基于 70 年报刊语料的现代汉语历时稳态词抽取与考察*

饶高琦，李宇明

北京语言大学，北京，100083

摘要：本文基于 70 年跨度的历时报刊语料库，使用 9 种统计方法计算了词语的历年使用情况，并通过对稳定性、覆盖度和时间区分性能的考察筛选获得了规模 3015 词的历时稳态词候选词集。该词集中动词与名词各占约三分之一（其余为形容词、副词与虚词），平均词长约 1.7 字，前密后疏得分布于历时语料库总频序表的前 7609 位，覆盖了近九成语料。该部分词语中包含大量构造句子结构的核心词语。它们塑造了稳态词在词长和词类上的特性。稳态词的提取可以加深对语言生活底层与基础词汇的认识，对汉语教学、中文信息处理和语言规划都具有重要意义。

关键词：稳态词，历时语料库，语言监测

Extraction and Investigation of Steady-state Words based on Diachronic Corpus of Newspapers across 70 Years

Rao Gaoqi，Li Yuming

Beijing Language and Culture University, Beijing, 100083, China

Abstract: Based on the diachronic corpus of modern newspaper across 70 years, 9 statistic methods were applied to compute the stability of word use to gain candidate sets of steady-state words. 3015 words were obtained by evaluating their corpus coverage, time sensitivity and diachronic classification. In this set, verb and noun take one third respectively (the rest consists of adjectives and function words). Word length is 1.7 characters in average. Steady-state words are distributed in top 7609 in frequency list, covering 90% of corpus. Basic morphemes and core words shape the features of the set in POS and length. Extraction of steady-state words benefits Chinese teaching, NLP and language planning.

Key Words: steady-state word, diachronic corpus, language monitoring

1 引言

汉语演变的历程中，词语使用受时间影响的程度不一，表现为词语在时间维度上的分布不同。词汇系统中很多词语使用稳定，受时间影响小，更新和变异缓慢，构成了现代汉语词汇系统的底层，起到基础和主干的作用。张普[1]先生的研究中将这部分词语被称作稳态词。

在自然语言处理中的直接相关研究较少。Fumiyo 等[2]使用卡方检验从 30 年的 MedLine 英文文档集中筛选具有时间显著性的名词。Degaetano-Ortlieb[3]等在 SciTex 中挖掘了与时间相关的词类串和用词特征，并使用相关分析协助选择。谢晓燕[4]则使用词语生命度方法计算了词在各时间点上的活跃程度，辅之以问卷调查从 30 年的《深圳特区报》中获得稳态词表。

本文从大时间跨度的历时语料中抽取稳态词候选集，以描述当代汉语报刊语言使用的底层，进而观察整个语言社团语言生活的基础。历时语料库可以被视作不同时期产生的文本集。抽取没有时间区分度的稳态词语等效于找出其中没有文本区分力的词。这恰好是文本分类任务的反面，因而可以借鉴其计算方法抽取在历时语料库中没有时间特征的词语。

本文使用了文本分类任务和停用词抽取任务中比较成熟的几种特征统计方法。语言监测中常用的“共用词”提取也是重要的方法。本文使用的语料为 BCC 历时检索系统[1][5, 6]中 1946

*国家自然科学基金项目（61300081、61170162）、国家社科重大基金项目（12&ZD173）；国家语委科研基金项目（YB125-42）；国家 863 计划重点项目（SQ2015AA0100074）；国家社科重点项目（大数据背景

年到 2015 年的《人民日报》语料[2]，时间跨度 70 年，规模 12 亿字，词种数约 220 万。

本文的组织架构如下：第二节介绍了 9 种词语使用情况的计量方法；第三节为本任务中各方法获得稳态词候选集的情况；第四节中通过覆盖度、重叠程度和文本分类精度评价了诸方法的性能，并获得了最优候选集；第五节分析了稳态词的基本性质；第六节为小节与展望。

2 稳态词的抽取方法

2.1 词频逆文档频

Sparck-Jones[7]和 Robertson[8]在信息检索中使用包含特定词的文档频率对单纯词频统计进行平衡，从而发展出词频逆文档频，后文简称 TF-IDF。本任务中，TF-IDF 值较小的词更倾向于成为没有时间区分度的稳态词。

由于不同时期的语料规模不一，总语料规模和总文档数量可以用来对词频与逆文档频值进行归一化。本文使用公式 1 到 3 对历时语料库中所有词的 TF-IDF 值进行了计算。

$$TF \cdot IDF(w) = TF(w) \cdot IDF(w) \quad (1) \qquad TF(w) = -\log(\frac{F_w}{N}) \quad (2) \qquad IDF(w) = \log(\frac{D_w}{D}) \quad (3)$$

其中 F_w 和 D_w 分别表示词 w 在整个语料库中出现的次数和整个语料库中包含词 w 的文档数。F 和 D 则是整个语料库的全部词次数与文档数。注意到文档数 D 的大小取决于历时语料库划分的颗粒度。如一年的语料视作一篇文档（年颗粒度），一个月的语料视作一篇文档（月颗粒度），或使用其他颗粒度，对 D 和 D_w 值会产生巨大影响。

2.2 互信息

在文本计算中互信息（Mutual Information）越大，特征 w 和类别 C 共同出现的可能性就越大，由此可以推断 w 和 C 的关联性越强[9, 10]。在这一部分中 w 即为语料库中的词，类别 C 则是包含 w 的语料所属的时间。本节在计算中采用 Y. XU[11]在文本分类任务中推导的词与分类的互信息计算公式，如式 4、5。

$$I(w) = \sum_{i=1}^{m} p(w, c_i) \log_2 \frac{p(w, c_i)}{p(w)p(c_i)} + \sum_{i=1}^{m} p(\overline{w}, c_i) \log_2 \frac{p(\overline{w}, c_i)}{p(\overline{w})p(c_i)} \quad (4) \qquad p(w, c_i) = \frac{count_{c_i}(w)}{N} \quad (5)$$

上式中 w 为词，c_i 是分类，即特定时间点的语料。这里使用最大似然估计来计算词 w 在 c 中的概率。$count_{c_i}(w)$ 是词 w 在语料 c_i 中的频次。N 为整个语料库的总词数，m 为语料库在一定时间颗粒度下的文本数。如年颗粒度下，m=70。

2.3 联合熵

相较于互信息，联合熵（Union Entropy, UE）的计算中同时体现了包含特定词的句子在文本中出现的概率和词语在该句子中出现的概率。在少数民族语言和现代汉语的语料库中，联合熵在获取停用词，过滤噪音词方面取得了较好的效果[12, 13]。联合熵的计算方法如下面一组公式所示

$$UE(w_i) = H(w_i) + H(s \mid w_i) \quad (6) \qquad H(w_i) = -\sum_{j=1} p_j(w_i) \log p_j(w_i) \quad (7)$$

下汉语语块数据库建设与应用研究）

[1] http://bcc.blcu.edu.cn/hc

[2] 由于种种原因，本文实验过程中没有获得 2003 年到 2008 年的《人民日报》语料，该部分由实验室积累的相应年份的《贵州日报》替补。

$$H(s|w_i)=-\sum_{l=1}^{n}p_l(s|w_i)\log p_l(s|w_i) \quad (8) \qquad p_j(w_i)=\frac{coun_j(w_i)}{\sum_{j=1}^{n}coun_j(w_i)} \quad (9)$$

$$p_l(s|w_i)=\frac{count_l(s|w_i)}{\sum_{l=1}^{m}count_l(s|w_i)} \quad (10)$$

这组公式中一个词 w_i 在语料库中的联合熵 $UE(w_i)$由其在句子中分布的熵值（8）和包含 w_i 的句子在特定时间的语料中分布的熵值（7）构成。其中（9）为 w_i 在某一句子中出现的概率，用最大似然估计计算。$count_l(w_i)$是 w_i 在该句子中出现的次数，n 为句子数。$count_l(s|w_i)$是包含 w_i 的句子 s 在文本 l 中出现的次数，m 为文本总数。

2.4 词项随机采样

词项随机采样（Term Based Random Sampling, TBRS）方法由 Lo T W[14]提出，用于在网页上自动探测停用词。该方法随机选取若干词，在包含这些词的文档中计算所有词的 *KL* 距离，并归一化。对每个词在其出现的每个文档中的 *KL* 距离值取平均，排序后选取得分较小的为停用词。本部分使用的 *KL* 计算公式如式 11、12 所示。

$$KL_c(w_i)=p_c(w_i)\log\frac{p_c(w_i)}{p(w_i)} \quad (11) \qquad TBRS(w_i)=\frac{\sum_{c=1}^{m}\frac{KL_c(w_i)}{\max_{c=1}^{m}(KL_c(w_i))}}{m} \quad (12)$$

其中 c 为特定时间的语料，$KL(w_i)$为在语料 c 中 w_i 分布和整个语料中 w_i 分布的 *KL* 距离，$p(w_i)$为 w_i 在 c 中出现的概率，$p(w_i)$为 w_i 在整个语料库中出现的概率，m 为有 w_i 出现的语料的份数，*max* 函数则在各份语料中取 *KL* 距离的最大值。

2.5 修正频率

修正频率（Korregierte Frequenz, *KF*）又称为调整频率（Adjusted Frequency）。*KF* 统计可以避免单纯统计频次时，集中于某些文档的某些高频词被误认为是整个语料的高频词。本文使用的 *KF* 计算公式[15, 16]如式 13 所示：

$$KF(w_i)=(\sum_{c=1}^{m}\sqrt{p(c)f(w_i)})^2=(\sum_{c=1}^{m}\sqrt{\frac{N_c}{N}count_c(w_i)})^2 \quad (13)$$

其中 w_i 的 *KF* 值为它在特定时间语料 c 中出现概率与语料概率的根平均数。$p(c)$为语料 c 在整个语料库中出现的概率，用最大似然估计计算，N_c 除为语料 c 的词数，整个语料库词数为 N，m 为按特定时间颗粒度划分整个语料的份数。

2.6 均根匀度

Huarui Zhang[17]研究了词汇使用的分布程度，提出了均根匀度（Square-mean-root Evenness, SE）用以获得特定文档的核心词表。其计算方法如式 14 所示：

$$SE(w_i)=\frac{\left(\frac{\sum_{c=1}^{m}(\sqrt{count_c(w_i)})}{m}\right)^2}{\frac{\sum_{c=1}^{m}count_c(w_i)}{m}}=\frac{\frac{(\sum_{c=1}^{m}(\sqrt{count_c(w_i)}))^2}{m}}{\sum_{c=1}^{m}count_c(w_i)} \quad (14)$$

其中 c 为特定时间的语料，m 为整个语料划分的份数，其余符号含义与前文相同。

2.7 变异系数

如果一个词是稳态词，那么它的使用情况随时间变化小，在统计上应表现为离散程度较小。但语料库中不同词语标准处的测量尺度相差很大，直接比较并不合适。变异系数（Coefficient of Variation, CV）是常用于这种场景的统计量。它是变量标准差与平均数的比，如式 15 所示。

$$CV(w_i)=\frac{sd(w_i)}{aver(w_i)} \quad (15)$$

对于词 w_i，离散程度低（标准差小），频率高（频率均值大）就更倾向于成为稳态词。因此选取变异系数较小的词作为候选词。

2.8 共用词

共用词是在一组文本中都出现的词。在《中国语言生活状况报告数据篇》[18]中使用共用词描述不同领域中词汇的重合程度，并间接显示领域间的词汇差异。借鉴到本任务中，年共用词是在各年语料中都出现的词语。该性质使得这部分词语可能成为较好的候选稳态词语。

对于本文中的历时语料，年度共用词即为文档频率 DF=70 时产生的词表。月、周、日共用词即为 DF 为相应数量时产生的词表。

2.9 累计频率

从整个语料库中抽取出的分词单元按频率高低进行排序，并计算其累计频率。由于词汇使用分布的不均衡性，少数超高频与高频词占据了语料库的大部分篇幅。累积频率达到一定阈值时的高频词更容易成为整个语料库中更具有通用性的词。该方法只能获取词语的全局频率，因而局限性明显，本文以其作为对照组。

3 稳态词候选词集的获取

3.1 通过拐点确定候选集

词语的 TF·IDF 值计算取决于文本的颗粒度，即 IDF 的大小。随着时间颗粒度的变化，词语的 TF·IDF 值和词汇的 TF·IDF 排序都会出现显著不同，如图 1 左所示。

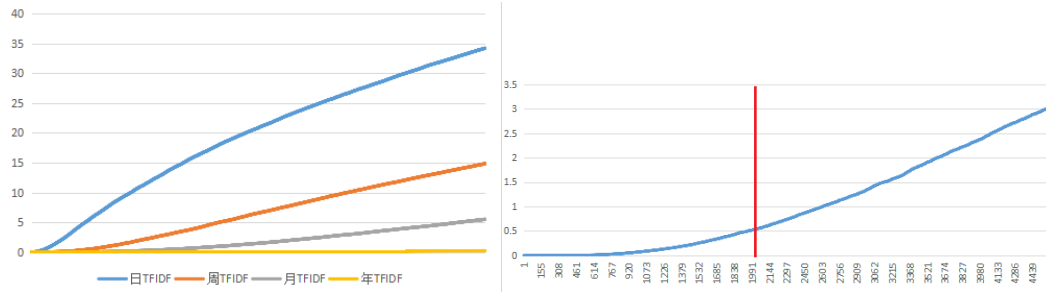

图 1 年、月、周、日四种颗粒度下词的 TF·IDF（升序前 15000 个）

从图 1 右中可以发现各颗粒度曲线的在拐点之后 TF·IDF 值的斜率趋于平缓。可以通过观察词 TF·IDF 值曲线的拐点来确定稳态词的数量。在以日为时间颗粒度的情况下拐点大致出现在 600 词的位置，在以周为时间颗粒度的情况下拐点大致出现在 2000 词的位置（如图

1 右），在以月为时间颗粒度的情况下拐点大致出现在 3000 词的位置，在以年为时间颗粒度的情况下拐点大致出现在 12452 词的位置。

以年为颗粒度的 TF-IDF 值的拐点取得是零跳变点，即在此排序中前 12452 个词的 TF-IDF 值均为 0。原因在于这些词在每一年的语料中都出现，造成 IDF=0。此时，依 TF-IDF 值选取出的候选词退化为了年共用词。各颗粒度的共用词将在 3.2 节中详细讨论。

大体上，四种时间颗粒度下的候选词表从大到小依次包含。各词表的交集为 599 词。

类似地，我们通过观察拐点的方式确定了基于互信息、联合熵、词项随机采样、修正频率、均根匀度、变异系数和对照组累计频率方法的候选词集的大小。由于数值的分布较差异较大，为了便于绘图和观察拐点，本部分对互信息、联合熵、修正频率值取以 10 为底数的对数。各颗粒度下候选集的规模如表 14 第二列所示。

3.2 基于共用词方法的候选集

表 1 年、月、周、日共用词数量及词类分布

	总数	名词	动词	形容词	虚词与其他
年共用词	12452	5208/0.418	4356/0.350	953/0.077	1932/0.155
月共用词	1821	576/0.317	633/0.348	169/0.093	440/0.242
周共用词	409	105/0.259	132/0.325	31/0.076	138/0.340
日共用词	15	1/0.067	1/0.067	1/0.067	12/0.800

本部分计算了历时语料库中年、月、周、日四种颗粒度下的共用词。年共用词有 12452 个，月共用词 1821 个，周共用词 409 个，而日共用词仅有 15 个。

日共用词按频率从高到低排分别是：的、了、在、和、是、一、个、中、上、为、地、到、人、大、下。它们都是单音节词。唯一的动词为“是”。“为”和“到”也有做动词的情况，但总体较少。唯一的名词是“人”，但“人”在数量结构中也可以做量词。唯一可以做形容词的是“大”。

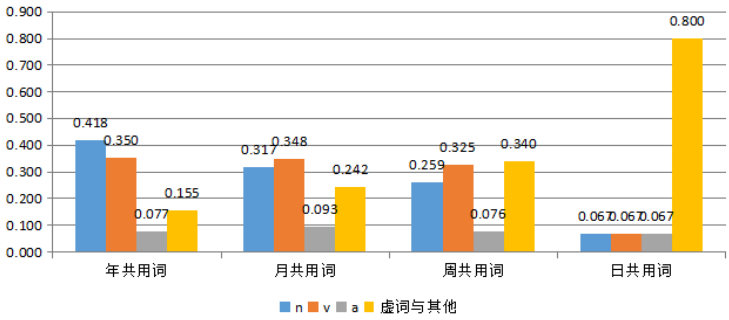

图 2 年、月、周、日共用词的虚词实词分布

构筑日常语用生活底层最稳定的部分是虚词。当“日常语言”细分到“每日必说”的程度时，名词、动词和形容词所占寡少。虚词所负载的主要是语法意义，是组织语言所必须的骨骼。而构成句子内容的实词则有很强的流动性甚至是偶然性。形容词在年、月、周、日共用词中的比例十分稳定，在 7%上下波动，是总量较少但不可或缺的一部分词汇。

由共用词的定义可知，年、月、周和日共用词表是逐层包含的。

4 选择一种稳态词表

前文综述了多种稳态词表的选取方法。本章采用以下办法对其进行评价：1.不同时间颗粒度下的重叠程度和时间敏感程度；2.对语料的覆盖程度；3.对历时文本分类的贡献。

4.1 重叠程度和时间颗粒度敏感程度

不同时间颗粒度下候选词表的重叠程度可以评价该方法的稳定性，重叠程度越高越好。

稳态词和非稳态词不会是泾渭分明的两个集合，而是渐变的连续统。前文使用了“年、月、周、日”四种时间颗粒度，获得了不同的稳态词候选词集。如果使用更大的时间颗粒度如“世纪”，则整个历时语料库都处在一个时间单元内。那么所有词都应该进入该颗粒度下的稳态词候选集，并包含其他颗粒度下的候选集。类似的，如果某方法在诸颗粒度下产生候选词集互相包含的程度高，它所筛选出的候选词集就更符合其作为连续统一部分的特性。

本节将重叠程度划分为四个等级[3]：完全重叠，即年、月、周、日四种时间颗粒度下产生的词表存在一者包含其他三者的关系；大部分重叠，即存在两者部分占各自很大比例（>80%）；部分重叠，即存两者重叠部分占各自较大比例（>30%）；较少重叠，即四者中任意两者重叠部分占各自比例较少（<30%）。表 2 为各种方法的重叠程度。累计频率法获得的结果和时间颗粒度无关，故不在本部分进行评价。

表 2 各方法产生的候选词集的重叠程度

	完全重叠	大部分重叠	部分重叠	较少重叠
TF-IDF		√		
互信息				√
联合熵	√			
词项随机采样				√
均根匀度			√	
修正频率		√		
变异系数		√		
共用词	√			

对于抽取稳态词集，本文倾向于选用对时间颗粒度的敏感性高的方法。这里使用由不同时间颗粒度下词表规模的大小差异来衡量不同方法对时间颗粒度的敏感程度。这里将年、月、周、日四种时间颗粒度下任意两候选词表间的词数之比取平均，来衡量这种差异。形式化描述如式 16 所示。

$$sen_t = aver(\sum_{s_j>s_i}\frac{s_j}{s_i}) \quad (16)$$

其中 sen_t 为某方法的时间敏感性，S_j 与 S_i 为两个不同时间颗粒度下获得的候选词表的规模，且有 $S_j>S_i$，$aver(\ )$ 为取算术平均值。

[3] 这一划分方法是针对本任务诸方法进行的，因而没有完备划分所有可能情况

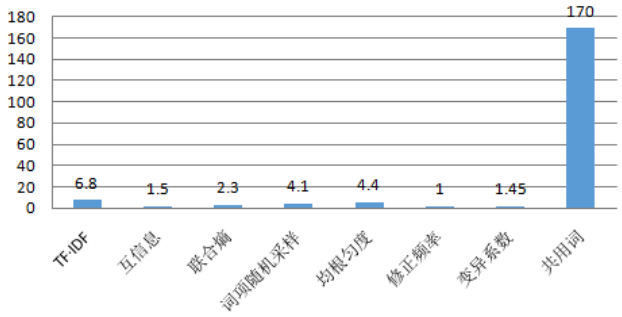

图 3 各方法的时间颗粒度敏感度

如图 3 所示，从时间敏感性和重叠度上来看，共用词和 TF-IDF 方法都是表现最好的稳态词获得方法。

4.2 在语料中的覆盖程度

如果一种方法获得的词表，其中词项在语料中的覆盖程度过小，说明它们充当语言生活底层的可能性不大，则其作为稳态词的可信度较低。各候选词表的语料覆盖程度如表 3 所示。TF-IDF、共用词、变异系数和修正频率在该项评价中表现较好，其中 TF-IDF 表现最佳。

表 3 各方法在总词汇中的覆盖度

	年	月	周	日	交集
TF-IDF	**0.876**	**0.762**	**0.725**	0.582	0.581
互信息	0.0005	0.004	0.015	0.079	-
联合熵	0.571	0.663	0.716	0.751	0.571
词项随机采样	0.06	0.0355	0.0415	0.16	0.02
均根匀度	0.368	0.509	0.588	0.841	0.364
变异系数	0.511	0.665	0.709	**0.765**	0.503
修正频率	0.663	0.662	0.662	0.662	**0.658**
共用词	**0.876**	0.703	0.517	0.239	0.239
累计覆盖率	0.8	-	-	-	-

4.3 对文本分类的贡献

根据对稳态词的特性，在历时语料库中去除候选词集后剩下的词汇应是与文本时间特性较为相关的。如果以这些词为特征对不同时间的文本进行以时间为类的文本分类，应有较好的分类效果。候选的稳态词集的质量越高，则去除该词集后语料库中保留下的具有较好时间敏感性的词越多，则最终文本分类的精确度越高。本节以此检验候选词集的质量。

本部分实验选取了文本分类任务中最经常用做基线实验的的朴素贝叶斯分类器[4]对测试数据集进行文本分类。

在历时语料库中均匀选取五分之一的年份（共 14 年），每年选取 2000 词的文本片段 20 篇，共 280 篇，56 万词作为文本分类任务的测试数据集。实验中去除频次为 1 的超低频词。再从测试集中去除第 3 章中各方法生成的候选词集，以剩余的词为特征，频次为特征值，进行分类实验。

[4] 本文使用 weka 数据挖掘平台[19]实现的朴素贝叶斯分类算法，版本 3.6.13

同时，设置一组对照实验。设测试数据集中包含各候选词集的词种为 m。作为对照，从测试数据集中随机去除 m 种词，以观察各方法生成的候选词集的质量。候选词集的质量越高，去除候选词集后的分类精度相较对照组提升的越大。

实验中，以年份为分类标准（14 类），10%交叉验证。实验结果如表 4 所示。

表 4 各方法在各颗粒度下产生的稳态词候选词集在文本分类任务中的表现

词集	词数	去除候选词精度（%）	随机去除词精度（%）	增幅（%）	词集	词数	去除候选词精度（%）	随机去除词精度（%）	增幅（%）
累积频率	3007	73.2	74.3	-1.48	修正频率.日	1000	80.4	75.7	6.21
					修正频率.周	1000	81.4	76.8	5.99
TF-IDF.日	600	78.9	76.8	2.73	修正频率.月	1000	83.6	76.8	8.85
TF-IDF.周	2000	**90**	75	20.00	修正频率.年	1000	81.8	76.1	7.49
TF-IDF.月	3000	**90.4**	72.1	**25.38**	修正频率.交集	967	81.1	76.8	5.60
TF-IDF.年	12452	87.5	70	**25.00**					
TF-IDF.交集	599	79.6	77.1	3.24	均根匀度.日	5000	87.1	73.6	18.34
					均根匀度.周	1000	81.1	76.8	5.60
互信息.日	1500	74.3	72.1	3.05	均根匀度.月	800	79.6	76.8	3.65
互信息.周	2000	73.2	76.1	-3.81	均根匀度.年	500	74.6	75	-0.53
互信息.月	2500	72.9	75.7	-3.70	均根匀度.交集	331	75.7	76.4	-0.92
互信息.年	3000	75	76.1	-1.45					
					共用词.日	15	73.6	76.8	-4.17
联合熵.日	2000	83.2	76.1	9.33	共用词.周	406	76.8	76.4	0.52
联合熵.周	1500	83.6	76.7	9.00	共用词.月	1818	**88.2**	74.3	18.71
联合熵.月	1000	80.4	76.8	4.69	共用词.年	12452	87.5	70	**25.00**
联合熵.年	500	78.6	73.2	7.38					
联合熵.交集	500	78.6	73.2	7.38	变异系数.日	3000	70.0	60.7	15.32
					变异系数.周	2500	74.3	67.1	10.73
TBRS.日	1600	76.4	73.6	3.80	变异系数.月	2500	75.4	62.9	19.87
TBRS.周	800	75	77.5	-3.23	变异系数.年	1500	75.7	65.0	16.46
TBRS.月	250	74.3	74.6	-0.40	变异系数.交集	962	74.3	60.7	22.41
TBRS.年	200	73.2	75	-2.40					
TBRS.交集	48	75.4	75.4	0.00					

从表 4 中容易看出，TF-IDF 方法在月颗粒度时形成的候选词集（后简称为 TF-IDF.m）帮助分类器获得了最好的分类精度，且该组实验相比于对照组也有最大的精度提升。相较于对照组，个别组的分类精度不增返降，说明这些词表中包含了具有较好时间敏感性的词。

经过比较，TF-IDF.m 完全包含了文本分类实验中分类精度第三名的月共用词词集，并且 TF-IDF.m 与精度第二的 TF-IDF 方法周颗粒度形成的候选词集（TF-IDF.w）仅有 12 个词的差异。也正因为这 12 个词，在重叠程度评价中 TF-IDF 方法没有成为完全重叠关系的方法。在覆盖度评价中，TF-IDF 方法形成的候选词集在年、月颗粒度中都优于其他方法。在

时间颗粒度的敏感程度上，虽与共用词有较大差异，TF-IDF 方法也超过其余诸方法。

TF-IDF 方法本质上是对单纯词频统计法的修正，其修正方式在于通过逆文档频 IDF 值描述了词分布的广泛程度。显然频率相同或相近的词中，分布更广泛的词所包含的信息量少，更有可能是构筑语言生活的底层的语言单位。而分布更窄的词对于了解其所在文档的特征具有更大价值。但 IDF 值的大小很大程度上取决于对整个语料库划分的粗细程度，亦即每份语料的规模。语料库默认以年为主要的时间计量单位。但每年语料的篇幅很大，词汇中不同词的词频波动范围很大。年颗粒度下的 IDF 取值（0 到 70），对中高频段的调节作用非常有限。实验表明以月为颗粒度进行划分对 IDF 值发挥调节作用较为合适。

月共用词的性能仅次于 TF-IDF.m。共用词的提取方法是选择每部分语料中都出现的词，而不考虑该词在各部分语料中出现频次的多寡。这实际上是 IDF=0（即文档频率 DF 取最大值）时 TF-IDF 的计算方法。因此共用词的取词方法是 TF-IDF 的一种极端情况。月共用词对语料的划分与月颗粒度 TF-IDF 的划分方法完全一致。这可以部分解释为何月共用词的性能仅次于月颗粒度 TF-IDF 方法。语料颗粒度主要确定词语在语料库中的分布程度，也就是历时的分布程度。颗粒度对候选词集的抽取有较大影响，说明词语是否进入时间分布层次的底层，分布广度较之频率高低更加重要。

类似的，在语言生活状况研究中常用的独用词是 TF-IDF 方法中 DF=1，IDF 取极小值时的另一种特殊情况。它所提取的恰恰是时间敏感性较强的词。

综合重叠性质、时间颗粒度敏感性、覆盖程度和文本分类性能三种情况，本文选择 TF-IDF 方法在月和周为颗粒度下形成的候选词集的并集（词与标点共 3015 个）为最优稳态词候选集（后文中简称为“词集”）。该词集完全包含月、周、日颗粒度下的共用词。

5 最优稳态词候选词集的性质

5.1 词类分布

词集中如“组织”“限制”“希望”等都是兼类词。表 5 前两行是以词在语料库中出现频次最多的词类为被统计词类。如果用兼类词在语料库中不同词类的出现概率对相应的词次数进行修正（如式 17 和 18）则可以得到修正词数和修正比例。

式中 $p_i(w)$ 为词 w 为词类 i 时的概率，通过最大似然估计获得。W 为整个词集，N_{wi} 为 w 以词类 i 在出现的次数，N_w 为 w 的总次数 N_w 为词类 i 修正后的词数。

$$N_i = \sum_{w \subset W} p_i(w) \qquad (17) \qquad p_i(w) = \frac{N_{wi}}{N_w} \qquad (18)$$

表 5 词集词类分布

词类	标点	形容词	动词	名词	虚词与其他
词数	15	254	1073	1038	633
比例	0.50%	8.43%	35.61%	34.45%	21.01%
修正词数	15	260.8	1025.5	936.9	774.8
修正比例	0.50%	8.65%	34.01%	31.07%	25.77%

词集中，动词略多于名词，且各占三分之一左右。形容词的比例和共用词方法中获得的相似，约占 8%。表 6 统计了词集中频次最高的无词类兼类现象的名词、动词和形容词。

表 6 词集中非兼类的高频名词、动词、形容词前十与频序

名词	人民/21	中国/30	国家/40	问题/46	群众/61	美国/66	企业/78	党/80	新华社/61	干部/82
动词	做/104	加强/124	参加/132	去/131	解决/141	举行/143	管理/152	表示/177	建立/178	实现/188
形容词	重要/121	伟大/236	友好/257	困难/313	严重/340	认真/378	重大/391	完全/403	坚决/411	正确/466

表 6 的频序分布呈现“名词<动词<形容词”的趋势，且差异明显。这主要是由于动词和形容词的兼类现象十分严重。总得来讲前十名的名词和形容词体现了很强的报刊语体特性。

5.2 词长分析

词集的词长分布可以按照词种和词项分别进行计算。前者不考虑该词在语料库中出现的频次，仅由词表计算生成；后者则用词集所覆盖的语料长度除以词数获得。

表 7 所示为按照两种方法计算所得的词长分布。平均词种词长为 1.69 字。双音词占据了词集的近七成，三音节、四音节的词占比不足 3%。仅有的两个超过四音节的词是“中国共产党”和“中华人民共和国”。双音词已成为构成报刊语言生活底层的主力。长词进入稳态词集需要强大而持续的社会外力。

表 7 词集词长分布

词长	单音节	双音节	三音节	四音节	五及以上
词种数	880	2026	88	19	2
比例	29.19%	**67.20%**	2.92%	0.63%	0.07%
词次数	232047655	219767262	6269166	1365739	251471
比例	**50.47%**	47.81%	1.36%	0.30%	0.05%

按照词项计算的平均词长为 1.52 字。与词种数分布产生巨大差异的地方在于单音节词和双音节词的对比发生了反转。单音节词的种数虽少，但平均词频很大，因而在词频比例上超过双音节词。如将标点符号也算入其中，则所占比例更大：单音节 57.85%，双音节 40.69%。

汉语的词汇变化总体而言是从单音向双音/多音演化。稳态词是随时间变化很小的词汇部分，因而保留了较多的单音节词。而在当代汉语中得以保存的这些单音节词多为基本构词语素或功能词，在组织语言的过程中不可或缺，因而频次很高，分布很广。

5.3 频位分布

对词集在语料库总词表中的频位分布进行统计可以发现 3015 个词与标点前密后疏地分布在频序 1 到 7608 之间，如图 4 所示。前 64 位和总词表频序完全相同。整个词集相对于总词频表的平均分布密度为 3011/7608=0.396。这一较低的平均分布密度表明了文档频率对纯粹词频的巨大修正作用。频序最低的稳态词频序位 7607，总词表在这里的累计频率为 0.889。

90%累积频率是对语料分布进行考察的常用阈值。它与稳态词集中最低频序词所对应的累计频率很接近。累计频率 90%的高频词基本可以包含质量较好的稳态词集。

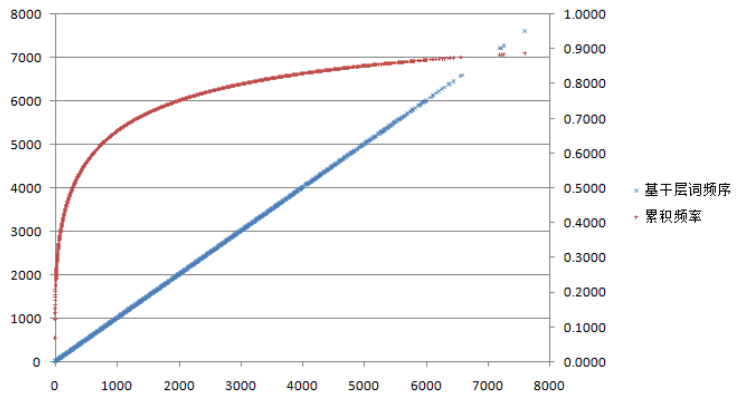

图 4 词集在总词表中的频序分布（左坐标轴）和对应的累积频率（右坐标轴）

将频序前 7609 位的词均分为十组，每组中稳态词出现的词数和密度如表 8 所示，频序越低越稀疏。

表 8 词集在总词表诸频序段的分布密度

分组	1	2	3	4	5	6	7	8	9	10
组内稳态词数	720	648	520	422	318	211	97	61	12	4
组内稳态词密度	0.947	0.853	0.684	0.555	0.418	0.278	0.128	0.080	0.016	0.005

汉语水平词汇等级大纲[20]甲级词（1035 个）中的 70%出现在词集中。谢晓燕[4]从《深圳特区报》中使用流通度理论提取了 26 年间使用较为稳定的稳态词，其中各时期稳态词交集的高频 3000 词与本文候选词集重合 2177 个，占 73%。由于其研究中的分词颗粒度和本文不同（如“一下子”），加之《深圳特区报》作为地方报纸对广东省外事务报导有限，造成中国和世界范围内的一些重要常用命名实体（如地名“山东省”“河北省”）没能进入其词表。排除这一部分（63 词），则与本文词集重合率为 75%。其余差异多由本文语料库更大的时间跨度和更广的新闻覆盖面造成。如“领袖”“耕地”“拥护”等词语是谢晓燕稳态词表中所不具备的。由于语料时间跨度差异巨大，计算方法并不完全相同，造成一些词语选择的差异并不奇怪。较高的重合率也从另一个角度验证了本文词集的性能。

6 结论与展望

本文在基于 70 年跨度的历时语料库，借鉴文本分类、停用词抽取等技术中的算法对各年度语料进行分析，获得稳态词的候选集。通过历时文本分类性能、时间敏感性、重叠性质和语料覆盖程度的考察遴选出了最优的算法和时间颗粒度设定：TF-IDF 方法和月颗粒度。

TF-IDF 方法统一了常用程度和时间分布两种重要的语言属性。逆文档频是对时间分布的刻画，对词频进行了时间的修正，从而在共时常用之外展现了时间维度的“常用性”。

最优候选集共包含 3011 个词，其中动词略多于名词，各占约三分之一，平均词长不足 1.7 字，前密后疏得分布于历时语料库总频序表的前 7609 位，覆盖了全部语料的近九成。稳态词中包含大量构造句子结构的核心词语。它们塑造了稳态词在词长和词类上的特性。稳态词的提取可以加深对语言生活底层与基础词汇的认识。稳态词的提取对于汉语教学、中文信息处理、语言规划和词典编纂都具有重要意义。

参考文献

[1] 张普．论语言的稳态[J]．郑州大学学报（哲学社会科学版），2008(02)．

[2] Fukumoto F, Suzuki Y, Takasu A. Timeline adaptation for text classification[C]// ACM International Conference on Information & Knowledge Management. 2013:1517-1520.

[3] Degaetanoortlieb S. Feature Discovery for Diachronic Register Analysis: a Semi-Automatic Approach[C]// Proceedings of International Conference on Language Resources and Evaluation (LREC'12):2786-2790.

[4] 谢晓燕. 基于 26 年《深圳特区报》的稳态词语提取与考察研究[D]. 北京语言大学博士学位论文，2010.

[5] 荀恩东，饶高琦，肖晓悦，臧娇娇. 大数据背景下 BCC 语料库的研制[J]. 语料库语言学，2016，3(1):93-118.

[6] 荀恩东，饶高琦，谢佳莉，黄志娥. 现代汉语词汇历时检索系统与应用研究[J]，中文信息学报，2015(3): 169-176.

[7] K. Sparck-Jones (1972). A statistical interpretation of term specificity and its application in

retrieval[J]. Journal of documentation, 28(1):11-21.
[8] S. E. Robertson, K. S. Jones. Relevance weighting of search terms[J]. Journal of American Society of Information Science, 27(3):129-146.
[9] C. E. Shannon, A mathematical theory of communication[J]. Bell System Technical Journal, vol. 27, pp. 379-423and 623-656, July and October, 1948.
[10] T. M. Cover and J. A. Thomas, Elements of Information Theory, 1991 John Wiley & Sons, Inc.
[11] Xu Y, Jones G J F, Li J T, et al. A study on mutual information-based feature selection for text categorization[J]. Journal of Computational Information Systems, 2007, 3(3): 1007-1012.
[12] 顾益军，樊孝忠，王建华,等. 中文停用词表的自动选取[J]. 北京理工大学学报，2005, 25(4):337-340.
[13] 关高娃. 蒙古文停用词和英文停用词比较研究[J]. 中文信息学报, 2011, 25(4):35-38.
[14] Lo T W, He B, Ounis I. Automatically Building a Stopword List for an Information Retrieval System.[J](2005). Journal of Digital Information Management, 3(1)3-8.
[15] 冯志伟，胡凤国. 数理语言学[M]. 北京：商务印书馆, 2012: 255.
[16] I. Rosengren, The quantitive concept of language and its relation to the structure of frequency dictionaries[J]. Etudes de Linguistiques Applique, 1971(1):103-127.
[17] Huarui Zhang, Churen Huang, Shiwen Yu, Distributional Consistency: A general method for defining a core lexicon[C]// Proceedings of International Conference on Language Resources and Evaluation (LREC'04).
[18] 教育部语言文字信息管理司. 中国语言生活状况报告[M]，北京：商务印书馆，2004-2015.
[19] Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition). Press. Morgan Kaufmann.
[20] 国家汉语水平考试委员会《汉语水平词汇等级大纲》，经济科学出版社，2001.

作者简介：饶高琦（1987），博士，主要研究领域为计算语言学、语言政策与语言规划，Email:raogaoqi-fj@163.com，手机 18813128566；李宇明（1955），通信作者，教授，主要研究领域为语言学理论、语法学、儿童语言学与语言规划，Email:liyum@263.net。