文章编号：1003-0077(2013)02-0065-09

词性对中英文文本聚类的影响研究

韩 普[1]，王东波[1]，刘艳云[2]，苏新宁[1]

(1. 南京大学 信息管理学院，江苏 南京 210093；2. 解放军理工大学 指挥自动化学院，江苏 南京 210007)

摘 要：不同词性特征在文本聚类中有不同的贡献度。该文对四组有代表性的中英文数据集，利用三种聚类算法验证了四种主要词性及其组合对中英文文本聚类的影响。实验结果表明，在中文和英文两种语言中，名词均是表征文本内容的最重要词性，动词、形容词和副词均对文本聚类结果有帮助，仅选择名词作为特征聚类的结果与保留所有词性聚类的结果相近，但可大大降低文本的维度；选用名词为文本特征不能实现最好的聚类效果；相对其他词性组合和单一词性，采用名词、动词、形容词和副词的组合特征往往可以实现更好的聚类效果。在词性所占的比例以及单一词性聚类的结果上，同一词性在中英文文本聚类中呈现出较大差异。相对于英文，不同词性特征及其组合在中文文本聚类中呈现的差异更为稳定。

关键词：词性标注；文本聚类；文本特征

中图分类号：TP391 **文献标识码**：A

Influence of Part-of-Speech on Chinese and English Document Clustering

HAN Pu[1], WANG Dongbo[1], LIU Yanyun[2], SU Xinning[1]

(1. School of Information Management, Nanjing University, Nanjing, Jiangsu 210093, China;
2. Institute of Command Automation, the PLA University of Technology & Science, Nanjing, Jiangsu 210007, China)

Abstract: Different part-of-speeches have different roles in document clustering. Using 4 popular English and Chinese datasets, the paper choose three clustering algorithms to investigate the influence of 4 major part-of-speeches as well as their combination on Chinese and English document clustering. The experimental result reveals that nouns are the most important in presenting the content of the document. Besides, verbs, adjectives and adverbs contribute to document clustering. Although similar result is obtained from the experiments, nouns. Using only nouns to characterize the document can not produce the best clustering result, but it can reduce the document dimensions to a great extent. The combination of 4 part-of-speeches produces the best clustering result. Single part-of-speech vary considerably in Chinese and English document clustering performance, and the differences are more consistent in Chinese document clustering.

Key words: part of speech tagging; document clustering; text feature

1 引言

通常认为，不同的词性在文本中发挥着不同作用，承担不同角色，重要度也不一样，例如，名词的重要性大于动词，动词的重要性大于副词。从语言学角度看，词性的变化，可以使语言表达更多信息，不同词性在文本内容表达上的功能是不同的，在句法结构中承担着不同角色。在文本处理时，选择重要角色的词性作为特征不但可以提高效率，还可能会

收稿日期：2011-11-11 **定稿日期**：2012-07-16

基金项目：863 计划项目“科技文献服务为主的搜索引擎研制”(2011AA01A206)；2011 年南京大学研究生科研创新基金资助项目“中英双语文本聚类技术及其应用研究”(2011CW12)

作者简介：韩普(1983—)，男，博士研究生，主要研究方向为信息处理、信息分析；王东波(1981—)，男，博士研究生，主要研究方向为自然语言处理与文本挖掘；刘艳云(1983—)，女，硕士，助教，主要研究方向为信息处理。

提升处理的效果。词性标注是自然语言处理中进行词性分析的一项基础工作，利用最大熵、条件随机场、SVM等算法[1-4]，该技术已经比较成熟，目前已在信息检索、自然语言处理、文本分类聚类等领域得到了广泛应用。

苏祺等[5]采用TREC数据集研究了词性标注对信息检索的影响，认为词性标注会对特定主题及相应文档集下的检索效果有所改进，但改进的效果不明显。Chua[6]对Reuters-21578数据集中的前10个类别，通过基于WordNet构建名词集合、动词集合、形容词集合、副词集合和混合词性集合，利用多项式朴素贝叶斯算法进行了文本分类实验，实验结果表明基于WordNet构建的名词集合的分类效果稍微好于其他四种词性集合，并认为名词特征集合可以更好地表达分类信息。Liu[7]等采用基于名词、动词和形容词共现的方法对Sougou文本分类语料中的五个类别进行了文本聚类比较，实验结果表明基于上述词性的特征选择方法要好于DF(Document Frequency)等特征选择方法。姚清耘[8]等利用Sougou语料对所有词性和只采用名词为特征进行中文文本聚类比较，结果表明只采用名词构建向量特征空间的聚类效果要明显好于所有词性参与聚类的效果。Rosell[9]基于四组瑞士语语料集，使用K-Means算法验证了词性选择对瑞士语文本聚类的作用，结果认为词性标注方法没有提高瑞士语文本聚类的结果，但得出结论认为，在瑞士语文本中，当选择名词和专有名词作为文本的特征时，可以取得和所有词性参与聚类的结果比较接近，但后者可显著降低文本特征维度，因此认为名词是瑞士语文本聚类的重要特征。Sedding[10]等通过采用词性标注对部分Reuters-21578语料中的多义词进行了先消歧再聚类，结果表明基于词性标注的消歧并不能提高聚类的效果。目前来看，词性选择在文本信息处理中已经普遍应用，将数词、冠词等词性进行过滤，不仅可以降低文本特征维度，还可以提高处理效果。名词、动词、形容词和副词在中英文中都是重要的词性，这四种词性对中英文文本聚类的影响尚需全面的实验验证。

目前已有的相关研究在词性选择研究时，一般选取一种语料或一种聚类算法进行比较，或仅比较分析其中的部分因素，带有一定片面性，其结论缺乏全面的论证。为了全面考察名词、动词、形容词和副词四类主要词性对文本聚类的贡献度，本研究利用四组有代表性的中英文数据集，尝试从更全面的角度验证四类主要词性对中文和英文文本聚类的影响。本研究的主要目的在于，全面地探讨四种主要词性及词性组合对中英文文本聚类的作用，为中英文文本挖掘和文本组织提供有价值的参考。

2 英汉词性标注集与数据集处理

2.1 词性标注集

在实验开始之前，首先需要确定词性标注集。英语词性标注集主要有Penn Treebank标注集、CLAWS5标注集和CLAWS7标注集，多数标注集是在Brown标注语料基础上改进而来。CLAWS5标注集和CLAWS7标注集适用于中型和大型语料库的标记，Penn Treebank标注集[11]适合于小规模语料标注，包含48个词性标记，是一个比较简单的词性标注集。汉语词性标注集比较有影响的有中国科学院计算技术研究所汉语词性标注集和北京大学汉语文本词性标注标注集。中国科学院计算技术研究所汉语词性标注集共有99个词性标记，北京大学汉语文本词性标注标注集共有68个词性标注。

根据语料的规模和性质，本文选择Penn Treebank标注集和中科院计算所汉语词性标注集标注英文语料和中文语料。Penn Treebank标注集和计算所汉语词性标注集都是为了语法分析的目的而构建的，在文本聚类特征选择时仍是过细的标注。如中科院计算所汉语词性标注集V3.0版，将名词细分为nr,nrf,ns,nt,nz,ntl等词性，这些细分的词性可以为深层的自然语言处理提供支撑，但选择更细的词性特征，会造成文本特征稀疏的问题更为突出。我们将英文和中文细分词性进行了合并处理，最终选择最能体现文本内容的四类词性—名词、动词、形容词和副词。词性标注集合并后的信息见表1和表2。

表1 宾州树库英文词性标注

类 别	包含词性
名词类(N)	NN,NNP,NNPS,NNS,FW
动词类(V)	VB,VBD,VBG,VBN,VBP,VBZ
形容词类(ADJ)	JJ,JJR,JJS
副词类(ADV)	RB,RBR,RBS

表 2　计算所汉语词性标记集汉语词性标注 V3.0

类别	包含词性
名词类(N)	n,nr,nrf,ns,nsf,nt,nz,nl,ng
动词类(V)	v,vd,vn,vf,vx,vi,vl,vg
形容词类(ADJ)	a,ad,an,ag,al
副词类(ADV)	d

2.2　数据集处理

实验所用中文和英文的数据集，不同语言分别采用不同处理方法，同一语言尽量保持一致。英文处理主要包括三部分：tokenization(断词)、词性标注和词形还原。对于 20Newsgroups 和 Reuters-21578 数据集，在使用前需要进行预处理等清洗工作。

20Newsgroups 由 Lang 收集来自 20 个不同新闻组的文本，Rennie 将 20Newsgroups 整理成了三个版本的语料[12]，本实验选择第二个版本 Bydate 训练语料部分，占总语料的 60%，该版本的语料去除了原始语料中的重复部分和文本的头部信息，更接近于真实的文本处理任务。Bydate 版本的训练语料还是存在一些问题，部分文档还包含 PGP 签名的加密信息，也有些文档含有乱码，预处理阶段去除了这些干扰信息。

Reuters-21578 共包含 21 578 篇文本，本实验选择 Lewis 基于 modApt 方法分割的训练语料[13]，去除了多分类标签文本，保留 8 个单分类下的文本。对于有些文档只有 TITLE，没有 BODY，以及长度<3 的短文本，本研究没有考虑入内。为了准确进行词性标注，在预处理等清洗过程中，尽量保持文本的原貌，如在词性标注之前，并没有将复合词进行处理，也没有进行停用词处理。

在预处理之后，英文需要 tokenization，其主要工作是根据空格断词，对于连写词“I'm”需要处理成“I'm”。英文词性标注选用 Stanford Log-linear Part-Of-Speech Tagger，由斯坦福大学自然语言处理小组基于最大熵算法开发，整个项目开源，目前使用较为广泛。由于英文存在词形变化，在词性标注后，通过词形还原将变化的词形还原生成基本词形。

英文数据集词性标注和词形还原完成后，实验还去除了长度小于 3 的单词，一般情况认为，长度小于 3 的单词往往没有多大意义。英文停用词采用 smart 系统中包含的 574 个停用词的词表。此外，文本中还包含一些数字和合成词，一并进行统一处理。由于词性识别受到上下文影响，Stanford Part-Of-Speech Tagger 将“数字-单词”、“数字-数字”等复合词识别为名词结构或形容词，如“53-year”、“8-k”，为解决该问题，处理后的复合词根据“-”、“_”进行断词处理，保留长度超过 2 的非数字单词。

中文语料本文选取了复旦文本分类语料和 TanCorp V1.0 语料。复旦文本分类语料分为训练语料和测试语料，我们选择了其中的训练语料部分。复旦语料中存在大量类内重复和类间重复文本，对于类内重复文本，仅保留一个副本；类间重复，一并去除，最终语料仅保留单标签文本。两组中文语料采用中国科学院分词工具 ICTCLAS 进行分词，词性标注采用中科院计算所汉语词性标记集二级词性标注，去掉数字、叹词、语气词、拟声词和各种标点符号。停用词表采用哈尔滨工业大学中文停用词表。处理后的各语料特征数量和所占比例如表 3 所示。下文将处理后的 20 Newsgroups 简称为 20NG，Reuters-21578 简称为 8RE，复旦分类语料简称为 FDCorp，TanCorp V1.0 简称为 TanCorp。

表 3　四种语料分布情况

20NG		8RE		FDCorp		TanCorp	
类别	数量	类别	数量	类别	数量	类别	数量
atheism	480	acq	1 429	Space	506	财经	819
graphics	583	crude	215	Energy	30	电脑	2 943
windows. misc	574	earn	2 662	Electronics	25	地域	150
ibm. pc. hardware	590	grain	38	Communication	25	房产	935
mac. hardware	578	interest	132	Computer	1 022	教育	808
windows. x	593	money-fx	168	Mine	33	科技	1 040
misc. forsale	585	ship	105	Transport	57	汽车	590

续表

20NG		8RE		FDCorp		TanCorp	
类别	数量	类别	数量	类别	数量	类别	数量
autos	594	trade	212	Art	547	人才	608
motorcycles	598			Enviornment	805	体育	2 805
sport. baseball	597			Agriculture	847	卫生	1 406
sport. hockey	600			Economy	1 468	艺术	546
sci. crypt	595			Law	51	娱乐	1 500
sci. electronics	591			Medical	51		
sci. med	594			Military	74		
sci. space	593			Politics	1 009		
religion. christian	599			Sports	1 204		
politics. guns	545			Literature	33		
politics. mideast	553			Education	58		
politics. misc	465			Philosophy	40		
religion. misc	377			History	465		
总计	11 284	总计	4 961	总计	8 350	总计	14 150

注：表中的四组语料次数均为除去停用词、标点符号和数字后的词汇数量，并且两组英文语料经过词形还原。

根据表3呈现的数据，8RE和TanCorp的类别分布较为不均衡，最大数量的类分别是最小数量类的70和40倍之多。相比之下，20NG和FDCorp是分布较为均衡的语料，尤其是20NG是四组语料中分布最为均衡的语料。四组语料均是文本聚类领域常用的数据集，既存在类别分布均衡的语料，也存在分布不均衡的语料，这样选择尽量避免单一类型语料的影响。

3 实验及结果分析

3.1 数据集词性分布

我们首先对四组语料中的词性分布进行了统计，为了研究四类主要词性及词性组合对文本聚类的影响，我们设计4组单一词性和5组混合词性共9组实验，每组特征统计结果见表4。

表4 四组语料中不同词性及词性组合统计

	20NG		8RE		FDCorp		TanCorp	
词性及组合	次数	比例/%	次数	比例/%	次数	比例/%	次数	比例/%
N	826 314	**65.29**	213 828	**68.79**	6 285 718	**44.92**	2 029 118	**41.33**
V	230 763	**18.23**	46 281	**14.89**	4 422 225	**31.60**	1 502 361	**30.60**
A	176 830	**13.97**	45 953	**14.78**	612 788	**4.38**	339 469	**6.91**
D	29 075	**2.30**	4 459	**1.43**	820 708	**5.86**	382 284	**7.79**
N-V	1 057 077	**83.53**	260 109	**83.68**	10 707 943	**76.52**	3 531 479	**71.93**
N-A	1 003 144	**79.26**	259 781	**83.57**	6 898 506	**49.30**	2 368 587	**48.25**
N-V-A	1 233 907	**97.50**	306 062	**98.46**	11 674 242	**83.42**	3 870 948	**78.85**
N-V-A-D	1 262 982	**99.80**	310 521	**99.90**	12 494 950	**89.29**	4 253 232	**86.63**
All-POS	1 265 563	**100.00**	310 844	**100.00**	13 994 201	**100.00**	4 909 377	**100.00**

注：表4中，N-名词，V-动词，A-形容词，D-副词，N-V表示N和V组合词性，All-POS表示所有词性，但不包括数词、标点和停用词。下表同。

为了降低单一聚类算法带来的影响，本文采用划分聚类和层次聚类两种常用的聚类算法进行聚类实验。考虑到初始种子选择对原始 K-means 算法影响较大，划分聚类算法选择 K-means Clustering 和 Bisecting K-means Clustering；层次聚类算法选择 Agglomerative Hierarchical Algorithms。K-means Clustering 和 Agglomerative Hierarchical 是常见的算法，在此不作赘述。Bisecting K-means Clustering 算法，也称为二分 k 均值算法。基本思想是：为了得到 k 个簇，将所有点的集合分裂成两个簇，从这些簇中选取一个继续分裂，如此下去，直到产生 k 个簇。

3.2 实验评价方法

本实验采用熵(Entropy)和纯度(Purity)两个评价方法来评价聚类结果。假设待聚类的文本集人工标注为 q 个类别。通过某一次聚类实验，得到 k 个结果簇，对于包含 n_r 个对象的簇 S_r 的熵 E 可以计算如式(1)所示：

$$E(S_r) = \frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \qquad (1)$$

n_r^i 是第 i 个类中被聚到第 r 个簇中对象的数量，整个聚类实验结果的熵计算如式(2)：

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \qquad (2)$$

同样，对于聚类结果簇 S_r 的纯度可以计算如式(3)：

$$P(S_r) = \frac{1}{n_r}\max_i(n_r^i) \qquad (3)$$

整个聚类实验结果的纯度如式(4)所示：

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \qquad (4)$$

纯度是正确聚类的文档数占总文档数的比例，表示某一个簇中占主导地位类别的数量与该簇数量的比值。其值在 0-1 之间，完全错误的聚类时值为 0，完全正确的聚类时值为 1。纯度的评价方法无法对退化的聚类方法给出正确的评价，如果聚类算法把每篇文档单独聚成一类，该方法认为所有文档都被正确分类，纯度为 1。比较公正的评价是与熵结合起来，熵是系统混乱程度的度量，值在 0 到 1 之间，越靠近 0 说明该类的成员越是由同一个类组成，越靠近 1 说明该类的成员组成越混乱，该值体现了结果簇中每个类的分布情况，其值越小，聚类整体效果越好。

3.3 实验结果分析

本文采用划分聚类和层次聚类的三种算法，对四组单标签中英文分类文本语料进行了聚类实验，以期更全面准确地比较词性对中英文文本聚类的影响。考虑在实际应用中，聚类结果簇的数目往往是未知的，实验时对每组语料选择 k=5、k=10、k=15 和 k=20 进行聚类。在三种聚类算法下共得到 108 (9×4×3)组实验结果，由于实验数据量较大，为了更全面展示多次聚类结果，最终聚类结果为每组实验在三种聚类算法下得到的平均值。聚类结果见图 1 至图 4，详细数据见表 5。

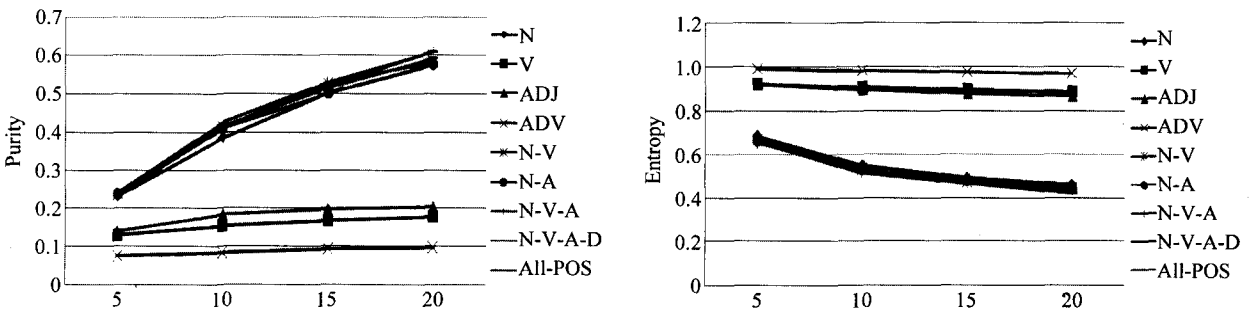

图 1　20NG 中不同词性和词性组合聚类结果

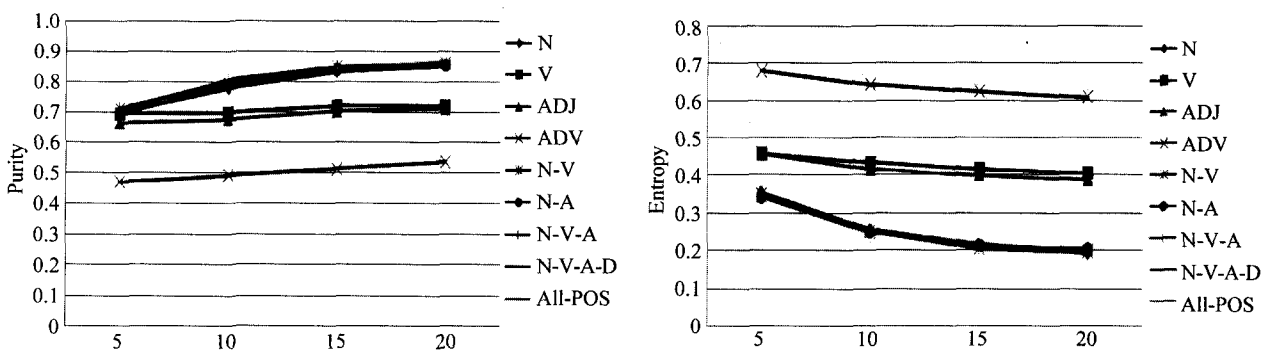

图 2　8RE 中不同词性和词性组合聚类结果

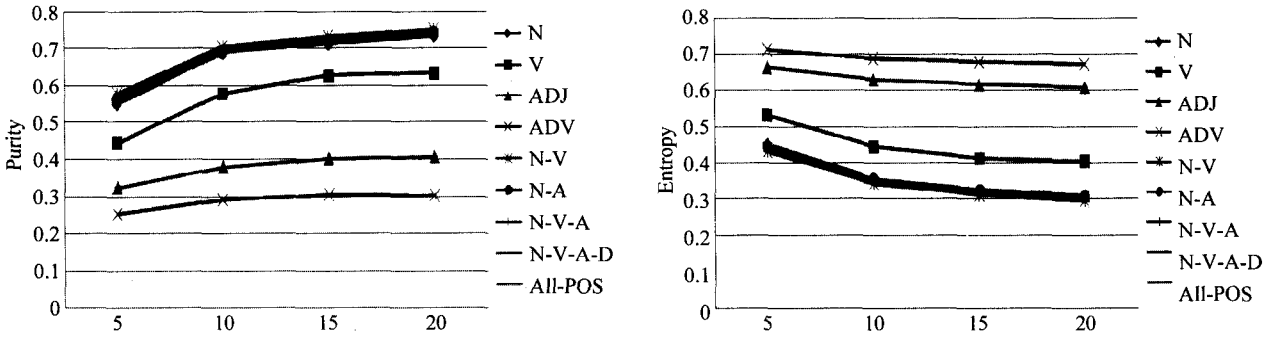

图 3 FDCorp 中不同词性和词性组合聚类结果

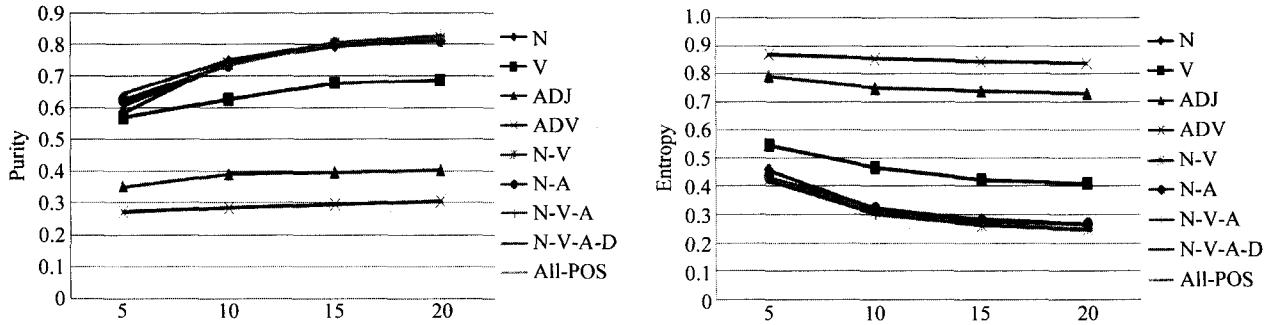

图 4 TanCorp 中不同词性和词性组合聚类结果

表 5 四组数据集的聚类结果

Dataset	POS	K=5		K=10		K=15		K=20	
		Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy
20NG	N	0.232	0.682	0.386	0.551	0.503	0.495	0.576	0.464
	V	0.129	0.920	0.153	0.905	0.167	0.896	0.178	0.886
	A	0.139	0.919	0.183	0.896	0.197	0.879	0.204	0.867
	ADV	0.074	0.991	0.082	0.983	0.092	0.976	0.096	0.969
	N-V	0.236	0.665	0.412	0.535	0.527	0.471	0.583	0.448
	N-A	0.236	0.664	0.410	0.534	0.500	0.481	0.577	0.448
	N-V-A	**0.240**	**0.656**	0.412	0.523	0.519	0.469	**0.611**	0.424
	N-V-A-D	**0.240**	0.661	**0.425**	**0.516**	**0.530**	**0.468**	0.609	**0.421**
	All-POS	0.238	0.663	0.417	0.534	0.514	0.477	0.594	0.435
8RE	N	0.699	0.353	0.779	0.256	0.832	0.215	0.852	0.193
	V	0.694	0.458	0.697	0.434	0.718	0.416	0.719	0.405
	A	0.664	0.458	0.676	0.416	0.704	0.399	0.713	0.387
	D	0.469	0.678	0.492	0.642	0.512	0.623	0.535	0.608
	N-V	**0.711**	0.347	0.792	0.250	0.850	0.206	0.861	0.196
	N-A	0.693	0.344	0.786	0.251	0.839	0.213	0.854	0.202
	N-V-A	0.702	0.346	**0.805**	**0.248**	0.846	0.210	0.852	0.205
	N-V-A-D	0.708	**0.343**	0.803	0.250	0.852	0.205	0.854	0.199
	All-POS	0.708	0.344	0.795	0.250	0.842	0.218	0.867	0.190

续表

Dataset	POS	K=5		K=10		K=15		K=20	
		Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy
FDCorp	N	0.549	0.451	0.690	0.355	0.712	0.324	0.735	0.309
	V	0.443	0.533	0.578	0.444	0.627	0.412	0.634	0.403
	A	0.323	0.664	0.380	0.630	0.401	0.616	0.406	0.608
	D	0.252	0.714	0.291	0.688	0.305	0.678	0.302	0.672
	N-V	0.578	**0.431**	**0.707**	0.343	0.734	0.312	**0.752**	**0.295**
	N-A	0.565	0.443	0.691	0.355	0.722	0.323	0.743	0.306
	N-V-A	0.563	0.439	0.700	**0.342**	0.716	0.316	0.750	0.299
	N-V-A-D	**0.571**	0.437	0.700	0.349	**0.735**	**0.311**	0.750	0.296
	All-POS	0.554	0.451	0.690	0.351	0.723	0.316	0.743	0.301
TanCorp	N	0.583	0.461	0.745	0.317	0.794	0.283	0.811	0.267
	V	0.570	0.546	0.626	0.469	0.677	0.426	0.686	0.411
	A	0.351	0.791	0.390	0.749	0.396	0.739	0.403	0.730
	D	0.271	0.870	0.284	0.855	0.294	0.845	0.305	0.839
	N-V	0.605	0.428	0.740	0.315	0.803	**0.262**	0.818	0.249
	N-A	0.623	0.434	0.736	0.323	0.803	0.278	0.812	0.266
	N-V-A	0.616	0.428	0.741	**0.302**	**0.805**	0.264	**0.827**	**0.246**
	N-V-A-D	**0.643**	**0.420**	**0.751**	**0.302**	0.804	0.265	0.817	0.248
	All-POS	0.618	0.432	0.736	0.314	0.802	0.266	0.820	0.249

实验数据说明，由于四组语料中均存在短文本，在选择单一词性为文本特征时，造成了部分文档长度为 0，实验中删除了长度为 0 的文本，所以在选择单一副词词性时，其文本总量略小于总文本数。

参考图 1～4、表 4 和表 5 数据，从词性比例、聚类结果的 Purity 和 Entropy，分别就四个单一词性和五组词性组合进行分析。

1）单一词性特征的数量比较

四种单一词性特征数量在中文和英文中的比例差异较大，但对于同一语种的两组语料，同一词性所占比例比较接近。根据表 4 显示，名词特征在英文语料中所占的比例远高于在中文语料中的比例；动词特征在中文语料中所占的比例远高于在英文语料中的比例；形容词在英文语料中的比例略低于动词，但在中文语料中的比例远低于动词；副词在英文语料中的比例非常低，在中文语料中的比例和形容词接近。在数量和比例上，四类词性特征是文本特征的重要组成部分，尤其是在英文语料中比重很大。

2）单一词性特征对文本聚类的影响分析

名词：表 5 和图 1～4 显示，在四个单一词性中，名词是对文本聚类影响最重要的词性。采用单一名词特征聚类的结果远好于其他单一词性特征聚类的结果，甚至与采用词性组合的特征所达到的结果十分接近。表 5 数据显示，虽然选择单一名词词性作为文本的聚类特征可以实现较好的聚类效果，但是仅仅采用名词特征还不能达到最优的聚类。

动词：四种词性中，在数量上，动词是除了名词之外比例最大词性，尤其是在两组中文语料中，动词所占的比例仅次于名词。但不同语种语料之间，动词比例存在较大差异。在聚类效果上，采用单一动词特征的聚类效果明显低于采用单一名词特征的聚类效果。在两组英文语料中，动词所占的比例远低于名词，仅选择动词特征会造成文本特征稀疏，这可能是造成单一动词特征在英文语料中聚类效果差的主要原因。在两组中文语料中，尽管动词的比例占总特征的 30%左右，但其聚类效果远低于名词的聚类效果，这表明，动词作为特征对文本的区分度不如名词。对于两组中文语料，单一动词为特征在 Purity

低于单一名词为特征时10%左右，在Entropy上高于单一名词为特征时10%左右。对于两组英文语料，由于动词比例较低，这两个差距变得更大。

形容词：在数量上，该词性在两种语言中分布差异较大。两组中文语料中，形容词数量是四种词性比例最小的，但在两组英文语料中，该词性的比例与动词所占比例较为接近。在聚类效果上，该词性在四组语料中也表现很大差异。对于两组英文语料，该词性在Entropy上均低于动词，但远高于名词和其他词性组合；在Purity上，该词性远低于名词和其他词性组合，在20NG中，其表现要好于动词，但在8RG中，该词性的表现略低于动词。通过两组英文语料，我们认为，在英文中，和动词相比，形容词对文本的类别有更好的区分能力。在两组中文语料中，该词性在Entropy和Purity上，都远不及动词，在Entropy上，高于动词20%以上，在Purity上，低于动词20%以上。在两语种语料中，形容词的表现差异很大，其根本原因是在中文语料中，形容词所占的比例非常低，仅为总词性特征的5%左右，但就两组英文语料来看，单一形容词比单一动词在文本类别上有更好的区分能力。

副词：在数量上，该词性特征所占比例较小。中文语料中，该词性比例略高于形容词，但在英文语料中，该词性的比例非常低，仅占总特征的2%左右。在聚类效果上，该词性在四组语料中的表现最差，这在英文中比较容易理解，仅选择该词性为特征时，造成文本的特征非常稀疏，不利于文本的聚类。在中文语料中，虽然该词性的比例高于形容词，但其聚类的效果却不及形容词。根据形容词和副词的聚类结果，我们认为，在中文中，副词在表征文本内容的区分度上不及形容词。

3）词性组合对文本聚类的影响

经过对单一词性在中英文文本聚类的结果比较，发现名词和形容词具有更好的文本类别区分度。为了进一步验证词性组合对聚类的影响，我们选择了N-V、N-A、N-V-A、N-V-A-D和All-POS共五组词性组合进行了实验。根据图1～4显示，五组词性组合在四组语料聚类的表现非常一致。但从表5数据上看，五组词性组合存在细微不同。

N-V和N-A：两词性组合在两组英文中的比例均在80%左右，从聚类的Purity和Entropy上，N-V的效果要好于N-A，虽然在20NG中，单一形容词词性作为特征时要优于单一动词，但同一语料中，N-V的效果略好于N-A，或者是很接近。在两组中文语料中，N-V的数量高于N-A 20%以上，N-V词性组合在数量上占有绝对优势，但N-V与N-A的效果却比较接近，我们认为主要是名词特征在起重要作用。

N-V-A和N-V-A-D：在数量上，这两种组合比例都很高，尤其在英文语料中，所占比例接近于All-POS。聚类结果上，N-V-A和N-V-A-D是往往能够实现最优聚类的词性组合。尤其是N-V-A-D词性组合，在四组语料的多次实验中，实现最优聚类的次数最多。根据多次实验结果，我们认为，名词是表征文本内容特征最重要的词性，其他三种词性动词、形容词和副词对文本内容表征也有不同的贡献度，对文本类别区分度均有正的影响。

All-POS：该词性组合是去除了停用词、数词和标点符号后，所有的词性特征组合。从数量上，在英文语料中，除了四种主要词性，其他词性数量几乎可以忽略，在聚类结果上，和N-V-A-D组合相比，All-POS为特征时聚类结果不升反降，表明，四类主要词性外的其他词性对文本类别区分度有负作用；在中文语料中，除了四种主要词性，其他词性大约占总特征的10%，在中文语料中，存在和英文语料类似的现象，虽然特征数量增加了，但All-POS聚类结果不及N-V-A和N-V-A-D的聚类结果。

虽然四组有代表性的数据集并不能涵盖所有的语料分布情况，但本文的研究可以反映大部分的情况。通过多次实验，我们发现，在中英文文本聚类中，词性是一个重要的影响因素。名词是表征本文内容的重要特征，在所有词性中，其类别区分度最高。仅采用单一名词特征聚类，可以实现较好的聚类结果，甚至与保留所有词性的聚类效果比较接近，但采用单一名词为特征，可使文本维度大大降低，对于英文，文本维度可以降低30%以上，对于中文，文本维度可以降低60%左右，在聚类的速度上很占优势。但仅仅采用单一名词作为文本特征，不能达到最优的聚类结果。在多数情况下，选用名词、动词、形容词和副词的组合特征得到的聚类结果，要好于单一词性和其他词性组合的聚类结果。四种主要词性之外的其他词性对文本聚类有负影响。

4 结论

本文选用四组有代表性的中英文语料，采用三种聚类算法验证了词性对中英文文本聚类的影响。通过实验我们得出如下结论：（1）名词、动词、形容

词和副词是文本特征的重要组成部分，但在中文和英文中，各词性所占的比例有很大差异；(2)在中文和英文中，名词均是最重要的语言知识体，是表征文本内容最重要的词性，在单一词性中其类别区分度最高，仅采用单一名词特征聚类的结果与保留所有词性时的结果相当。动词、形容词和副词对文本聚类均有不同的贡献度，同一词性贡献度在两语种之间存在差异。相对于英文，不同词性特征及其组合在中文文本聚类中呈现的差异更为稳定；(3)通常情况下，采用去除停用词，保留所有特征参与文本聚类的方法，并不能实现最优的聚类结果；(4)在中英文文本聚类中，多数情况下，采用名词、动词、形容词和副词四类词性组合特征得到的聚类结果，要好于其他词性组合的聚类结果。在下一步工作中我们将研究词性之外的因素对文本聚类的作用，在一些常用特征的基础上再考虑不同词性对于聚类结果的影响；下一步还要对不同词性特征进行加权，进一步挖掘对聚类有重要作用的因素。

参考文献

[1] J Gimenez, L Marquez. Fast and accurate part-of-speech tagging: the SVM approach revisited[A]//Proceedings of the 4th RANLP, Bulgaria, 2003: 158-165.

[2] 王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注[J]. 中文信息学报，2009,23(4):16-21.

[3] Y C Wu, J C Yang, Y S Lee. Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff 2008[C]//Proceedings of the SIGHAN, 2008.

[4] A Chen, Y Zhang, G Sun. A Two-Stage Approach to Chinese Part-of-Speech Tagging[C]//Proceedings of 6th SIGHAN Workshop on Chinese Language processing. Indian, 2007:82-85.

[5] 苏祺，眷红英，胡景贺，等. 词性标注对信息检索系统性能的影响[J]. 中文信息学报，2005，19(2):58-65.

[6] S Chua. The Role of Parts-of-Speech in Feature Selection[C]//Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong. 2008.

[7] Z T Liu, W C Yu, Y L Deng. A Feature Selection Method for Document Clustering Based on Part-of-Speech and Word Co-Occurrence[C]//Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010). Yantai, China.

[8] 姚清耘，刘功申，李翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程，2008,34(18):39-41.

[9] M Rosell. Part of speech tagging for text clustering in swedish[C]//Proceedings of the 17th Nordic Conference of Computational Linguistics. Odense, Denmark. 2009.

[10] J L Sedding, D Kazakov. Wordnet-based text document clustering[C]//Proceedings of the Third Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND). Geneva, 2004:104-113.

[11] M P Marcus, B Santorini, M A Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank [J]. Computational Linguistics, 1993,19(2):313-330.

[12] J Rennie. 20 Newsgroups dataset [EB/OL]. [2012-03-16]. http://people. csail. mit. edu/jrennie/20Newsgroups/.

[13] D Lewis. Reuters-21578 dataset [EB/OL]. [2012-03-16]. http://www. daviddlewis. com/resources/testcollections/reuters21578/.