effectively.
Key words: story link detection;topic detection and tracking;multi-vector mode

摘 要:针对网络新闻的特点,从人名、时间名、地点名、组织机构名、内容五个方面抽取特征词形成特征向量。在此基础上,分别进行了相似度计算,其中,人名、组织机构名、内容采用余弦夹角的方法,时间和地点向量,相似度计算采用了引入报道时间和关联度计算。最后,使用这5个相似度作为特征,使用SVM进行训练,并在测试集上进行了测试。测试结果表明,这种方法可以有效地改善系统的性能。
关键词:报道关系识别;话题检测与跟踪;多向量表示模型
DOI:10.3778/j.issn.1002-8331.2008.33.044 **文章编号:**1002-8331(2008)33-0141-03 **文献标识码:**A **中图分类号:**TP391

1 引言

随着互联网信息的膨胀,人们很难从众多信息中快捷地获取自己需要的信息。话题检测与跟踪(Topic Detection and Tracking,TDT[1],作为一项旨在帮助人们应对信息过载问题的研究,以新闻专线、广播、电视等媒体信息流为处理对象,将语言形式的信息流分割为不同的新闻报道,监控对新话题的报道,并将涉及某个话题的报道组织起来以某种方式呈现给用户。报道关系识别(story link detection)是TDT的一个子任务,目标是判断两篇随机的报道(story)是否描述了同一个话题(topic)。这里的话题是指发生在特定时间、地点的一个核心事件或活动,以及所有与之相关的事件或活动。该任务被认为是TDT其他子任务例如新事件检测、话题跟踪的核心技术。

目前已有很多机器学习的算法应用到报道关系识别系统中,主要分为两类:基于向量空间模型的方法和基于概率模型的方法。两者各有优缺点,向量空间模型的局限在于其独立性假设,即向量特征之间是相互独立的。而概率模型由于新闻报道简短精练,使模型原本就有的数据稀疏问题更加严重。Lavrenko V,Allan J[2]等人提出话题检测与跟踪的相关性模型,使用概率的方法获得相关性文档。Francine Chen,Ayman Farahat[3]等人提出使用报道的消息来源和采用不同的相似度计算方法计算报道对的相似度值,然后共同作为特征项来训练支持向量机的方法进行识别。

采用多向量的空间模型来表示文本,根据语义特征将特征词划分为5类(人名、时间名、地点名、组织机构名、内容),每一类构成一个子向量。对每个子向量进行独立的相似度计算,把这些相似度值一起作为SVM的特征输入,进而对报道间的关系进行训练和测试。实验表明该方法可以有效地改进系统的性能。

2 多向量文本表示模型的构建

预处理中,每篇新闻报道都进行分词、词性标注、停用词过滤、之后将获得的用于向量空间模型表示的特征候选集合、特征的频率以及报道的长度在预处理之后统计得出。

2.1 特征抽取与划分

对文本进行特征抽取并划分是建立多向量模型的第一步。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60475022);山西省自然科学基金(the Natural Science Foundation of Shanxi Province of China under Grant No.20041041);山西省回国留学人员基金(No.2002004)。
作者简介:王强(1983-),男,硕士研究生,主要研究方向:中文信息处理、话题检测与跟踪;张永奎(1945-),男,教授,博士生导师,主要研究方向:人工智能、中文信息处理。
收稿日期:2008-07-02 **修回日期:**2008-09-26

多向量模型的特征抽取方法和单向量模型的方法一样,对报道进行分词、词性标注及停用词过滤后,以词及其词性标记为特征,对同一个字串,如果标记为不同词性,则认为是不同特征。然后根据词性从候选特征集合中抽取出人名、时间名、地点名、组织机构名。候选特征集剩余的词表示成另外的内容子向量,用 5 个子向量表示选出的这 5 类特征子集构成本文的多向量文本表示模型。

2.2 特征权重的计算

文中所有子向量的特征都采用改进的 tf-idf 来计算,向量的特征是切分后的词加上其词性标记,对于同一个词,如果标记为不同词性,则认为是不同特征。向量中的每一维表示该特征在报道中的权重。每个新闻报道 d_i 表示成一个范化的特征向量:

$$v(d_i)=(t_1,w_1(d_i);t_k,w_i(d_i);\cdots t_n,w_n(d_i))$$

其中 t_k 为特征项,$w_k(d_i)$为 t_k 在文档 d_i 中的权值。

传统的 tf-idf 的计算方法如公式(1)、(2)、(3)所示:

$$w_k(d_i)=tf\times idf \tag{1}$$

$$tf=n/(n+0.5+1.5dl/dl_{avg}) \tag{2}$$

$$idf=\log(N+0.5)/df)/\log(N+1) \tag{3}$$

其中,n 是特征在报道中出现的次数,dl 是报道的长度,dl_{avg} 是背景语料(已处理过的报道对所处的源文件及同时段源文件集合)中文档的平均长度,N 代表背景语料中所有报道的个数,df 是背景语料中出现该特征的报道个数。tf 词频代表了特征能够反映报道内容信息的程度,idf 是集合中逆文本词频的对数值,它可用来去掉那些经常在报道中出现的不代表主要意义的字和词。

改进的 tf-idf 计算主要是引入了位置信息,论文仅考虑了报道的标题和报道开头的第一句话,公式如下:

$$tf=n\times f\_loc/(n\times f\_loc+0.5+1.5dl/dl_{avg}) \tag{4}$$

$$f\_loc=\begin{cases}\alpha(\text{特征 } t \text{ 出现在标题和报道的首句中})\\ 1(\text{特征 } t \text{ 出现在其他的位置})\end{cases}$$

α 是经验值取 1.3。

2.3 多向量文本表示模型

在分析新闻结构特点的基础上,采用了基于语义特征分组的多向量空间模型。每个语义组由语义相近的词组成,如人名、时间名、地点名等。将一篇新闻文档用五个独立的向量模型来表示。实验结果表明[4]:在表示一篇新闻报道时,如果在模型转换的过程中有信息丢失,那么一定会造成系统检测代价提高和性能下降:表示模型在信息含量不变的情况下,信息区分的越细越好,并且每类信息间的相似度计算方法以及多个信息相似度整合的方法也会对系统性能有很大的影响。

多向量文本表示模型的内容包括:一、对文本进行特征抽取,然后按照某种划分方式把抽取出的特征集合划分为多个不相交的子集,每个子集由一个向量表示;二、模型之间对应子向量的相似度计算;三、对多个相似度进行整合,从而判断两个模型之间的相似性。

3 向量的相似度计算

在相似度计算这部分,对于不同的子向量采用不同的相似度计算方法。对于人名、组织机构名、内容词构成的子向量采用余弦相似度的计算方法。时间向量和地点向量采用独立的相似度计算方法。

3.1 余弦相似度计算

两文档的余弦相似度计算公式为:

$$\mathrm{Sim}(\alpha,\beta)=\frac{\sum_{i=1}^{t}W_{i,\alpha}\times W_{i,\beta}}{\sqrt{\sum_{i=1}^{t}W_{i,\alpha}^2\times\sum_{j=1}^{t}W_{j,\beta}^2}} \tag{5}$$

其中 α 和 β 分别是两个待测的向量空间模型,$W_{i,\alpha}$ 和 $W_{j,\beta}$ 为两向量空间模型中词的权重,$\mathrm{Sim}(\alpha,\beta)$为 α 与 β 的相似度。

3.2 基于时间距离的相似度计算模型

在计算时间相似度的时候考虑到每篇报道都有报道时间,同时报道时间可以有效的提高区分度,所以引入报道时间距离。报道时间的格式:2008 年 06 月 27 日 09:51,对报道时间首先进行规范化和形式化,都写成以上的时间形式,时间有秒的话也舍弃,考虑到精确到秒影响漏检率,然后形式化为 200806270951:

$$distime(a,b)=|time_a-time_b| \tag{6}$$

$time_a$ 为报道 a 的报道时间,$time_b$ 为报道 b 的报道时间。得到时间距离后,使用下面公式计算两文档时间的相似度:

$$score_{a,b}=\mathrm{sim}(\alpha,\beta)+\theta distime(a,b)$$

θ 为经验值取 1.5,sim (α,β) 为报道 a,b 对应时间向量 α 和 β 的余弦相似度值。

3.3 地点相似度计算

对地点向量进行相似度计算时考虑到精确匹配可能导致关系的丢失,比如说是山西和太原。所以采用关联度[4]来计算地点向量之间的相似度。即:

$$relate(entity1,entity2)=\frac{wei(entity1)\times wei(entity2)\times cooccur(entity1,entity2)}{sioccur(entity1)\times sioccur(entity2)} \tag{7}$$

$$sioccur(entity1)=occur(entity1)-cooccur(entity1,entity2) \tag{8}$$

$$sioccur(entity2)=occur(entity2)-cooccur(entity1,entity2) \tag{9}$$

其中 $wei(entity)$指该实体词在子向量中的特征权重,$cooccur(entity1,entity2)$指两个实体词在背景语料中共同出现过的报道个数;$sioccur(entity)$是相对于另一个实体词来说该实体词在背景语料中单独出现的报道个数;$occur(entity)$是背景语料中包含该实体词的报道个数。有了实体词关联信息之后,命名实体子向量之间的关联度则是两个向量全关联后所有关联度的平均值,如公式(10)所示:

$$Vecrelate(entityVector_1,entityVector_2)=\frac{1}{size1\times size2}\sum_{1\leqslant i\leqslant size_1,1\leqslant j\leqslant size_2}relate(entityVector_{1i},entityVector_{2j}) \tag{10}$$

其中,$entityVector_1$,$entityVector_2$ 是经过压缩后的向量表示,即去除了权重为零的特征,$size$ 即是特征向量中权重不为零的特征个数,$entityVector_{1i}$ 是第一个向量中的第 i 个特征,$entityVector_{2j}$ 是第二个向量中的第 j 个特征。

4 基于 SVM 的报道关系识别

将报道表示成 5 个向量表示模型之后,用多相似度的计算方法计算出对应子向量之间相似度,那么如何将这些相似度整合起来得到两文本间的相似度又是一个非常关键的步骤。

采用一种新的机器学习的算法,即支持向量机学习器(SVM)。实验证明比基于其他的例如决策树的方法在这个系统上有很好的性能。在多向量模型中使用SVM进行多相似度的整合,就是以多个相似度构成的向量<link/not linked,sim1,sim2,…sim5>作为输入特征进行训练,得到训练模型然后自动地判断两篇新的新闻报道对是否相关。

5　实验结果与分析

5.1　评测语料

目前已经在互联网上收集了2000年~2007年的3 000余篇突发事件新闻,给出了15个待测话题,共有1 386个报道对,其中900个用于训练,486用于测试。

5.2　评测公式

采用TDT2003的评测方法对系统性能进行评价。TDT建立了完整的评测体系,评测标准是利用系统漏检率和错检率计算损耗代价(C_{Det})Norm作为系统的评价指标,此值越小则系统性能越好。TDT评测公式定义如下:

$$C_{Det}=C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot P_{non-target} \quad (11)$$

检测开销通常被归一化为0和1之间:

$$(C_{Det})_{Norm}=\frac{C_{Det}}{\min(C_{Miss}P_{target},C_{FA}P_{non-target}} \quad (12)$$

其中C_{MISS}和C_{FA}分别代表漏检率和错检率的代价系数,在TDT2003中C_{MISS}和C_{FA}分别取1和0.1;P_{MISS}和P_{FA}分别是系统漏检率和错检率;P_{target}和$P_{non-target}$是先验目标概率($P_{non-target}=1-P_{target}$),取0.02。

5.3　实验结果分析

为验证上述方法在报道关系识别中的有效性,共实现了三个报道关系识别系统:基准系统,多向量表示模型的系统,改进的tf-idf结合多相似度计算的多向量表示模型系统。他们根据单个相似度或多相似度整合值对报道对之间的话题相关性作出判断。其中基准系统采用单向量表示模型和余弦相似度计算方法,首先在训练样本中获取使系统性能达到最优的阈值,用该阈值评测基准系统在测试样本中的性能;基于多向量表示模型的系统采用了5个子向量来表示文本,对应子向量采用余弦相似度的方法和SVM多值整合方法;在此基础上采用多相似度的计算方法,即对时间向量和地点向量采用不同的相似度计算方法。表1中是这三种方法下报道关系识别系统的评测值。

表1　三种方法在各项指标下的评测值

评价指标	基准系统	基于多向量模型的系统	提出的系统
P_{miss}	0.038 2	0.042 5	0.040 1
P_{fa}	0.015 2	0.007 8	0.007 2
C_{det}	0.002 3	0.001 6	0.001 5
$(C_{det})_{norm}$	0.115 0	0.080 0	0.075 0

从实验结果可以看出,后两个报道关系系统在$(C_{det})_{norm}$都比基准系统下降了不少,但是从单个指标来看,多向量模型表示使得系统的漏检的概率有所增加,经分析可能是在按照语义特征将特征向量划分为多向量模型的过程中,割裂了各个子向量特征间的关系,进而导致漏检率上有所增加。

6　总结

在分析新闻文档的基础上,根据语义特征将特征词划分为五类,构建了一个新的多向量文本表示模型。同时在计算特征权重的时候考虑到标题和报道开头的一句话对文章之间的区分度比较大,引入了位置信息。最后在进行相似度的时候,采用多相似度的计算方法,即针对不同子向量采用不同的相似度计算方法。实验结果表明,上述方法有效地改进了系统的性能。将来的工作主要是在计算报道关系的时候考虑报道源特征,因为同样的事件不同的网站和作者的用词风格都有区别。

参考文献:

[1] 李保利,俞士汶.话题识别与跟踪研究[J].计算机工程与应用,2003,39(17):7-10.

[2] Lavrenko V,Allan J,DeGuzman E,et al.Relevance models for topic detection and tracking[C]//Proceedings of Human Language Technologies Conference,HLT,2002:104-110.

[3] Chen F,Farahat A,Brants T.Multiple measures and source-pair information in story link detection[C]//Proceedings of HLT-NAACL,2004:313-320.

[4] 张晓艳,王挺,陈火旺.基于多向量和实体模糊匹配的话题关联识别[J].中文信息学报,2008,22(1):9-14.

[5] 宋丹,王卫东,陈英.基于改进向量空间模型的话题识别与跟踪[J].计算机技术与发展,2006,16(9).

[6] Farahat A,Chen F,brants T.Optimizing story link detection is not equivalent to optimizing new event detection[C]//Proceedings of ACL,2003:232-239.

(上接109页)

[3] Lee W B,Wu C C,Tsaur W J.A novel deniable authentication protocol using generalized ElGamal signature scheme[J].Information Sciences,2007,177:1376-1381.

[4] Dwork C,Naor M,Sahai A.Concurrent zero-knowledge[J].Journal of the ACM,2004,51(6):851-898.

[5] Aumann Y,Rabin M.Efficient deniable authentication of long messages[EB/OL].Hongkong:City University of Hongkong,(1998)[2006-10-11].http://www.cs.cityu.edu.hk/dept/ video.html.

[6] Fan L,Xu C X,Li J H.Deniable authentication protocol based on Diffie-Hellman algorithm[J].Electronics Letters,2002,38(4):705-706.

[7] Lu R X,Cao Z F.A new deniable authentication protocol from bilinear pairings[J].Applied Mathematics and Computation,2005,168(2):954-961.

[8] Lu R X,Cao Z F.Non-interactive deniable authentication protocol based on factoring[J].Computer Standards & Interfaces,2005,27(4):401-405.

[9] Joux A.A one-round protocol for tripartite Diffie-Hellman[C]//Bosma W.LNCS 1838:Proceedings of ANTS IV 2000.Algorithm Number Theory Symposium-ANTS-IV,Leiden,The Netherlands,July 2-7,2000.Berlin:Springer-Verlag,2000:385-394.