第32卷 第20期 计 算 机 工 程 2006年10月
Vol.32 № 20 Computer Engineering October 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)20—0183—02 文献标识码：A 中图分类号：TP311

不同粒度下的文档分类

赵欣欣，朱铁丹，刘玉树
(北京理工大学信息科学技术学院计算机科学工程系，北京 100081)

摘 要：提出了句子空间模型及基于句子空间模型的分类算法。比较了从词、句子两个不同粒度对文档进行表示的向量空间模型和句子空间模型在对同一问题进行分类时的召回率和准确率。实验表明，与向量空间模型相比，句子空间模型在许多情况下具有较好的分类性能。
关键词：粒度；向量空间模型；句子空间模型

Document Classification in Different Granularity

ZHAO Xinxin, ZHU Tiedan, LIU Yushu
(Department of Computer Science & Engineering, School of Information Science & Technology, Beijing Institute of Technology, Beijing 100081)

【Abstract】This paper proposes a sentence space model(SSM)and a classification algorithm based on SSM. It compares the vector space model and the sentence space model in classifying the same document with the recall and the precision from different granularity, word granularity and sentence granularity. Experiments show SSM has a better classification performance than vector space model in many circumstances.
【Key words】 Granularity; Vector space model; Sentence space model

人类智能的一个公认特点，就是能从极不相同的粒度上观察和分析同一问题。人们不仅能在不同粒度的世界上进行问题求解，而且能够很快地从一个粒度世界跳到另一个粒度世界，往返自如，毫无困难。这种处理不同粒度世界的能力，正是人类问题求解的强有力的表现[1]。

张钹和张铃提出了信息粒度的概念[2]，并且对信息粒度这一概念作出了非常精辟和透彻的论述。信息粒度，是对信息和知识细化的不同层次的度量[6]，这一概念的提出，是因为人工智能和认知科学研究者们观察到人类智能的一个公认特点，那就是在认知和处理现实世界的问题时，常常采用从不同层次观察问题的策略，往往从极不相同的粒度上观察和分析同一问题。

词粒度上的模型——向量空间模型和句子粒度上的模型——句子空间模型就是从不同粒度上对文档进行表示，从而完成分类的两个模型。

1 向量空间模型

向量空间模型(Vector Space Model)是20世纪60年代末由GeraldSalton等人提出的。VSM是一种知识表示方法[5]，也是最早最出名的一个数学模型。在该模型中，文档被看作是由一组正交词条向量所组成的向量空间，每个文档D表示为其中的一个范化特征向量D=(t_1,w_1(D)；t_2,w_2(D)；…；t_n,w_n(D))，其中t_i为词条项，w_i(d)为t_i在D中的权值[3]。t_i可以是D中出现的所有单词，也可以是D中出现的部分单词，w_i(d)一般被定义为t_i在D中的出现频率的函数。然后通过计算文本相似度的方法来确定待分类样本的类别。也就是说，向量空间模型是从词的粒度上表示文档，从而完成对文档的分类。

事实上，在实际的问题求解中，粒度的选择应该视具体问题而定。不同粒度的划分是为了研究、分析问题的方便。句子是表达完整语义的基本单位，据此，以句子为单位，也就是在句子这一粒度上表示文档，然后对文档进行分类。并提出了可以保证句子的完整性、不需要进行降维处理的句子空间模型。

2 句子空间模型

2.1 几个定义

给定一个 C 分类问题 $C=(C_1,C_2,...,C_k)$，N 个句子组成输入文本 $T=(S_1,S_2,...,S_N)$，n 个单词组成每个句子 $S_i=(w_{i1},w_{i2},...,w_{in})$。

句子的贡献度：称后验概率 $P(C_j|S_i)$ 为句子 S_i 对类别 C_j 的贡献度 x_{ij}。

句子的类向量：句子的类向量是指给定句子 S_i 的一个 $|C|$ 维向量表示，形如 $S_i=(x_{i1},x_{i2},...,x_{i|c|})$，其中，$x_{ij}$ 是 S_i 对类别 C_j 的贡献度。

句子－类矩阵：称输入文本 T 的 $N\times|C|$ 矩阵表示为 T 的句子－类矩阵，矩阵的每一行是一个句子的类向量，形如

$$T=\begin{pmatrix}S_1\\S_2\\\vdots\\S_N\end{pmatrix}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1|c|}\\x_{21}&x_{22}&\cdots&x_{2|c|}\\\vdots&\vdots&\ddots&\vdots\\x_{N1}&x_{N2}&\cdots&x_{N|c|}\end{pmatrix}$$

句子空间模型：称基于矩阵－类矩阵表示的文本分类模型为句子空间模型。

2.2 句子空间模型的分类算法

算法步骤如下：首先计算每个句子对各个类别的贡献度；然后将待分类文本表示成句子－类矩阵的形式；最后得到文本的类别判断。

算法描述如下：

输入：待分类的文档T，C分类问题

作者简介：赵欣欣(1978－)，女，博士生，主研方向：Web挖掘，个性化服务；朱铁丹，博士生；刘玉树，教授、博导
收稿日期：2005-10-18 **E-mail：**zhaoxxtm@bit.edu.cn

输出：分类结果

首先计算句子对C的贡献度x_{ij}。由贝叶斯公式，有

$$x_{ij} = P(C_j/S_i) = \frac{P(C_j)P(S_i/C_j)}{P(S_i)}$$

在独立性假设的前提下，有

$$x_{ij} = \frac{P(C_j)\cdot\prod_{k=1}^{n}P(w_{ik}/C_j)}{\prod_{k=1}^{n}P(w_{ik})} = \alpha\cdot P(C_j)\cdot\prod_{k=1}^{n}P(w_{ik}/C_j)$$

其中，α 是与类别C_j C_j 无关的常数。事实上，只需得到每个贡献度的相对大小，故不需要计算 $P(C_j|S_i)$ 的值，也就是说，计算x_{ij}等价于计算 $P(C_j)\cdot\prod_{k=1}^{n}P(w_{ik}/C_j)$ ，即计算句子的贡献度时，只需要从训练数据集中学习两组参数：先验概率 $P(C_j)$ 和类条件密度 $P(w_{ik}|C_j)$ 。

然后，取每个文档的前 N 个句子，求出这 N 个句子对每一类的贡献度，并将待分类文本表示为句子－类矩阵的形式，构造句子－类矩阵。

最后，进行文本的类别判断。进行类别判断的方法有合成法和投票法。合成法的主要思想是将矩阵的各行相加，求其均值，将这个均值作为文本的类向量，然后根据向量各维的数值大小判断文章的类别归属；而投票法是根据矩阵的每一行判断该句子的类别，对每个句子选择一个类别进行投票，最后将票数多的类别作为文本的类别。合成法和投票法均没有考虑上下文的信息。而实验中采用的是双投票法，即在句子 S_i 对各个类别的贡献度中，找出贡献度的最大值x_{ij1}和次大值 x_{ij2} ，其对应的类分别是 C_{j1} 和 C_{j2} ，计算权重 $weight=map(x_{ij_1}/x_{ij_2})$ ，其中，map是实数空间到整数空间[1,m]的映射(实验中，取m为 5)。若$weight$=1，则 S_i 对C_{j1}和C_{j2}各投一票；若$weight$>1，则 S_i 对C_{j1}投$weight$票。

句子空间模型中的双投票法步骤如下：

```
Begin Input T,C
Input T,C
初始化投票计数器 Vi=0，(i=1,2,…,|C|)
  Foreach sentence S in T
  Begin
    计算 S 对 C 中每个类别的贡献度
    取贡献度最大值 xi 和次大值 xj
    计算权值 weight=map(xi/xj)
    If (weight=1) then
        Vi=Vi+1； Vj=Vj+1
    Else
        Vi=Vi+weight；
  End
  Return Ci，(i=argmax Vi)
End
```

3 实验结果

实验分 6 组进行，每组有 5×100 篇文档(5 类，每类 100 篇)，进行 10 次实验。数据源是南加州大学的UCI资源库[4]。实验的目的是对比不同粒度下，即词粒度下的向量空间模型和句子粒度下的句子空间模型，对文档进行分类的准确率和召回率。表 1 列出了实验对比结果。从实验结果可以看出，向量空间模型在有的文本集上做得很好，而句子空间模型在大多数文本集上可以取得较好的分类效果。

表 1 不同粒度下的分类模型的实验数据对比

Newsgroups	向量空间模型		句子空间模型	
	准确率	召回率	准确率	召回率
comp.sys.ibm.pc.hardware	73.06%	84.33%	75.87%	89.17%
rec.autos	87.22%	75.67%	91.36%	83.17%
rec.sport.hockey	96.32%	93.70%	96.98%	95.60%
sci.space	75.75%	83.33%	81.52%	87.50%
talk.politics.misc	72.92%	92.50%	77.03%	92.17%
平均值	81.05%	85.91%	84.55%	89.52%

4 结论

对事物或物理过程的定性描述是问题求解的关键，也是人类智能的一种表现[1]。在一个复杂的环境中，从根本上说，很难掌握全部、完全的信息。本文比较了词粒度下的向量空间模型和句子粒度下的句子空间模型对文档分类的准确率和召回率。实验表明，与向量空间模型相比较，句子空间模型在许多情况下具有较好的分类性能。

但是，句子空间模型的分类算法尚有许多不足。首先，句子的贡献度是在条件独立性假设下计算的，但是条件独立性假设在大多数情况下是不成立的。如果能够做到贡献度的估计值的相对大小与真正贡献度的相对大小一致，如果能构造适当的函数进行相邻行的计算，则分类结果会更理想。下一步将应用自然语言理解的方法对文档进行句法分析，以得到更好的分类效果。也将考虑从段落这一粒度上对文档进行分类研究，进一步研究不同粒度下的文本分类。

参考文献

1 Hobbs J R. Granularity[C]. Proc. of IJCAI, Los Angeles, 1985: 432-435.

2 张 钹, 张 铃. 问题求解理论及应用[M]. 北京: 清华大学出版社, 1990.

3 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

4 Murphy P M, Aha D W. UCI Repository of Machine Learning Databases[Z]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1995.

5 Salton G. Developments in Automatic Text Retrieval[J]. Science, 1991, 253(5023): 974-979.

6 邵 健. 基于 Rough Sets 的信息粒度计算及其应用[D]. 北京: 中国科学院自动化研究所, 2000.

(上接第 176 页)

5 结束语

通过对自驱动安全管理策略模型的研究，我们开发了具有自主知识的安全产品，建立了以资源为主题的、以安全威胁所驱动的、以安全策略作实施的、以分析结果为评估的动态自适应的信息安全集成管理平台，应用于政府、企业等多个领域。对安全策略的研究将会更好地提高安全产品的易用性和实用性，更好地抵御攻击与防范安全风险，使信息化建设在一个更加稳固的基础之上。

参考文献

1 林东岱, 曹天杰. 企业信息系统安全——威胁与对策[M]. 北京: 电子工业出版社, 2004: 5-31.

2 漫谈网络安全策略[Z]. http://searchsecurity.techtarget.com.cn/tips/403/1889903.shtml.

3 李 铭. 互联网安全技术: 现状和未来—基于时间的安全防护体系[J]. 信息网络安全, 2002, 18(6).

4 申雅琴. P^{2}DR模型——网络安全管理的指南[J]. 微电脑世界, 2001,

31(1).