文章编号：1003-0077（2011）00-0000-00

先秦词汇的时代特征自动获取及文献时代的自动判定*

刘 浏[1]，李 斌[1,2]，曲维光[3]，陈小荷[1]

（1. 南京师范大学 语言信息科技研究中心，江苏 南京 210097；

2. 南京大学 计算机软件新技术国家重点实验室，江苏 南京 210093；

3. 南京师范大学 计算机科学与技术学院，江苏 南京 210097）

摘要：词汇的时代特征能反应词汇在一个时代发展变化的规律。本文将先秦分为前春秋、春秋和战国三个时代，获取并研究这三个时代的时代独有词、时代特征词及时代发源词。本文提出两种自动判断先秦文献时代的方法，分别基于向量相似度和朴素贝叶斯分类器，在 25 种先秦文献上后者的分类性能更稳定。最后本文使用朴素贝叶斯分类器验证了《列子》并非成书于先秦。

关键词：先秦词汇；时代；向量空间模型；朴素贝叶斯分类器

中图分类号：TP391　　**文献标识码：**A

The Automatic Acquisition of Pre-Qin Word's Property of Times and The Automatic Classification of Document's Times

Liu Liu[1], Li Bin[1,2] , Qu Wei-guang[3], Chen Xiao-he[1]

(1. Research Center of Language and Informatics, Nanjing Normal University, Nanjing, Jiangsu 210097, China; 2. State Key Laboratory for Novel Software Technology, Nanjing University Nanjing, Nanjing, Jiangsu 210093, China; 3. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: Words' property of times shows rules of how a word changes in a particular times. We divide the Pre-Qin times into three parts as Pre-Chunqiu, Chunqiu and Zhanguo. We find out and focus on three kinds of words which are only in a times, popular in a times and arised in a times. We also propose methods using VSM and Naive Bayes Classifier to decide the times of a text with which we experiment on 25 texts of Pre-Qin. The latter one's result turn out much better. With the same method we verified that "LieZi" is not written in Pre-Qin.

Key words: Pre-Qin Words, Times, VSM, Naive Bayes Classifier

1 引言

先秦在史学中是秦朝以前时代的统称，以公元前 221 年秦始皇统一六国为界限。这一时代看似简略，却包含了夏、商、西周，以及春秋、战国长达 1800 年的历史。由于是华夏文明的开始阶段，时代久远，大部分历史都只能从古代流传的史籍和现代出土的文物中寻得一丝端倪，真正留下大量典籍反映当时文化的是春秋战国时代。我们使用的 25 种先秦文献也主要是春秋战国时代的文献，对于先秦词汇时代特征的研究，主要就是对春秋战国这一时期

* **收稿日期：2013-6-30　　定稿日期：2013-7-15**

基金项目：国家社科基金（10CYY021、10&ZD117）；江苏省哲社重点研究基地课题（2010JDXM023）；南京大学计算机软件新技术国家重点实验室开放课题（KFKT2011B03）；中国博士后基金（2012M510178）；江苏省博士后基金（1101065C）；江苏高校优势学科建设工程；江苏省普通高校研究生科研创新计划项目（CXLX12_0357）

作者简介：刘浏（1989—），男，硕士生，主要研究方向为计算语言学；李斌（1980—），男，副教授，主要研究方向为计算语言学；曲维光（1964—），男，教授，主要研究方向为计算语言学。

词汇时代特征的研究。但即便是春秋战国时代，也包含了从公元前 770 年到公元前 221 年共 550 年的历史，这一历史时期汉语词汇依然存在着变化发展，对这一历史时期词汇时代特征的研究，也就是寻求这一时期汉语词汇的变化特征，以期能够在此基础上发现更多社会文化发展变化的特征。

词汇的时代特征是词汇意义重要的组成部分，对于其进行定量研究以助于更深一步的语义知识挖掘很有价值和必要性，本文立足于先秦文献，通过定量方法研究先秦词汇特点，分别基于向量相似度和朴素贝叶斯分类器，在 25 种先秦文献进行分类实验，发现在面向开放语料时，后者的性能更为稳定。

2 相关研究

有关先秦词汇的研究丰富而多样，主要见于以下几种类别：从词汇看语言的发展变化，如文献[1]；对词汇本身进行研究，如文献[2][3]；还有利用词汇信息研究古籍的成书年代，如文献[4]。综观这些研究，可以发现对于先秦词汇的研究目前还仅限于古汉语或词汇学等本体语言学领域，从语言信息处理角度看待并研究先秦词汇的并不多见，陈小荷在[5]中的《词汇概貌》一章详细介绍了利用语言信息处理手段获得的先秦词汇知识，是这一领域难得的研究成果。其中没有提及词汇在时代特征方面的研究，而这是本文主要研究目的所在。

有关词汇的时代特征，语言学界已有许多研究，这些研究多着力于发掘并描述词汇所具有的时代特征本身的性质或意义，如罗曼·雅克布森[6]曾指出：“语言社会往往把时间轴包括在那些可以直接感知的语言因素之内，比如，人们会感觉到语言系统中的陈旧成分是古旧的，新鲜成分是时髦的。”杨振兰[7]认为词语的时代特征“是词所体现出的某个历史时代特殊的时代氛围和时代气息，是社会的变化发展在语言词汇中的投影和映射。”“必须是反映了比较重要的社会历史内容的词，才具备一定的时代气息。”沈孟璎[8]认为时代色彩具备如“高频率、时效性、选择性、系列化、言文趋同化”等特征。王吉辉[9]认为，词语的时代特征不仅以其理性意义为基础，更与词语的使用状况紧密联系。利用词汇时代特征的性质特点，自动发掘词汇时代特征的研究见于文献[10]，其提出了对现代汉语词汇的时代特征自动获取的方法。

3 语料资源及时代划分

3.1 语料资源

我们选取了汉达文库[11]共 25 种先秦文献，包括《楚辞》《公羊传》《管子》《谷梁传》《国语》《韩非子》《老子》《礼记》《论语》《吕氏春秋》《孟子》《墨子》《商君书》《诗经》《孙子兵法》《吴子》《孝经》《荀子》《晏子春秋》《仪礼》《周礼》《周易》《庄子》《尚书》《左传》。25 种文献类型不一，成书的时代信息详尽程度也不一，为了保持时代数据的一致性，保证后续时代特征获取实验能够顺利完成，我们对每部文献的成书时代划定一个大致的区间。文献时代信息可考的，参照考证时代；不确切的，参考作者的时代；若作者时代不可考，对于史书可参考史书记录的时代；对于成书年代尚存疑的文献，如《孝经》《周礼》，我们选取较为可信的观点。这 25 种文献的成书时代大致情况见图 1：

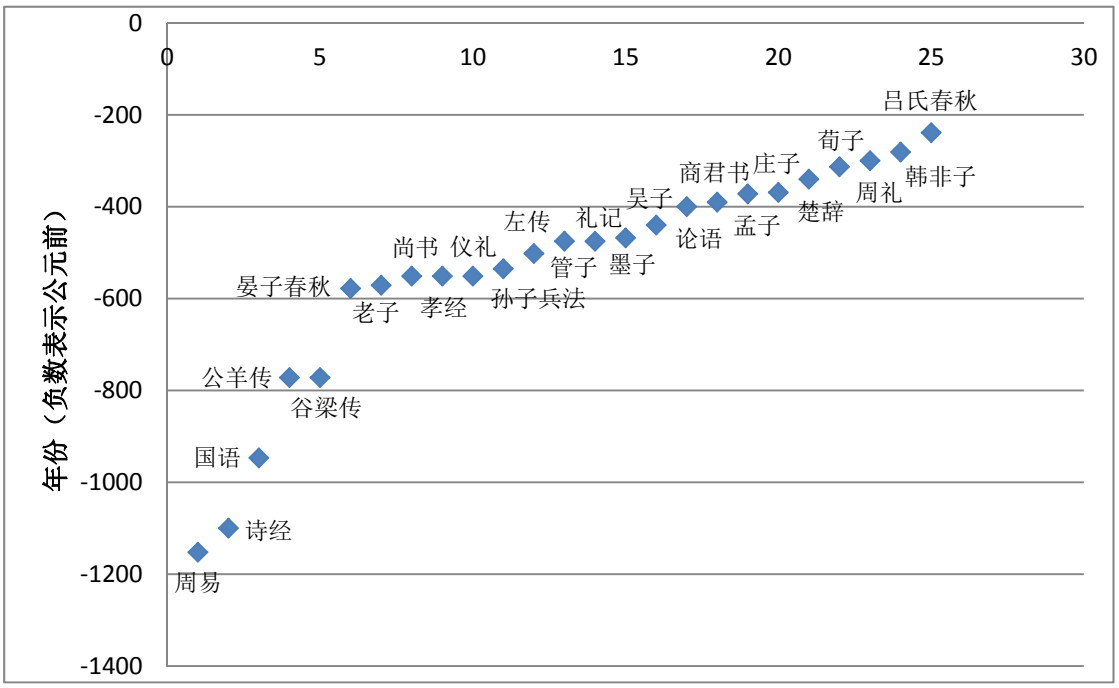

图 1：25 种先秦文献大致成书时代

3.2 时代划分

先秦文献年代的判定并不都是十分精确的，而且由于文献语料规模的限制，词汇的特征及其变化很难如现代汉语般鲜明地表现在一年甚至一个月上，因此我们按照先秦时代本身的特点以及语料规模的特点，将这 25 种按时代分为前春秋、春秋和战国三个时代区间。这三个时代各自包含的文献语料如表 1 所示：

表 1：25 种先秦文献时代划分

时代	文献
前春秋	《诗经》《周易》《国语》
春秋	《公羊传》《谷梁传》《孙子兵法》《孝经》《晏子春秋》《仪礼》《尚书》《左传》
战国	《楚辞》《管子》《韩非子》《老子》《礼记》《论语》《吕氏春秋》《孟子》《墨子》《商君书》《庄子》《吴子》《荀子》《周礼》

其中前春秋 3 部，占 12%；春秋 8 部，占 32%；战国 14 部，占 56%。

3.3 词汇概貌

对文献语料的分词以及词性标注是对文献词汇的研究的前提和基础。我们参照了石民[12]的方法，使用 CRF 模型对文献进行了分词以及词性标注。并在此基础上对划分出的先秦三个时代所包含的文献分别进行了词汇的频次统计，得到一个先秦文献的词频数据库，对于先秦文献词汇的时代特征研究都是基于该数据库进行的，见表 2：

表 2：先秦词频次

词	词性	前春秋	春秋	战国
之	u	2853	6936	24148
不	d	2925	8044	22207
也	y	2638	8690	20349
而	c	1965	5297	19999

25 部先秦文献总规模为 1221202 词，其中前春秋频次为 113238，占总频次的 9.27%；春秋词频次为 361188，占 29.58%；战国频次为 746776，占 61.15%。总的频次分布与各个时代文献数目大致相符合。

4 时代词语的获取

4.1 时代独有词

时代独有词，顾名思义，表示只属于一个时代的那些词汇。这样的词汇，其特征在于，从其产生到消亡的整个过程只是出现在一个时代的区间里[1]。我们获取这样的词汇，只需要严格按照定义，寻找那些在各个所属时代出现频率为 0 的那些词即可。这样的词汇，由于其具有的“独有性”的特点，对于古汉语尤其是词汇学方面的研究，具有特殊的研究价值，因此我们单独将这些词语摘录并建立数据库，见表 3：

表 3：春秋独有词示例

词	词性	词频次
爲	v	641
衛	ns	131
佐食	n	116
曷為	r	115

我们统计的各时代独有词中，前春秋独有词有 3291 例，春秋独有词有 10388 例，战国独有词有 20318 例，分别占各个时代总词次比例为：2.9%，2.88%和 2.72%。这是一个很有意思的现象，由于时代和语料两方面规模的限制，我们还不能够下一个确定的结论。但就已有语料的数据可以大胆猜测，那就是各个时代独有词汇占各个时代总词汇的比例是大致固定的，这个比例可能在 2.5%到 3%之间。

但从各个时代独有词汇的比例规模来看，独有词汇的数量还是比较庞大的，这与我们语感上预期的情况不太一致，原因在于这些独有词中，大部分词语的出现频次很低。我们统计各时代独有词中出现频次小于各时代总频次的 0.001%的那些词，发现三个时代这些“低频独有词”所占比例分别为 74.81%、88%和 96.8%。为何“低频独有词”会在独有词中占如此大比例的一部分，我们分析其主要原因是低频独有词中含有大量人名、地名等命名实体。这些命名实体往往是只会出现在一个时代的，若是不重要的往往只会出现少数几次或一次。

4.2 时代特有词

[1] 当然也有可能某些词汇在之后的某个时代又再度出现，这里的独有仅限先秦这一更大的时代区间而论。

时代特有词汇从概念上说，应该是显著包含并表现了这个时代所特有信息的一类词。根据[7][8][9][10]等人的研究，词语的时代性（在这里就表现为时代特有词），主要是体现在词语的高词频这一特点上的。这类词不应在各个时代都是高频，应该只是在这些时代中的某一个时代区间内高频率。据此，我们筛选每个时代那些词频是别的时代词频 5 倍以上的词[2]。得到先秦三个时代各自的时代特有词，并建立数据库，如表 4 所示。

表 4：春秋特有词示例

词	词性	前春秋词频	春秋词频	战国词频
拜	v	0.0002649	0.0041336	0.0003964
賓	n	0.0003886	0.0037294	0.0002732
西	f	0.0002031	0.0028849	0.0002424
主人	n	0.0000088	0.0028960	0.0002544

该方法得到的词语，不仅具有高频率的特点，而且限制了高频率的时代区间，因此获得的词语都满足“时代特有”这一特性及条件。通过该方法，我们得到前春秋特有词 367 个，春秋特有词 138 个，战国特有词 86 个。三个时代特有词呈逐步减少的趋势，可能是词汇的传承造成的。比如战国时代许多词语是从春秋时代沿袭下来的，这些词语一旦固化成常用词语，词频就不会发生太大的变化，因此通过词频比较的方法，也就很难从战国时代找到太多的时代特有词。这也说明战国时代较之春秋时代虽然社会生活发生了剧烈的变化，但语言尤其是词汇方面，依然表现出了一种稳定的延续性和传承性[3]，这种延续性和传承性在下文的“时代发源词”中将会进一步分析。

4.3 时代发源词

有些词汇是从某一个时代开始才出现的，这类词在发源的时代之前词频基本为 0，从某一个时代开始词频会有显著的提升，比如：“然後 c”在前春秋时代词频为 0；到了春秋时代，词频为 0.0000775；到了战国时代，其词频增长为 0.0001553。从这个例子可以看出“然後 c”这个词发源与春秋时代并逐渐通行的特点。通过词频的筛选，我们就可以获取这些时代发源词。由于先秦的时代我们只划分为三个时代，我们很难通过上述提出的方法，严格界定出发源于前春秋时代或战国时代的词语。因此这里我们只就发源于春秋时代的词语进行讨论。

通过我们的方法得到了一个时代发源词表如表 5 所示：

表 5：春秋发源词示例

词	词性	前春秋词频	春秋词频	战国词频
觶	n	0.0000000	0.0009053	0.0000121
某	r	0.0000000	0.0004928	0.0001433
然後	c	0.0000000	0.0000775	0.0001553
下士	n	0.0000000	0.0000028	0.0002759

观察词表可以发现，时代发源词也分几种情况：有些词语发源于并流行于一个时代，并

[2] 我们还排除了那些词频为 0 的词语，因为这些词语的特点已经时代独有词中体现了，并且还将在下文所述的“时代发源词”中进一步分析。

[3] 之所以不将前春秋词汇纳入这一比较范围，是因为其语料规模与另外两个时代相差较大。春秋和战国时代的语料规模基本相当，这更便于我们得出以上的结论。

在之后的时代继续保持着一定的使用频率，这样的词语从发源开始逐渐成为常用词的一部分，如“然後 c”；有些词语发源于并流行于一个时代，在之后的时代中虽也见使用，但频率远不及其发源的时代，这种词汇与时代独有词和时代特有词均有相似的成分，但却又有明显的区别，因此我们并未将其算作时代独有词或时代特有词的特殊情况，而是作为时代发源词的一类，这类词如“鼾 n”；有些词语发源于某一个时代，但真正流行却是在之后的时代，这类词也有成为常用词或以后某个时代的时代特有词的可能性，如“下士 n”。

5 文献时代判定

文献时代的判定可以看作一种将文献划分为不同时代类别的文本分类的任务。我们实现并比较了两种文本分类的方法，一种是基于向量相似度的计算，一种是使用朴素贝叶斯分类器。下文将就这两种分类方法进行详细的说明和分析。

5.1 向量空间模型及文档特征选择

5.1.1 向量空间模型

向量空间模型（VSM）由 G.Salton[13]首先提出。给定任意一个文档 D，D 可以表示为 D(t1,w1;t2,w2; …;tn,wn)，其中各特征项 tk (1<k<n)互异且各特征项 tk 无先后顺序关系。可以把特征项 t1,t2,…,tn 看作一个 n 维坐标系，权重 w1,w2,…,wn 看作每个特征项在对应维度上的坐标值，于是，便可以把一个文档表示为 n 维空间中的一个向量。

5.1.2 x^2 统计量

文档的特征项可以由字、词、短语等来表示，不论选取哪一种作为特征项，一篇文档的特征维度都会是非常高的，这样高维的向量不利于此基础上的进一步计算，因此特征项的选择至关重要。目前已有许多成熟的特征选择方法，如利用信息增益（IG）、x^2 统计量、互信息（MI）等方法[14]。本文通过实验比较，将词作为文档特征项，使用 x^2 统计量进行文档特征的选择。

“x^2 统计量（CHI）衡量特征项 ti 和类别 Cj 之间的关联程度，并假设 ti 和 Cj 之间符合具有一阶自由度的 x^2 分布。特征对于某类的 x^2 统计值越高，它与该类之间的相关性越大，携带的类别信息也较多，反之则越少。”[15][16][4]

“令 N 表示训练语料中文档的总数，A 表示属于 Cj 类且包含 ti 的文档频度，B 表示不属于 Cj 类但包含 ti 的文档频度，C 表示属于 Cj 类但不包含 ti 的文档频度，D 是既不属于 Cj 也不包含 ti 的文档频度。表 6 表示了这 4 种情况。”

表 6：特征与类关系示意图

特征项 \ 类别	Cj	~Cj
ti	A	B
~ti	C	D

特征项 ti 对 Cj 的 CHI 值为[5]：

[4] 关于 x^2 统计量的公式及表格均参考文献[18]。

$$\chi^2(t_i,\ C_j)=\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (1)$$

基于 CHI 统计量的的特征提取方法可以分别计算 ti 对每个类别的 CHI 值，然后在整个训练语料上计算，见公式（3），其中 M 为类别数。

$$\chi^2_{MAX}(t_i)=\max_{1<j<M}\{\chi^2(t_i,\ C_j)\} \quad (2)$$

通过计算 25 部文献每个词对于三个时代的 x 2 统计量，我们从训练语料 45238 个词例中选取了 6240 个词例作为特征项，这些特征项的值均大于 3.5[5]。

5.2 文献时代判定

5.2.1 基于向量相似度计算

我们把每一个文献都看做一个文档 Di,把每一个时代也看做一个文档 Dj,那么某一部文献是否属于一个时代，就可以用文档 Di 和文档 Dj 两个向量的相似度来计算。某一篇文献向量与哪一个时代向量的相似性最高，那么它就是属于这一个时代。向量相似度可以用向量夹角的余弦值来表示，如公式（3）所示：

$$\mathrm{Sim}(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n} w_{1k}\times w_{2k}}{\sqrt{(\sum_{k=1}^{n} w_{1k}^2)(\sum_{k=1}^{n} w_{2k}^2)}} \quad (3)$$

使用之前选定特征项，并用词频作为向量特征项的权重。计算这每一部文献向量与各个时代向量的相似度，选取相似度最高的那个时代作为该文献的成书时代，如公式（4）所示，T(Di)表示文献的成书时代，j 表示时代，得到的实验结果见表 7，8：

$$T(D_i)=\mathrm{argmax}_j\ \mathrm{Sim}(D_i,D_j) \quad (4)$$

表 7：25 部文献成书年代判定（基于向量相似度，封闭测试）

	前春秋	春秋	战国	微平均[6]	宏平均[7]
正确率	75.00%	100.00%	86.67%	87.50%	87.22%
召回率	100.00%	62.50%	92.86%	84.00%	85.12%
F 值	85.71%	76.92%	89.66%	85.71%	86.16%

表 8：25 部文献成书年代判定（基于向量相似度，开放测试）

	前春秋	春秋	战国	微平均	宏平均
正确率	0.00%	50.00%	73.33%	56.00%	41.11%
召回率	0.00%	37.50%	78.57%	56.00%	38.69%
F 值	0.00%	42.86%	75.86%	56.00%	39.86%

[5] 取该值综合考虑了特征项占总数的比例以及特征项统计量值的分布。
[6] 关于微平均和宏平均的计算，参阅文献[18]:p352-353。
[7] 同上。

由于语料规模较小，我们的开放测试是从 25 部文献中抽取 24 部训练之后，再对剩余的一本进行分类测试，测试结果是对 25 部文献分别用此方法得到的结果。

该实验中，判定结果取的是相似度最大的值对应的时代。根据夹角余弦公式，相似度为 1 时，两向量完全相同，为 0 时完全不相关，因此相似度值越接近 1，两向量越相似。我们的实验中得到的判断时代的相似度最大值的平均值在封闭测试中为 0.76，在开放测试中为 0.68，均比较接近于 1，这也验证了实验的有效性。

从封闭测试来看，该方法在前春秋和战国两个时代的文献判定上召回率很高，在春秋时代的正确率很高，而春秋时代的召回率较低，这个现象可能是由于春秋和战国两个时代之间词汇分布的差异并不是非常明显造成的。从开放测试来看，整体效果是不尽如人意的，原因很显然，是因为前春秋和春秋两个时代的语料规模远小于战国时代，特征项在战国时代的噪音信息较大，甚至掩盖了其在前春秋和春秋时代有价值的信息。若能提供更大规模的训练语料，该分类方法的性能会有显著的改善。

5.2.2 基于朴素贝叶斯分类器

朴素贝叶斯分类器是文本分类研究中最为普遍的一种分类器，其基本思想时利用特征项与类别的联合概率估计给定文档的个别概率，并且假定每个文档中的词与词之间是相互独立的，文本中词的出现只依赖与文本类别，不依赖其他词及文本长度。根据贝叶斯公式，文档 Doc 属于 Ci 类的概率见如下公式：

$$P(C_i|Doc) = \frac{P(Doc|C_i)\times P(C_i)}{P(Doc)} \quad (5)$$

使用词频 TF 表示向量 V 的特征权重，则该公式可以改写为：

$$P(C_i|Doc) = \frac{P(C_i)\prod_{t_j\in V} P(t_j|C_i)^{TF(t_j,Doc)}}{\sum_j P(C_j)\prod_{t_i\in V} P(t_i|C_j)^{TF(t_i,Doc)}} \quad (6)$$

其中 TF(ti,Doc)是文档 Doc 中特征 ti 出现的频度，P(tj | Ci)是 Ci 文档中特征 tj 出现条件概率的拉普拉斯概率估计。上述公式可以简化为只求解$P(C_i)\prod_{t_j\in V} P(t_j|C_i)^{TF(t_j,Doc)}$部分，得到最大值的$P(C_i|Doc)$，即是文档 Doc 的年代。为了便于计算，我们将计算$P(C_i)\prod_{t_j\in V} P(t_j|C_i)^{TF(t_j,Doc)}$的最大值转换为计算$-\log_2 P(C_i)\prod_{t_j\in V} P(t_j|C_i)^{TF(t_j,Doc)}$的最小值。

根据朴素贝叶斯分类器对 25 部文献进行的时代分类实验结果如下表 9,10 所示：

表 9：25 部文献成书年代判定（基于朴素贝叶斯分类器，封闭测试）

	前春秋	春秋	战国	微平均	宏平均
正确率	100.00%	100.00%	82.35%	88.00%	94.12%
召回率	100.00%	62.50%	100.00%	88.00%	87.50%
F 值	100.00%	76.92%	90.32%	88.00%	90.69%

表 10：25 部文献成书年代判定（基于朴素贝叶斯分类器，开放测试）

	前春秋	春秋	战国	微平均	宏平均
正确率	50.00%	60.00%	76.47%	69.23%	62.16%
召回率	66.67%	37.50%	92.86%	72.00%	65.67%
F 值	57.14%	46.15%	83.87%	70.59%	63.87%

封闭集和开放集的选取同基于向量相似度计算的实验。从表中可见，不论是面向封闭语料还是开放语料，朴素贝叶斯分类器的分类性能都要远优于单纯利用向量相似度的计算方法。但是朴素贝叶斯分类器也表现出了明显的对语料的依赖性，这与基于向量相似度的计算是类似的，语料规模最大的战国时代总体性能远好于规模较小的另外两个时代，但语料的分布不均匀也影响了分类实验的结果。但我们预计在更优质的语料条件下，该分类方法的性能还有很大的提升空间。

朴素贝叶斯分类器之所以比向量相似度方法的性能高出很多，原因在于向量相似度的计算方法需要将每个文献文本与每个时代文本做相似度计算，由于我们的语料规模限制，每个时代文本的质量并不高，因此判断某一个文献是否属于某个时代的准确率也就不高了。朴素贝叶斯分类器不存在这种问题，其直接利用条件概率估算每个文献文本“符合”各个时代的条件概率，即使语料规模并不大，也能胜任我们的分类任务。

5.3《列子》的成书年代判定

关于《列子》一书，学界一直存有争议，主要在于现存《列子》究竟是战国列子原著，还是魏晋之士伪作，甚或是东晋张湛自作自注[17]。

由于目前基于向量相似度计算的方法面向开放语料分类效果并不很好，因此我们使用朴素贝叶斯分类器判定方法对《列子》[8]年代进行判定。使用向量相似度计算，《列子》与三个时代的相似度分别为 0.42,0.38,0.48，虽然与战国时代的相似度最高，但由于这个相似度的值远小于之前实验的平均值 0.60，因此通过此方法，《列子》成书于先秦的可能性不高。使用朴素贝叶斯分类器，求得的列子成书于先秦三个时代的概率的负对数值分别为 3.9E5，4.2E5 和 4.3E5，也远大于之前开放测试的平均最小值 5.9E4，因此该方法也验证了《列子》成书于先秦的可能性不高。两种方法均认为《列子》成书于先秦的可能性不高，如果有可靠的魏晋时代分词语料，将可以更有力的判定《列子》的成书年代是否真的是魏晋时代，限于论文篇幅，这里不做赘述。

6 结论

本文从时代独有词、时代特有词和时代发源词三个角度分别研究了先秦词汇时代特征。将先秦分为前春秋、春秋和战国三个时代，自动获取了各个时代具有时代特征的三类词语，并对这些词语的分布和特征做了进一步的分析和讨论。之后文章使用向量空间模型和朴素贝叶斯分类器分别自动判定文献时代，封闭测试的结果是较好的，但由于语料规模的限制，开放测试虽逊色于封闭测试，但基本也是令人满意的。

由于我们选取的语料限于 25 种文献，得出的某些结论较为有限，但是本文主要是旨在探索一种获取词汇的时代特征及文本时代判定的方法。在语料的规模及范围得到充分拓展的情况下，本研究提出的方法将可以得到更加科学和严谨的结论。有关文献年代的自动判定，使用其他文本分类的方法是否效果更好，这也是以后要研究的主要方面。

[8] 这里所用是《列子》字频表，而不是词频表。

参考文献

[1] 谭书旺.从《孟子章句》看战国至东汉的语言发展[J].古汉语研究，2001，2：62-66.
[2] 吴宝安，黄树先.先秦“皮”的语义场研究[J].古汉语研究，2006，2：69-72.
[3] 叶南.《尔雅》与先秦语言研究[J].西南民族学院学报（哲学社会科学版），1996，s6：74-77.
[4] 谢祥娟.从词汇角度看《晏子春秋》的成书年代[J].中南大学学报(社会科学版)，2011，8：207-210.
[5] 陈小荷,冯敏萱,徐润华等.先秦文献信息处理[M].北京:世界图书出版公司北京公司,2013：146-168.
[6] 罗曼•雅克布森，泼沫斯卡.雅克布森文集[C]//钱军.北京：商务印书馆，1980：130-144.
[7] 杨振兰.词的时代色彩初探[J].山东大学学报，1988，3：102-106.
[8] 沈孟璎.论词语时代色彩的主要特征[J].内蒙古民族师院学报，1991，3：24-29.
[9] 王吉辉.词语的时代色彩与词语的使用[J].理论与现代化，2001，2：72-77.
[10] Liu Liu, Li Bin, etc. Automatic Acquisition of Chinese Words' Property of Times[J]. Chinese Lexical Semantics. Lecture Notes in Computer Science, 2013, Volume 7717:154-165.
[11] 汉达文库.先秦文献[DB/OL].http://www.chant.org/.2010.
[12] 石民，李斌，陈小荷.基于 CRF 的先秦汉语分词标注一体化研究[J].中文信息学报，2010，2：39-45.
[13] Salton G. The SMART Retrieval System. Experiments in Automatic Document Processing[M]. Prentice Hall, 1971:115-411.
[14] Yang, Y., Pedersen, J.P.. A Comparative Study on Feature Selection in Text Categorization[C]. Proceedings of 14th International Conference on Machine Learning, 1997:412-420.
[15] Dunning, T. Accurate Method for the Statistics of Surprise and Coincidence[J]. Computational Linguistics, 1993, 19(1):61-74.
[16] 宗成庆.统计自然语言处理[M].北京:清华大学出版社，2008：340-353.
[17] 王光照，卞鲁晓. 20 世纪《列子》及张湛注研究述略[J]. 安徽大学学报(哲学社会科学版)，2008，3：14-19.

作者联系方式：
姓名：刘浏
地址：南京师范大学 语言信息科技研究中心，江苏 南京 210097
联系方式：13655197379
邮箱：liuliu1989@gmail.com