第33卷 第22期 计 算 机 工 程 2007年11月
Vol.33 No.22 Computer Engineering November 2007
·博士论文· 文章编号：1000—3428(2007)22—0045—03 文献标识码：A 中图分类号：TP391

基于句类向量空间模型的自动文本分类研究

张运良[1,2]，张 全[2]
(1. 中国科学院研究生院，北京 100039；2. 中国科学院声学研究所，北京 100080)

摘 要：向量空间模型是自动文本分类中成熟的文本表示模型，通常以词语或短语作为特征项，但这些特征项通常只能提供较少的局部语义信息。为实现基于内容的文本分类，该文用HNC理论中的句类作为特征项，通过混合句类分解等技术对句类向量空间降维，使用tfc算法对特征项进行权重计算，用KNN算法进行分类。该分类器的平均准确率和召回率都是可接受的，对类别的抽象程度无要求，即抽象度较高和较低的类别可以同时分类。通过使用更好的机器学习算法和其他的HNC语言理解技术，性能可以进一步提高。
关键词：文本分类；句类；向量空间模型；HNC理论

Research of Automatic Text Categorization Based on Sentence Category VSM

ZHANG Yun-liang[1,2], ZHANG Quan[2]
(1. Graduate School, Chinese Academy of Sciences, Beijing 100039; 2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080)

【Abstract】 Vector space model is a mature model of text representation in automatic text categorization. Words and phrases are commonly used as feature items, but these items provide little local semantic information. This paper uses sentence categories, which include more semantic information, as feature items. To reduce the dimensionality of sentence category vector space, it divides mixed sentence categories and reformes the weights by tfc algorithm and buildsthe classifier by KNN algorithm. The average precision and recall of the classifier are acceptable, especially in the case of categories having different abstraction. The performance can be improved by HNC techniques and machine learning algorithm.
【Key words】 text classification; sentence category; vector space model (VSM); HNC theory

随着社会信息化程度加深，特别是互联网的发展，电子文档越来越多。若仅依靠人工处理海量的电子文档，无法保证处理的速度和一致性，所以电子文档的自动处理势在必行。而对文档进行自动分类往往是处理的首要工作，文本自动分类技术同时也能为信息检索、问答系统等提供支持。

向量空间模型(VSM)[1]是最常用的文本表示方法，但随着对文本分类要求的提高，研究者发现，单纯的以词语、汉字或短语为特征项的文本表示方法，难以提供复杂的语义信息，无法满足基于内容和理解的文本处理的需求。获取语句级句子语义的信息将消解部分词汇模糊性，能为文本分类提供更加详尽和准确的语义信息。句类理论就是黄曾阳研究员在HNC(概念层次网络)理论框架内提出的语句级信息表述和获取方式[2]，该理论历经近20年理论探索和工程实践检验，成为本文的基础之一。VSM和很多成熟的机器学习算法已经广泛应用到文本分类领域。本文尝试利用HNC理论获得高语义内涵的文本句类表示，再结合VSM，利用机器学习算法形成分类器。

1 向量空间模型及文本分类应用

向量空间模型是20世纪60年代由G. Slaton提出的，被广泛应用于文本分类、信息检索等应用的文本表示。VSM的基本思想是将文本离散化，处理为某种特称项表示的向量。常用的特征项有词、短语、术语[3]等，也有用汉字的。VSM具有表达简明、处理容易、一致性好的特点。

使用VSM进行文本分类的基本步骤包括：文本预处理，向量生成，向量降维，分类计算[4]。文本预处理阶段将文本离散化，同时除去对分类贡献较小的特征项。向量生成主要根据文本中各特征项的出现频率及不同的权重方法生成向量。目前主要的权重方法有布尔权重、特征项频率权重、tfidf权重、tfc权重、熵权重等。向量降维能够减少计算的时间复杂度和空间复杂度。目前降维主要有两种思路：一种是利用特征选择技术将一些对分类区分度不大的项去掉；另一种是将类似的项以某种方法归并。特征选择可以使用的方法包括阈值选择、信息增益、互信息、最大熵、CHI等方法[5~6]；归并方法可以利用某些语言本体，如wordnet[7]、知网(Hownet)、同义词词林等，也可以利用某些统计算法，如潜在语义索引(LSI)[8]。机器学习方法用于获取各个类别在向量空间模型下的表示，并确定一个新文本的类别所属。目前主要有Rocchio算法、Naïve Bayes方法、KNN算法、决策树方法、人工神经网络方法等[9]。

2 HNC理论句类知识介绍

句类是HNC理论的重要概念之一，HNC理论认为语句无限而语句的类型是有限，并提出57组共307种基本句类，任何一个语句都可以归结为基本句类及其混合[10]。句类是由语义块构成的，一个实例句类表示式如式(1)所示。根据对句类表达的关联性，可以将语义块分为主语义块和辅语义块，主

基金项目：国家“973”计划基金资助项目“自然语言理解的交互引擎研究”(2004CB318104)；中科院声学所知识创新工程资助项目
作者简介：张运良(1979－)，男，博士研究生，主研方向：文本分类，HNC理论；张 全，研究员、博士生导师
收稿日期：2006-11-30 **E-mail：**yunlzhang@mails.gucas.ac.cn

语义块是语句表达所必需的，即使在语句中不出现，也可以通过上下文补足，而辅语义块是可选的。主语义块又分为特征语义块和广义对象语义块(构成要素中包含作用者A，对象B和内容C)，一个句类对应的主语义块是确定的。如式(1)所示的信息转移句T31J由 3 个语义块构成：TA表示信息转移的发出者；T表示信息转移这一活动；T3C表示信息转移的内容。句类提纲擎领地表达了句义，如式(2)、式(3)两个句子，在词汇层面上不同，但从内容上看都是T31J。

T31=TA+T+ [#T3C#] (1)

李晓明 ‖ 宣布 ‖ [# 傅老师 ‖ 重返 ‖ 科院课堂#]。

T31J[#T3C#]=[#T2b0J#] (2)

周济 ‖ 指出 ‖ [# 教学评估 ‖ 是 ‖ \{ 提高 | 教育质量 } 的关键 /#]。

T31J[#T3C#]=[#jDJ#DC=\{!31XY401*211J}/#] (3)

句类突破了文本的表层形式，深入到文本内部，因而能够提供更多语义的信息，更能代表文本的内容，有更利于发现同类文本之间的共性。

混合句类是语言表述的客观需要和客观存在，如上例中的提高教学质量这一句，句类代码 XY401*211J 是由作用句 XJ 和一般效应句 Y401J 构成的。3 个句类的句类表示式分别如式(4)～式(6)所示。

XJ=A+X+B (4)

Y401J=YBC+Y (5)

XY401*211J=A+XY401+YBC (6)

其中，A，X，B分别表示作用句类的作用者、具体作用、作用对象；YBC表示效应的对象和内容，如“ 教育质量 ”中“ 教育 ”对象“ 质量 ”是内容，Y表示效应；“ * ”是混合句类的标记，后面有 3 位数字取得的广义对象语义块的数量及其来源，与理论HNC相应的句类分析技术和分析平台也在不断发展[11]。

3 句类向量空间模型

句类向量空间模型是在HNC句类理论基础上，利用向量空间模型的思想而产生的文本分类的模型，句类向量空间表示如式(7)所示，其中SC_j是句类向量空间中第i个句类。

$$V=(SC_1,SC_2,\cdots,SC_n) \quad i\in N,\ 1\le i\le n \tag{7}$$

基本句类有 57 组，共 307 种，两两混合的句类有P(307,2)= 93 942 种，而考虑 3 种以上的基本句类混合则总句类数远远超过 10 万种，因此必须对句类向量空间进行降维。本文的策略是将混合句类归结到基本句类，凡是出现混合句类，按照混合句类中出现的基本句类平均分配出现频度，再进行权重计算。在 HNC 理论中，不同层次的句类重要性不同，但为了计算的简便，也受限于目前不同句类重要性的量化困难，目前对不同位置出现的句类不加以区分。

4 分类器相关算法

本文权重计算使用tfc算法，其基本思想是在tfidf算法的基础上，进行归一化，以充分消除不同长度文本的影响。文档d_j 在句类句类向量空间$\boldsymbol{V}$下可以用向量 $\boldsymbol{d}_j$表示如式(8)所示，其中a_{ij}是文档j使用tfc权重算法的句类向量空间上第i维度上的表示，计算方法如式(9)。式(9)中TF_{ij} 是句类i在文档j中出现的等效次数，N 是文本集合中的文本总数， DF_i是出现句类i的文档数量。

$$\boldsymbol{d}_j=(a_{ij},a_{2j},\cdots,a_{nj}) \tag{8}$$

$$a_{ij}=\frac{\log(TF_{ij}+1.0)*\log(N/DF_i)}{\sqrt{\sum_k[\log(TF_{kj}+1.0)*\log(N/DF_k)]^2}} \tag{9}$$

本文分类算法使用 KNN，该算法的基本思想是：判断在训练文本集中与待判定文本距离最近的 K 个文本，再根据这 K 个文本所属的类别确定待判定文本所属的类别，具体的算法如下：

Step 1 对训练集合文档进行句类分析，获得训练文本向量集合；

Step 2 根据待判定文本句类分析结果，确定其向量表示；

Step 3 在训练文本集中选出与待判定文本最相似的K个文本，这里用余弦向量距离计算，如式(10)所示，其中，a_{ik}表示向量$\boldsymbol{d}_i$特征向量的第k维；

$$\mathrm{sim}(\boldsymbol{d}_i,\boldsymbol{d}_j)=\frac{\sum_{k=1}^{n}a_{ik}\times a_{jk}}{\sqrt{(\sum_{k=1}^{n}a_{ik}^2)(\sum_{k=1}^{n}a_{jk}^2)}} \tag{10}$$

Step 4 在待判定文本的K个邻居中，依次计算每类的权重，计算式如(11)，其中，$\boldsymbol{d}_x$为待判定文本的特征向量，$sim(\boldsymbol{d}_x,\boldsymbol{d}_i)$ 为相似度计算公式，与上一步骤的计算公式相同，而 $p(\boldsymbol{d}_i,C_j)$ 为类别属性函数，即，如果$\boldsymbol{d}_i$属于类 C_j ，那么函数值为 1，否则为 0；

$$p(\boldsymbol{x},C_j)=\sum_{\boldsymbol{d}_i\in KNN}sim(\boldsymbol{d}_x,\boldsymbol{d}_i)\,p(\boldsymbol{d}_i,C_j) \tag{11}$$

Step 5 比较各类的权重，将文本分到权重最大的那个类别中。

5 分类器的表现

本文使用来自互联网的 1 610 篇新闻语料(语料 A)及其子集(语料 B，1 078 篇)来测试分类器的性能，单篇语料长度为 600 字～5 000 字之间。语料 A 被分为 15 个类别，包括 10 个高层类别，如政治、经济等，同时包括 5 个高层类别的特定子类，如政治领域的选举活动、经济领域的农村问题等。语料 B 仅有 10 个高层次类别。测试包括两语料上的封闭测试及开放测试，开放测试用 80%的语料作为训练集，另 20%的语料作为测试集。

通常使用精确率和召回率来评价分类器的表现[12]，对每一类别c_i的精确率和召回率计算公式分别如式(12)、式(13)所示，其中，N_{cci}表示被正确分到c_i类别的文档数量；N_{cti}表示分类器分到c_i类别的文档数量；N_{tci}表示应该被分到c_i类别的文档数量。

$$precision(c_i)=\frac{N_{cci}}{N_{cti}} \tag{12}$$

$$recall(c_i)=\frac{N_{cci}}{N_{tci}} \tag{13}$$

对分类器评价，本文采用宏平均的精确率和召回率，即先求出每个类别的精确率和召回率，然后算术平均，其计算方法如式(14)、式(15)所示，分类器的测试表现如表 1 所示。

$$precision_a=\frac{1}{m}\sum_{i=1}^{m}precision(c_i) \tag{14}$$

$$recall_a=\frac{1}{m}\sum_{i=1}^{m}recall(c_i) \tag{15}$$

表 1 分类器的测试表现

测试类型		封闭测试/(%)	开放测试/(%)
语料 A	$Pricision_A$	86.58	72.99
	$Recall_A$	88.27	74.31
语料 B	$Pricision_B$	89.42	80.16
	$Recall_B$	91.78	75.43

6 问题讨论

本节将对 3 个影响分类器性能的重要问题加以分析和讨论。

一个长句中，很可能出现多个句类。那么这相互关联的多个句类是否应该赋予不同的权重呢？事实上，这是一个复杂的问题，本文的简化处理方法是赋予它们相等的权重。由于句类特征语义块存在E_g-E_l，E_p-E_r，E_1-E_2等 3 种关系，因此句类之间的关系也必然是这 3 种关系及其复合。研究此问题需要充分研究句蜕、块扩和复句。

关于混合句类如何向基本句类转化，使得句类向量空间降维，也是一个值得继续探讨的问题，目前的处理方法是将混合句类进行简单平均分配到基本句类。针对此问题，有两种不同观点：(1)混合句类包含更加复杂的信息；(2)混合句类和基本句类包含同等程度的信息。两种观点的差别在于混合句类总信息量的多少，但无论哪一观点，都需要确定每一基本句类的具体分配的权重。

自动的句类分析技术仍然在不断完善，文本分类是一项在不完全句类分析技术下可行的应用，因为人工分类也不可能对全文包含各个细节都有完全的理解和掌握，很多时候仅从文档的部分章节就可以确定其所属的类别。同样，本文文本分类中使用的句类分析技术，可以进一步完善，但是分析到什么程度就不会影响本文的分类；句类分析系统会给出一些不能确定的分析结果，此种情况下是将可能分析错误的结果置之不理还是以某种特定概率分布算法进行判定，以上种种，都需要进一步研究。

7 结论和下一步工作

句类向量空间模型利用 HNC 句类理论提供较完整、较深入的语义信息。本文使用该模型进行文本表示并形成一种分类器。本分类器在降维中使用了混合句类向基本句类转化的方法，在权重计算上使用 tfc 算法，能够消除文本长度的影响；在训练和分类中使用 KNN 算法。本分类方法在初步的测试中获得了基本可以接受的结果，而且能够在分类类别抽象度不同的情况下，仍有较好的分类表现，因而本分类方法可以用来获取某些用户订制的信息。

研究并解决在第 6 节中提到的问题，尝试不同的机器学习算法，并应用更多 HNC 语言理解技术，获得性能更好的分类器，将是下一步研究的目标。

参考文献

1 Salton G, Lesk M E. Computer Evaluation of Indexing and Text Processing[J]. Journal of the ACM, 1968 15(1): 8-36.

2 黄曾阳. HNC(概念层次网络)理论[M]. 北京: 清华大学出版社, 1998 .

3 庞剑锋,卜东波. 基于向量空模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001, 18(9): 23-26.

4 Aas K, Eikvil A. Text Categorization: A Survey[R]. Norwegian Computing Center, Technical Report: #941, 1999.

5 Yang Yiming, Pedersen J O. A Comparative Study on Feature Selection in Text categorization[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997: 412-420.

6 周 茜, 赵明生. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17-23.

7 张 剑, 李春平. 基于 WordNet 概念向量空间模型的文本分类[J]. 计算机工程与应用, 2006, 42(4): 174-178.

8 王天江, 叶卫国. LSI 和 kNN 相结合的文本分类模型研究[J]. 华中科技大学学报(自然科学版), 2004, 32(4): 59-60, 86.

9 Shi Yongfeng, Zhao Yanping. Comparison of Text Categorization Algorithms[J]. 武汉大学学报(自然科学版), 2004, 9(5): 798-804 .

10 苗传江. HNC 句类知识研究[D]. 北京: 中国科学院声学研究所, 2001.

11 韦向峰. 基于 HNC 理论的扩展句类分析平台研究[D]. 北京: 中国科学院声学研究所, 2005.

12 宋枫溪, 高 林. 文本分类器性能评估指标[J]. 计算机工程, 2004, 30(13) : 107-109, 127.

(上接第 41 页)

参考文献

1 Mobasseri B G. Digital Modulation Classification Using Constellation Shape[J]. Signal Processing, 2000, 80(2): 251-277.

2 周 明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 2002-05.

3 李敏强, 寇纪淞, 林 丹, 等. 遗传算法的基本理论与应用[M]. 北京: 科学出版社, 2002-03.

4 Duda R O, Hart P E, Stork D G. 模式分类[M]. 李宏东, 姚天翔, 译. 北京: 机械工业出版社, 2003-09.

5 徐 勇, 荆 涛. 神经网络模式识别及其实现[M]. 北京: 电子工业出版社, 1999.

6 Wu Y X, Ge L D, Liu F F. Comprehensive Features Based Digital Modulation Identification Using a Neural Tree Network[C] //Proceeding of 2005 International Conference on Communications, Circuits and Systems. 2005, 2: 748-752,.

7 喻寿益, 郭观七. 一种改善遗传算法全局搜索性能的小生境技术[J]. 信息与控制, 2001, 30(6): 526-530.

(上接第 44 页)

试验表明，BBMML 算法能在较少的样本上获得满意的学习效果。在经过 100 个循环的训练后根据 Q 表基本能找到最优路。标准的 Q 学习需要到 500 步之后才能使 Q 估计值接近真实值。这是因为 BBMML 算法通过开关函数协调了彼此的样本学习路径，避免了无效的 Q 表更新。

参考文献

1 Wyatt J. Exploration and Inference in Learning from Reinforcement[D]. Department of Artificial Intelligence, University of Edinburgh, UK, 1997: 33-34.

2 罗 清, 李智军, 吕恬生. 复杂环境中的多智能体强化学习[J]. 上海交通大学学报, 2002, 36(3): 224-230.

3 杜春侠, 高 云, 张 文. 多智能体系统中具有先验知识的 Q 学习算法[J]. 清华大学学报(自然科学版), 2005, 45(7): 443-447.

4 Sutton R. Learning to Predict by the Methods of Temporal Differences [J]. Machine Learning, 1988, 3(1): 9-44

5 于存贵, 自 勇, 马志文, 等. 基于黑板模型的多属性决策模式[J]. 南京理工大学学报, 2004, 24(4): 334-339.

6 韩 伟. 电子市场环境下的多智能体学习与协商[D]. 上海: 华东师范大学, 2006-04.