“甚”的语法功能的聚类分析与词性演变模型的建立

王秋月

（清华大学）

摘要：本文运用多元统计分析方法，对先秦两汉时期至魏晋南北朝时期的十二部古典文献中的“甚”字的语法功能进行了聚类分析。分析表明，在古典文献中，“甚”的语法功能主要有两大类：一是在先秦两汉时期，“甚”的主要语法功能是作谓语；二是在魏晋南北朝时期，“甚”的语法功能主要是作状语。从“甚”的语法功能演变可以推论出“甚”的词性演变，即在先秦两汉时期“甚”是形容词，而到了魏晋南北朝时期，“甚”的词性主要是副词。聚类分析方法不仅适用于汉语词性演变的建模研究，而且对古籍鉴定也有重要的参考价值.

关键词：聚类分析方法，汉语词性演变模型，古籍断代鉴定

Study on modeling of SHEN’s character development by analysis of its grammatical function using hierarchical clustering method

Wang Qiuyue

(Tsinghua University)

Abstract: Statistical analysis on the grammatical function of SHEN used in 12 ancient books from Qin and Han Dynasty to Wei and Jin Dynasty is carried out by hierarchical clustering method. The results show that SHEN has two grammatical functions: (1) predicate during Qin and Han Dynasty; (2) adverbial modifier in the Wei and Jin Dynasty. The model of SHEN’s character development is inferred from the evolution of its grammatical function. SHEN is an adjective in Qin and Han dynasty, while it is an adverb during Wei and Jin dynasty. The hierarchical clustering method is useful not only in modeling the Chinese character development, but also in determining the publishing period of ancient book.

Key words: Hierarchical clustering method, Chinese character development model, division of ancient book into periods

一. 引言

“物以类聚，鸟以群分”。分类学是人类认识世界的一门基础学科。分类学发展至今经历了从多半凭经验和本学科的专业知识来进行分类的定性分类阶段到利用数学方法进行更科学的分类的数值分类学两个阶段。近几十年来，多元统计分析技术的引进，从数值分类学中逐渐分离出了聚类分析这个新分支。聚类分析已广泛应用于天气与气侯预报、服装标准的制订、考古以及传统的地学研究[1] [2] [3]。

聚类分析应用于语言学研究则是随着计算语言学的诞生而得到迅速发展。美国伊利诺斯大学的郑锦全教授开创了计算机在汉语音韵和方言学上的应用研究[5]。特别是大型汉语语料库的建立，基于语料库的统计和规则相结合的汉语计算语言学研究方法，为汉语的自然语言理解与处理研究提供了强有力的工具，并取得了丰硕的成果[6]。

本文试图采用聚类分析方法， 对先秦两汉至魏晋南北朝时期的十二部典籍中的“甚”字的语法功能进行统计分析， 由“甚”的语法功能演变推论出其词性演变规律， 从而以“甚”为例，探讨建立汉语词性演变模型的方法， 为汉语的计算语言学研究与古文献鉴定等奠定基础。

二. “甚”的语法功能的统计与分析

研究汉语词性演变规律可以从考察词的语法功能的演变入手。以“甚”字为例。为了建立“甚”的词性演变模型， 对先秦两汉至魏晋南北朝时期的十二部典籍中的“甚”字的全部用例，按照语法功能进行统计， 结果如表 1 所示。其中， 1- 8 号文献是先秦时期的作品； 9 号文献是东汉时期的作品； 10-12 号文献是魏晋南北朝时期的作品。

表 1 的统计结果表明，在上古汉语中， “甚”的语法功能大体上可分为两大类：一是在先秦两汉时期， “甚”的语法功能主要是作谓语和状语，作谓语的占用例的 50%以上，例如：天之爱民甚矣（《左传 · 襄公十四年》），作状语的占用例的 30%以上；例如：动刀甚微（《庄子 · 养生主》），二是在魏晋南北朝时期， “甚”的语法功能发生了重大变化，其作状语的用例高达 90%以上，例如：佛法甚微（《法显传》）；众生甚苦（《百喻经》）。值得注意的是，《谷梁传》在先秦两汉作品中有特殊性的一面，其中“甚”的语法功能主要是作谓语。

在古代汉语中，形容词既能作谓语，又能作状语。因此， “甚”在先秦两汉时期是形容词。而到了魏晋南北朝时期， “甚”已演变成了副词[4]。

上述关于“甚”字的语法功能分类与词性演变分析主要是基于定性分类方法的，能否应用多元统计分析中的聚类分析方法得出同样的结论呢?

表 1. “甚”的语法功能统计表

文献号	文献名	作谓语	作宾语	作定语	作补语	作状语	合计
1	左传	48	3	3	3	27	84
2	公羊传	11			3	5	19
3	古梁传	23			1		24
4	论语	3	1		1		5
5	孟子	16	4	1	1	3	25
6	墨子	13		2		26	41
7	庄子	15	2	1		21	39
8	荀子	22	2	2	2	27	55
9	论衡	27	4	4	16	18	69
10	法显传				1	7	8
11	世说新语		1		4	135	140
12	百喻经	2				16	18

三. “甚”的语法功能的聚类分析

系统聚类法(Hierachical Clustering Methods)是一种寻找样品或变量的自然类别的数值分类法。对样品的聚类称作 Q 型聚类，对变量的聚类称作 R 型聚类。系统聚类法的基本思想是：先将 n 个样品各自看成一类，然后计算各类之间的归类指数。根据归类指数的大小衡量两类之间的密切程度。将关系最密切的两类并成一个新类，直至所有样品都成为一类为止，形成一幅表示各样品之间亲疏关系的谱系图。通常选用距离系数作为归类指数，并且在两类并成一个新类时，以两类之中的归类指数的最小者作为新类的归类指数，即最小距离法。

设每个类(或样品)中有 P 个变量，我们可把一个类(或样品)看作 P 维空间的一个点，那么 n 个类(或样品)就是 P 维空间中的 n 个点。因此，可用 P 维空间中的 n 个点之间的欧几里德(Euclid)距离来表示 n 个类(或样品)之间的靠近程度。定义类与类之间的距离：

$$\overline{d_{ij}} = \sqrt{\sum_{k=1}^{p}\left(x_{ki} - x_{kj}\right)^2} \qquad (1)$$

其中 $i, j = 1,2,\ldots,n$； n 为类(或样品)数目； $k = 1,2,\ldots,p$； p 为每类中的变量数目。

在实际应用中，通常取距离系数

$$d_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{p}\left(x_{ki} - x_{kj}\right)^2} \qquad (2)$$

作为距离系数。距离系数越小，两类之间的关系越密切，可以并成一个新类。

对于表 1， 按式（ 2 ）的距离系数定义，分别计算各文献之间的距离系数 d_{ij}，得到“甚”的语法功能距离系数表，即表 2 。最后，按最小距离法聚类后得到“甚”的语法功

能距离系数 Q 型聚类谱系图，如图 1 所示。表 3 给出距离系数 Q 型聚类分析图联接表。

由图 1 可见，“甚”的语法功能主要分两类：一是在先秦两汉时期的作品中“甚”的语法功能比较接近，可归为一类；二是在魏晋南北朝时期的作品中“甚”的语法功能比较接近，归为另一大类。同时，聚类分析也将“甚”在《谷梁传》中的特殊语法功能分离出来。聚类分析的结果与语法专家的定性分析结果完全一致。

表 2. “甚”的语法功能距离系数表

文献号	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00000	0.06473	0.22610	0.17854	0.11024	0.18178	0.12932	0.10781	0.12295	0.35875	0.38521	0.32794
2	.06473	.00000	.21294	.14931	.11432	.21613	.16840	.14236	.09737	.37699	.41075	.35647
3	.22610	.21294	.00000	.19671	.16908	.40435	.35330	.33329	.29410	.58154	.60803	.54948
4	.17854	.14931	.19671	.00000	.09466	.33603	.28253	.25918	.16502	.48399	.52085	.47099
5	.11024	.11432	.16908	.09466	.00000	.28139	.22537	.20518	.16078	.45034	.47907	.42420
6	.18178	.21613	.40435	.33603	.28139	.00000	.05811	.07771	.20103	.18791	.20629	.14812
7	.12932	.16840	.35330	.28253	.22537	.05811	.00000	.02883	.16246	.16246	.25795	.20045
8	.10781	.14236	.33329	.25918	.20518	.07771	.02883	.00000	.13576	.25220	.27797	.22173
9	.12295	.09737	.29410	.16502	.1650	.20103	.16246	.13576	.00000	.33118	.37288	.32662
10	.35875	.37699	.58154	.48399	.45034	.18791	.23669	.25220	.33118	.00000	.05886	.07505
11	.38521	.41075	.60803	.52085	.47907	.20629	.25795	.27797	.37288	.05886	.00000	06148
12	.32794	.35647	.54948	.47099	.42420	.14812	.20045	.22173	.3266	.07505	.06148	.00000

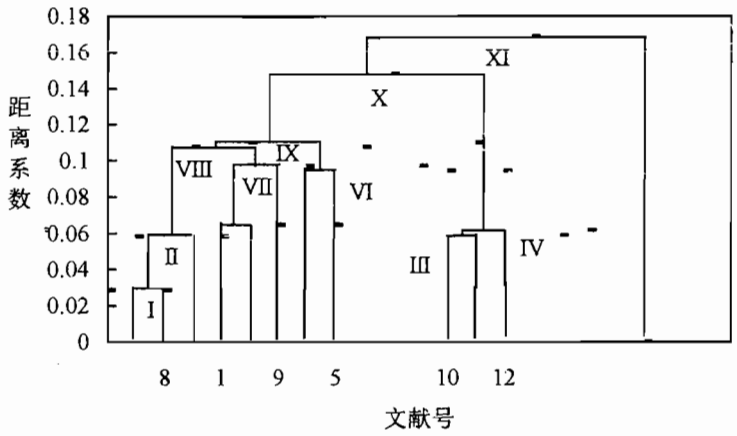

图 1. 甚”的语法功能距离系数 Q 型聚类谱系图

表 3. “甚”的语法功能距离系数 Q 型聚类分析图联接表

聚类序号	文献号	文献号	距离系数
I	7	8	0.02883
II	6	I	0.05811
III	10	11	0.05886
IV	12	III	0.06148
V	1	2	0.06473
VI	4	5	0.09466
VII	9	V	0.09737
VIII	II	VII	0.10781
IX	VIII	VI	0.11024
X	IX	IV	0.14812
XI	X	3	0.16908

四. 用聚类分析方法建立汉语词性演变模型

大型汉语语料库的建立为汉语的计算语言学研究开辟了新路。由于语料库中蕴含丰富的经过分词与词性标记处理的不同历史时期的熟语料，并提供字、词、句的统计与分析功能，所以基于语料库的汉语词性演变研究不仅简便易行，而且具有足够的典型性与可信度。

我们知道，词性的演变集中体现在其语法功能的变化。因此， 只要能建立词的语法功能演变模型， 再根据语法规则， 就可推论出词性演变模型。汉语词性演变模型的建立可分以下三个步骤:

（ 1 ）基于语料库， 统计词的语法功能分布;

（ 2 ）基于聚类分析方法， 建立词的语法功能演变模型;

（ 3 ）基于规则， 运用语法知识， 将词的语法功能演变模型转化为词性演变模型。

以“甚”的词性演变建模为例，首先在大型语料库上，统计出各不同历史时期的代表性作品中“甚”的语法功能分布，得到象表 1 一样的“甚”的语法功能统计表；其次，应用聚类分析技术，将“甚”的语法功能分布进行聚类，可知在上古汉语中，“甚”的语法功能可分为两大类，即在先秦两汉时期“甚”的语法功能主要是作谓语和状语；在魏晋南北朝时期，“甚”的语法功能主要作状语。最后，根据古代汉语语法知识，建立推理规则，即:

规则 1： 如果一词的语法功能既能作谓语，又能作状语，则这种词是形容词;

规则 2： 如果一个词的语法功能主要是作状语，即作状语占用例的 80%以上，则这种词是副词。

由规则 1 和规则 2 可推出“甚”的词性演变模型为: “甚”在先秦两汉时期是形容词，“甚”在魏晋南北朝时期已演变成副词，用数学形式表示为:

$$f(t)=\begin{cases}\text{形容词}, & t=\text{先秦两汉}\\ \text{副词}, & t=\text{魏晋南北朝}\end{cases} \qquad (3)$$

其中， $f(t)$ 为“甚”的词性函数； t 为自变量，表示年代。

五. 基于词性演变模型的古文献断代方法

中华民族五千年灿烂文化为我们留下了大量珍贵的古典文献。保护和整理这些宝贵的文化遗产历来受到国家的重视。但是，由于老专家奇缺，为古籍鉴定与整理工作带来了很多困难。如何应用现代的计算机与信息技术进行这方面 工作已显得十分重要。

在古籍鉴定与整理工作中，断代是一项重要的工作。以往的断代方法主要凭老专家的经验，如从装裱方式、纸张质地、印刷方式、篇章内容等方面对古籍进行鉴定。能否从计算语言学的角度，对文献进行计算机处理，从而进行断代呢?汉语词性演变模型的建立为计算机辅助古籍鉴定工作的开展奠定了基础。

我们仍以“甚”字为例。“甚”在先秦两汉时期是形容词；“甚”在魏晋南北朝时期是副词。如果我们对一本古籍中所出现的“甚”的语法功能进行统计，得出其词性规律，则可根据“甚”的词性演变模型判断该古籍的写作年代，为古籍鉴定提供数值分析方法。

当然，为了将词性演化模型用于古籍鉴定，不仅需要细化词性演化模型，即进一步建立年代更为详尽的词性演化模型；而且要建立多个词的词性演化模型，综合寻优，给出正确的鉴定结果。

六. 结论

系统聚类法为汉语词的语法分类研究与词性演变建模提供了数值分析手段。基于语料库的词的语法功能统计和应用聚类分析方法的数值分类，以及根据语法知识的规则推理，为词的词性演化模型的建立提供了计算语言学方法。同时，词性演化模型的建立也为古籍鉴定提供了特征模式。

这种基于语料库与语法规则的聚类分析建模方法不仅适用于汉语词性演变模型的建立，而且也适用于其它自然语言的词性演化模型的建立。

参考文献

[1]. 张尧庭，方开泰，《多元统计分析引论》，科学出版社，1982 年.

[2]. 陆巍： 北辛文化和山东龙山文化陶器成分的聚类分析，《考古》，1996 年第 11 期.

[3]. 苏炳凯，《大气科学中的统计诊断与预测》，南京大学出版社，1996 年.

[4]. 李杰群，“甚”的词性演变，《语文研究》，1986 年第 2 期.

[5]. 陆致极，计算机在汉语音韵和方言学上的应用，《国外语言学》，1986 年.

[6]. 罗振声， 清华大学 ZN 大型通用汉语语料库的研究，《中文信息》，1996 年 1 月.