基于隐马尔科夫模型的语义倾向性研究

章栋兵,姚寒冰,颜 昕
(武汉理工大学 计算机学院,湖北 武汉 430063)

摘 要:以网络评论为研究对象,试图把隐马尔科夫模型从已经成功应用的模式识别领域推广到语义倾向性分析系统。与传统倾向性识别系统不同的是,此理论通过建立隐马尔科夫分类模型,将未知文本进行状态序列化,得到文本中所有的词语所对应的倾向性,然后选定多数词的倾向性来作为文本的总体语义倾向。实验表明,当训练数据越全面、规模越大时,识别率越高。

关键词:语义倾向性;隐马尔科夫模型;序列化

中图分类号:TP391.1 **文献标识码**:A **文章编号**:1674-7720(2010)17-0071-03

Research of semantic orientation based on hidden Markov models

ZHANG Dong Bing, YAO Han Bing, YAN Xin
(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: Regarding network comments as the research object, this paper tries popularizing the hidden Markov model from the pattern recognition domain that has already been successfully applied to the semantic orientation analysis system. Different from the traditional orientation recognition system, this theory serializes the state of the unknown texts through the establishment of a hidden Markov model. After getting the corresponding tendencies of all the words in the text, it then chooses the one of the majority as the general tendency of the text semantic. Experiments show that the more comprehensive the training data is and the larger the scale is, the better recognition there will be.

Key words: semantic orientation; hidden Markov model; serialization

网络媒体被公认为是继报纸、广播、电视之后的“第四媒体”,成为反映社会舆情的主要载体之一。人们希望能快速高效地在浩如烟海的网络信息中提取对于诸如人物、事件、传媒、产品等有价值的评价信息。如何有效地提取文本信息,推断其语义倾向,已经成为当前自然语言与信息安全研究领域的热点问题[1]。

当前流行的语义倾向性分析系统可以分为两个步骤:首先是识别词汇的语义(短语)倾向性[2],然后利用不同的策略根据词汇(短语)的倾向性给出整个文本的语义倾向评价。目前主要有三种研究思路:(1)对所有词汇的倾向性评分进行统计求和,根据最终的得分正负来评价文本的倾向性[3]。(2)采用机器学习的方式根据词汇的倾向性训练出语义倾向分类器[4],这是目前比较流行的思路,总体效果比统计求和要好。这两种思路是基于概率统计的,领域性限制小。(3)基于“格语法”分析的思路。该思路很难全面反应样本空间规律,具有一定的领域限制性。

本文利用隐马尔科夫模型 HMM(Hidden Markov Models)在文本处理方面的优势,首先对其理论进行介绍,然后根据现有学者对 HMM 在文本分类中的应用和文本分类技术在倾向性分析中应用的研究结果,提出将 HMM 应用于文本倾向性研究的理论,并用实验证明此理论的可行性。

1 理论基础

1.1 隐马尔科夫模型

隐马尔可夫模型[5]作为一种统计模型,非常适合处理时变信号,用于动态过程时间序列建模并具有强大的时序模式分类能力,理论上可处理任意长度的时序。HMM 是一个双重随机过程,其中之一是 Markov 链,其基本随机过程为描述状态的转移;另一个随机过程描述状态与观察值之间的统计对应关系,只能看到观察值,而不能看到状态,即通过一个随机过程去感知状态的存在

及其特性。

1.2 HMM 在文本分类中的应用

罗双虎[6]把待分类文本描述成一系列状态演化的隐 Markov 过程，其中状态以特定的概率产生代表文本的特征项。用序列模式来描述文本类，文本序列通过与隐 Markov 模型的匹配，求出其对应状态序列和最大输出概率，以比较各个文本类的结果，达到文本分类的目的。

龙丽君[7]对关键字所在的句子构成的词序列建立 HMM，以判断句子所属的类别。为了建立 HMM，将词语所属的类别理解为状态，将所选择的关键字理解为输出值。这样就把要判定一个观测序列(一个句子)的整体所属的类别转换为已知模型和观测序列，求出全局最优的整体序列。观测序列的整体所属类别即为关键字所属类别，或者说观测序列的整体类别即为状态序列中居多数的状态对应的类别。

1.3 文本分类技术在倾向性分析中的应用

1997 年，Hatzivassiloglou 和 McKeown 尝试使用监督学习的方法对词语进行语义倾向判别，通过对训练语料的学习进行语义倾向判别，准确率约 82%，在加入篇章中形容词之间的接续信息后，准确率提升到约 90%[2]。2003 年，Turney 在其论文[8]中提出了利用统计信息对单词进行语义倾向判断的新方法。文本的语义倾向判别也可被看作一个褒贬的分类问题，因此，文本分类中的方法同样被应用到了语义倾向判别研究中。

2 HMM 在语义倾向性研究的应用

本文是针对网络评论，判断其表达的是支持(褒义)、反对(贬义)还是中立(中性)的语义倾向性。

2.1 建立模型

本文所定义的 HMM 初始参数如下：

(1)N 为模型中 Markov 链中的状态数目，即所定义的词语类别的数目，由于倾向只有支持(褒义)、反对(贬义)和中立(中性)3 个类别，因此 N 的值为 3。记 N 个状态为 S_1, S_2, S_3。t 时刻 Markov 链所处状态 q_t，显然 $q_t \in \{S_1, S_2, S_3\}$。

(2)M 为每个状态对应的可能的观测值数目，即要进行分类的词语的个数。由于词语数量比较庞大，选出语料库中前 1 000 个权重最高的词语，每个句子只选择这些词序列作为一个观测符号序列。记 1 000 个观测值为 $V_1, V_2, V_3, \cdots, V_{1\,000}$，$t$ 时刻观测到的观测值为 O_t，其中 $O_t \in \{V_1, V_2, V_3, \cdots, V_{1\,000}\}$。

(3)π 为初始状态概率矢量，即初始时每个类别的概率经过统计之后的结果。对所选的全部文档而言，对应上文中的三个类别，π_i 的值分别为 0.15、0.15、0.7，其中 $\pi_i = P(q_1 = S_i)$，$1 \leqslant i \leqslant 3$。

(4)A 为状态转移概率矩阵，即从一种词语类别转移到另一种词语类别的概率。本文采用 Baum-Welch 算法[5]对初始模型进行训练，$A = \{a_{ij}\}$，其中 $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$，$1 \leqslant i, j \leqslant 3$。

(5)B 为观测值概率矩阵，即观测序列在状态 j 时接受某一观测符号的概率，$B = \{b_j(k)\}$，其中 $b_j(k) = P(O_t = V_k | q_t = S_j)$，$1 \leqslant j \leqslant 3$，$1 \leqslant k \leqslant M$。

2.2 实验系统框架

系统整体框架如图 1 所示，整个系统分为训练阶段和识别阶段。

2.2.1 语料库准备

训练语料库是国内还没有公开的文本倾向语料库。本实验全部由人工收集，然后对所提取的所有的句子进行分词、标注之后，去掉连词、助词和代词等不具倾向性的无用词，得到最终的语料库。

否定词表：带否定意义的词，如：不、不是、非等。

2.2.2 训练阶段

首先根据初始参数建立初始模型，然后使用 Baum-Welch 算法[5]对参数进行训练，得出最终分类模型。

2.2.3 识别阶段

将未知评论文本经预处理得到字串 $(W_1, W_2, W_3, \cdots, W_n)$ 作为上文中训练得到的 HMM 分类模型的观察序列，通过维特比(Viterbi)算法[5]得到最优状态序列 S，然后使用以下算法得出整个语句的语义倾向性，如图 2 所示。

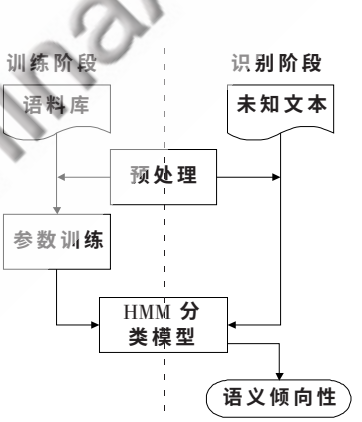

图 1 系统整体框架

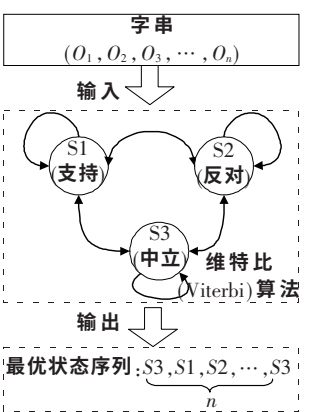

图 2 HMM 模型识别结构图

```
Array<Word> W;                          //字串
Array<State> S;                         //最优状态序列
Dictionary Deny;                        //否定词表
Integer Length;          //字串长度，即字串中词语的数目
Procedure getOrientation ()
  //句子倾向性为状态序列中具有倾向（非中立）的状
                          态占多数的状态所对应的倾向
  //由于网络评论中作者的倾向多数是在句首，取首个
具有语义倾向的状态对应的倾向为整个句子的语义倾向性
  Orientation orientation="中性";
```

```
Integer numP=0;//S1(支持)的数量
Integer numN=0;//S2(反对)的数量
Orientation firstOrientation;//记录句子中首个非中性的状态
For i ← 0 to Length-1 do
  If S[i]! =S3 then
    If i>0 and W[i-1] ∈ Deny then
      //此状态不是句首且此状态对应的观察值是否定词时
      //状态类别以相反类别计数
      S[i]==S1?numN++;numP++;
    Else
      S[i]==S1?numP++;numN++;
    End If
    If firstOrientation==NULL then
      firstOrientation =(S[i]== S1?"支持":"反对");
    End If
  End If
Repeat
If numP>numN then
  orientation="支持";
Else If numP<numN then
  orientation="反对";
Else
  orientation=firstOrientation;
End If
end getOrientation
```

2.3 应用举例

例句："我同意你的观点"。

经分词结果为："我/r 同意/v 你/r 的/u 观点/n"。去除无用词得到观察值序列为："同意/v 观点/n"，最后经过识别得出最优状态序列为：S_1，S_3。由于 S_1 出现 1 次，而没有出现 S_2，故这个句子的倾向性为 S_1 的倾向类别：支持。

3 实验结果及分析

实验文本是来自不同网站上下载的各种评论共 2 000 条，所有的评论都经过分词、标注和去无用词处理，然后手工分为：支持(褒义)、反对(贬义)和中立(中性)3 个类别。然后在每个类别中分别取 200、300、400、500 条，共 600、900、1 200、1 500 条作为本实验的训练数据，进行封闭测试并对剩余的评论进行开放测试。实验结果如表 1、表 2 所示。

表 1 各类别识别率(封闭实验)

总训练数据数/条	支持(褒义)/%	反对(贬义)/%	中立(中性)/%	平均/%
600	83.61	84.42	70.06	79.31
900	85.44	84.29	86.66	85.46
1200	84.67	85.92	80.35	83.65
1500	82.44	81.29	82.76	82.16

表 2 各类别识别率(开放实验)

总训练数据数/条	支持(褒义)/%	反对(贬义)/%	中立(中性)/%	平均/%
600	41.61	40.42	44.92	42.32
900	49.44	50.87	53.35	51.22
1 200	62.33	61.48	62.78	62.20
1 500	70.22	68.94	69.72	69.63

从表中结果可以看出，封闭测试可以达到很高的识别率，可见训练语料库的规模将直接影响分析结果。当语料更全面、覆盖面更广泛时，识别率将大大提高，因此建立一个良好的训练语料库的识别方法将有很好的应用前景。

本文从单个句子出发，研究其倾向性分析方法，从实验结果数据可以看出，此方法有很好的识别率，但需面对两个问题：(1)网络文本的复杂性：如语句的语气、具有倾向性的词语所针对不同的评价对象和网络新词的频繁出现等情况；(2)语料库的整理：语料库的完整性和准确性将直接影响分析方法的准确率，而国内还没有公开的文本倾向语料库。这些问题将做进一步地研究和改进。

参考文献

[1] 宋火尧，刘功申.基于主题相关性分析的文本倾向性研究[J].信息安全与通信保密，2009(3):77-78.

[2] HATZIVASSILOGLOU V, MEKEOWN K R. Predicting the semantic orientation of adjectives[A]. In: Proceedings of the 35th Annual Meeting of the Association for Computational Liguistics and the 8th Conference of the European Chapter of the ACL, 1997:174-181.

[3] PETER T. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[A]. In: Proceedings of the 40th Annual Meeting of the Association for Computational Liguistics, 2002.

[4] 徐琳宏，林鸿飞，杨志豪.基于语义理解的文本倾向性识别机制[J].中文信息学报，2007，21(01)：98-102.

[5] 宗成庆.统计自然语言处理[M].北京：清华大学出版社，2008.

[6] 罗双虎，欧阳为民.基于隐 Markov 模型的文本分类[J].计算机工程与应用，2007，43(30):179-181.

[7] 龙丽君.网络内容监管系统中基于局部信息的语义倾向性识别算法[D].南京.南京理工大学，2004.

[8] PETER T，MICHAEL L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003，21(4):315-346.

(收稿日期：2010-03-03)

作者简介：

章栋兵，男，1984 年生，硕士研究生，主要研究方向：信息检索。