用主成分分析法研究短语字段的判别因素

俞小娟[1],胡金柱[1],李 琼[2],周毕吉[2]
(1.华中师范大学 计算机科学系,湖北 武汉 430079;
2.华中师范大学 语言教育研究中心,湖北 武汉 430079)

摘 要:对复句层次结构和层次关系进行分析和研究之前,首先要确定有标复句中分句的数量,即有标复句中的哪些字段是分句,哪些字段只是加了标点符号的句法成分(文中称之为短语字段)。结合语言学的相关理论,提取出识别短语字段的因素,并对这些因素进行主成分分析,从而得出进行识别的综合影响因素以及与原始的具体因素之间的关系。结果表明,前三个主成分所包含的信息量接近85%,已包含原有因素大部分的信息,在今后的研究中,这三个主成分将取代原来的多个变量,从而简化研究的复杂度。

关键词:主成分分析;变量;相关系数
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2008)10-0116-04

Studying Factors of Judging Phrase Fields by Method of Principal Component Analysis

YU Xiao-juan[1],HU Jin-zhu[1],LI Qiong[2],ZHOU Bi-ji[2]
(1.Department of Computer Science, Huazhong Normal University,Wuhan 430079,China;
2.Centre for Language Education, Huazhong Normal University,Wuhan 430079,China)

Abstract:Before studying the hierarchical structure and relations of the complex sentence, the number of the clauses in a complex sentence which has been tagged should have been confirmed, and which fields are the complete clauses and which are not but phrases due to the complexity of using punctuations in Chinese are the same. In the meanwhile it increases the difficulty of the later classifying the category and the layer of the tagged complex sentences. According to the linguistics theory this thesis tells the method of how to dig out and analyze the factors of finding out the phrase fields. By principal component analysis, it can not only reduce the complexity of the problem but also find out the relations of the factors. The result suggests that first three main components contain most of the information, so can use these three main components instead of the original variables to reduce the complexity of further work.

Key words:principal component analysis; variable; coefficient

0 引 言

复句由两个或两个以上意义上相关、结构上互不作句子成分的分句组成。分句是结构上类似单句而没有完整句调的语法单位[1]。复句中的各分句之间一般有停顿,书面上用逗号或分号、冒号表示;但是用逗号或分号、冒号结尾的字段并不一定就是分句。汉语中标点符号使用的灵活性,一方面增强了语言的表达效果,另一方面也使一些句子成分独立成一个字段,给确定复句中分句的数量带来了困难。

在研究复句的层次结构和层次关系之前,首先要理清分句的结构[2],包括分句的数量和每个分句包含的字段数。目前的研究是先将含谓语的分句和单独成一字段的句子成分区分开来,从识别非分句字段着手。为研究方便,将复句中独立成为一个字段的句子成分定义为短语字段。

1 规则提取

对短语字段中的明显形式标志进行研究,总结了一些基于词类信息的规则[2],用以提取识别短语字段的因素。

规则如下:

① 以时间副词或方位词结尾的状语性成分单独为一个字段,则标注该字段为短语字段。

收稿日期:2008-01-09
基金项目:软件工程国家重点实验室开发基金资助项目(SKLSE04-018);湖北省重点科技攻关资助项目(2003AA101C26)
作者简介:俞小娟(1982-),女,硕士研究生,研究方向为计算语言学;胡金柱,教授,博士生导师,研究方向为软件工程和计算语言学。

② 以介词或类此介词的成分开头的字段,一般情况下,也作为状语成分独立成一个字段,则该字段也标为短语字段。

③ 字段中不含任何动词和形容词,该字段一般是独立成字段的句子成分,标为短语字段。

④ "的 + 名词"结构结尾的字段中"的"前面没有动词的字段,一般不作为一个完整的分句,同样标为短语字段。

⑤ "的 + 名词"结构结尾的字段中"的"前面内容有引号包含,一般不作为一个完整的分句,同样标为短语字段。

2　基于规则的主成分分析过程

2.1　变量设置

根据目前总结出来的判断短语字段的规则,抽取相应的属性作为主成分分析[3]的变量,用已经做好分词和词性标注的大规模语料库中随机抽取的 5000 句复句中的短语字段作为训练集[4],分析各变量之间的关系以及变量组合用以短语字段判断所含的信息量。

设置变量如下:

① $x1$:以时间词或方位词结尾;

② $x2$:以介词或类似介词的成分开头;

③ $x3$:不含动词和形容词;

④ $x4$:本字段中以"的 + 名词"结构结尾;

⑤ $x5$:本字段中以"的 + 名词"结构结尾并且该结构前不含动词;

⑥ $x6$:本字段中以"的 + 名词"结构结尾并且该结构前的部分用引号包含;

⑦ $x7$:所含动词的数量。

2.2　变量属性值的数值化

在考察短语字段的明显形式标志时,都是些概念性的描述,因此抽取出相关变量后,需要对所设的变量的属性值进行数值化处理。

采用计算机和人工统计结合的方法,将具有上述每种属性的字段分别搜索出来放入文本中,再对每个文本的材料做人工的判断是否为短语字段,统计出具备每种属性同时为短语字段的成功率,将这个概率作为该属性对应变量的属性值[5]。

相应的数据统计如下:

$x1$:以时间词或方位词结尾。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 79.6%。

$x2$:以介词或类似介词的成分开头。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 56.1%。

$x3$:不含动词和形容词。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 85.2%。

$x4$:本字段以"的 + 名词"结构结尾。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 16.4%。

$x5$:本字段中以"的 + 名词"结构结尾并且该结构前不含动词。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 83.9%。

$x6$:本字段中以"的 + 名词"结构结尾并且该结构前的部分用引号包含。

统计结果:具有该属性的字段为短语字段的数量占具有该属性的字段总数的 76.7%。

$x7$:所含动词的数量。

该属性值按照各字段的具体情况设置。

例如复句:[1930 年 /t10 月 /t,/w] [在 /p 纽约 /ns 举行 /v 的 /u 一 /m 次 /q 数学 /n 讨论 /v 会上 /t ,/w] 数学家 /n 科尔 /nr 通过 /p 演算 /vn 证明 /n 226/m 一/w 一1/m 是 /a 合数 /n ,/w 而 /c 不 /d 是 /v 二 /m 百年 /q 来 /f 一直 /d 被 /p 人们 /n 怀疑 /v 的 /u 质数 /n 。/w

短语字段 1:1930 年 /t 10 月 /t,/w 属性值如下:

$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$
79.6%	0	85.2%	0	0	0	0

2.3　主成分分析过程

以大规模语料库中随机抽取的 5000 句复句中的短语字段及其对应的属性值作为主成分分析的数据源,对其进行主成分分析。

表 1 给出的是部分原始矩阵数据。

表 1　部分原始数据

编号	典型短语字段	X1%	X2%	X3%	X4%	X5%	X6%	X7
1	10 月 /t31 日 /t,/w	79.6	0	85.2	0	0	0	0
2	11/n 一 /w12 月 /t,/w	79.6	0	85.2	0	0	0	0
3	"/w 三八 /n 画会 /n"/w 的 /u 作品 /n, /w	0	0	85.2	16.4	83.9	76.7	0
4	在 /p 泄露 /v 了 /u 即将 /d 戒严 /v 的 /b 机密 /a 之后	79.6	56.1	85.2	0	0	0	2

根据原始矩阵和以下公式求得相关系数矩阵 $\boldsymbol{R}$:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_i)'$$

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n}x_{ki}$$

$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$ 其中$(i = 1,2,\cdots,p)$

注:$X = (x_{ij})$ $S = (s_{ij})$

得出相关系数矩阵 $\boldsymbol{R}$ 见表 2。

表 3 为所求得的主成分提取分析表。

表 3 主成分提取分析表

Component	Total	of Variance %	Cumulative %
1	2.766	39.511	39.511
2	1.962	38.022	67.533
3	1.156	16.512	84.045
4	0.558	7.977	92.022
5	0.269	3.842	95.864
6	0.193	2.762	98.626
7	9.619E-02	1.374	100.00

表 4 为所求得的初始因子载荷矩阵。

3 分析所得的数据

3.1 相关系数矩阵

相关系数矩阵如下:

① 变量 $x4$ 和 $x5$、$x6$ 的相关性非常密切,即短语字段的属性中以"的 + 名词"结构结尾的属性和以"的 + 名词"结构结尾并且"的"前面没有动词的属性、以"的 + 名词"结尾且该结构前用引号包含的属性关系是非常密切的,也就是说目前的语料中提取出来的短语字段如果是以"的 + 名词"结构结尾的多数是该结构前面不含动词的和该结构前面用引号包含的。

② 变量 $x1$ 和 $x4$、$x5$、$x6$ 是负相关性比较强的,显然如果短语字段中以时间词或方位词结尾,则不具备以"的 + 名词"结构结尾这个属性。

③ 变量 $x3$ 和 $x7$ 是负相关性比较强的,这个特征比较明显,即如果该字段不含动词和形容词,则该字段的动词的个数就是 0 个。

3.2 主成分分析表和初始因子载荷矩阵

各主成分分布如下:

$F1 = -0.923x1 + 9.829 * 10^{-2}x2 - 0.151x3 + 0.959x4 + 0.745x5 + 0.636x6 + 3.602 * 10^{-2}x7$

方差贡献率 39.511%,累计贡献率 39.551%

$F2 = -2.77 * 10^{-2}x1 + 0.384x2 - 0.892x3 - 2.84 * 10^{-2}x4 - 0.389x5 + 9.662 * 10^{-2}x6 + 0.926x7$

方差贡献率 28.022%,累计贡献率 67.533%

$F3 = 2.291 * 10^{-2}x1 + 0.836x2 + 0.221x3 - 1.70 * 10^{-2}x4 + 0.399x5 - 0.490x6 + 8.539 * 10^{-2}x7$

方差贡献率 16.512%,累计贡献率 84.045%

$F4 = 0.198x1 + 0.338x2 + 0.190x3 - 0.119x4 - 0.102x5 + 0.584x6 - 5.55 * 10^{-2}x7$

方差贡献率 7.977%,累计贡献率 92.022%

$F5 = 0.236x1 - 0.168x2 + 0.127x3 + 9.205 * 10^{-3}x4 + 0.268x5 + 5.354 * 10^{-2}x6 + 0.306x7$

方差贡献率 3.842%,累计贡献率 95.864%

$F6 = -0.154x1 - 8.83 * 10^{-3}x2 + 0.284x3 + 8.201 * 10^{-2}x4 - 0.214x5 - 3.83 * 10^{-2}x6 + 0.187x7$

方差贡献率 2.762%,累计贡献率 98.686%

$F7 = 0.167x1 + 2.409 * 10^{-2}x2 - 1.19 * 10^{-2}x3 + 0.242x4 - 7.71 * 10^{-2}x5 - 3.53 * 10^{-2}x6 - 4.44 *$

表 2 相关系数矩阵

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007
VAR00001	1.000	-0.049	0.191	-0.879	-0.605	-0.472	-0.032
VAR00002	-0.049	1.000	-0.132	0.042	0.178	-0.122	0.358
VAR00003	0.191	-0.132	1.000	-0.147	0.277	-0.183	-0.730
VAR00004	-0.879	0.042	-0.147	1.000	0.688	0.537	0.045
VAR00005	-0.605	0.178	0.277	0.688	1.000	0.206	-0.248
VAR00006	-0.472	-0.122	-0.183	0.537	0.206	1.000	0.049
VAR00007	-0.032	0.358	-0.730	0.045	-0.248	0.049	1.000

表 4 初始因子载荷矩阵

	Component						
	1	2	3	4	5	6	7
VAR00001	-0.923	-2.77E-02	2.291E-02	0.198	0.236	-0.154	0.167
VAR00002	9.829E-02	0.384	-0.836	0.338	-0.168	-8.83E-03	2.409E-02
VAR00003	-0.151	-0.892	0.221	0.190	0.127	0.284	-1.19E-02
VAR00004	0.959	-2.84E-03	-1.70E-02	-0.119	9.205E-03	8.201E-02	0.242
VAR00005	0.745	-0.389	0.399	-0.102	0.268	-0.214	-7.71E-02
VAR00006	0.636	9.662E-02	-0.490	0.584	5.354E-02	-3.83E-02	-3.53E-02
VAR00007	3.602E-02	0.926	8.539E-02	-5.55E-02	0.306	0.187	-4.44E-02

$10^{-2}x7$

方差贡献率 1.374%，累计贡献率 100%

前面三个主成分的方差贡献率接近 85%，说明前三个主成分已经包括了全部指标的大部分信息，则这三个综合指标在之后的研究中就可以用来取代原先的七个指标，降低了问题的复杂度。

3.3 模型分析

在第一主成分的表达式中，变量 $x1$、$x4$、$x5$、$x6$ 的系数较大，则这个主成分主要是用来反映短语字段是否以时间词或方位词结尾、是否以“的 + 名词”结构结尾、是否以“的 + 名词”结构结尾并且“的”前面没有动词、是否以“的 + 名词”结构结尾并且“的”前面是用双引号包含的这四个属性的，这反映了在当前考查的语料中，以时间词或方位词结尾的和以“的 + 名词”结构结尾的短语字段占所有短语字段的绝大多数。

在第二主成分的表达式中，变量 $x3$ 和 $x7$ 的系数较大，则这个主成分是用来反映短语字段关于动词和形容词的数量，即从短语字段中的谓语成分的有无来判断是否为短语字段。

在第三主成分的表达式中，变量 $x2$ 的系数较大，并且 $x3$、$x5$、$x6$ 的系数占一定的大小，则这个主成分是用来反映短语字段是否以介词打头并以方位词或名词结构结尾这个特征的。

4 结束语

将数学中多元统计的方法和语言学的理论结合起来，来分析复句中用以判断短语字段的因素，从而找出各因素之间的相关性和对判断的综合影响，降低了问题的复杂性，为今后实现短语字段的自动识别做好基础工作。由于材料的有限性和提取领域的局限性，数据在很大程度上会失真，还存在很多的不足，这也是今后努力完善的方向。

在完善的基础上，针对更多的更全面的语料，利用主成分分析所得的结果，欲利用提取出来的主成分采用聚类的方法[6]对语料库中的短语字段进行分类识别，努力实现短语字段的自动识别，为之后的复句的层次结构和关系标注做好准备工作。

参考文献：

[1] 邢福义. 汉语复句研究[M]. 北京:商务印书馆, 2001.

[2] 邢福义. 汉语语法学[M]. 长春:东北师范大学出版社, 1996.

[3] 罗积玉，邢　瑛. 经济统计分析方法及预测[M]. 北京:清华大学出版社,1987.

[4] 俞士汶. 计算语言学概论[M]. 北京:商务印书馆,2003.

[5] 刘　颖. 计算语言学[M]. 北京:清华大学出版社,2002.

[6] 谷　波，李济洪. 基于 COSA 算法的中文文本聚类[J]. 中文信息学报,2007(6):65－70.

(上接第 115 页)

实例比较分析证实此算法减少了候选集的数量，提高了运算速度，节省了系统空间，在一定程度上突破了传统算法的性能瓶颈，在预测分析和智能决策中具有一定的意义。

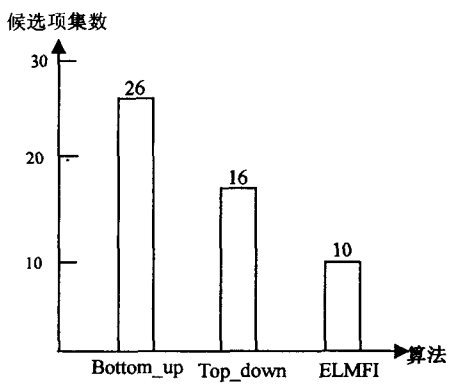

图 1　算法比较

参考文献：

[1] 毛国君，段立娟，王　实，等. 数据挖掘原理与算法[M]. 北京:清华大学出版社,2005.

[2] Kantardzic M. 数据挖掘[M]. 北京:清华大学出版社,2003.

[3] 王　丹，张　浩，陆剑峰. 针对高项频繁集的关联规则改进算法[J]. 计算机工程,2006,24(32):29－31.

[4] 王创新. 关联规则提取中对 Apriori 算法的一种改进[J]. 计算机工程与应用,2004(34):183－185.

[5] 李海军. 数据挖掘在 GIS 中的应用[D]. 北京:北京化工大学,2004.

[6] 邵峰晶，于忠清. 数据挖掘原理与算法[M]. 北京:水利水电出版社,2003.

[7] 章　艳，刘美玲，张师超，等. Apriori 算法的三种优化方法[J]. 计算机工程与应用,2004(36):191－193.

[8] Han J, Kamber M. 数据挖掘:概念与技术[M]. 北京:机械工业出版社,2001.

[9] 宋　雨，赵建利，王保义. 关联规则挖掘中最大频繁集的双向查找算法[J]. 华北电力大学学报,2005,32(2):67－71.

[10] 李清峰，杨路明，张晓峰，等. 数据挖掘中关联规则的一种高效 Apriori 算法[J]. 计算机应用与软件,2004,21(12):84－86.

[11] 刘桂庆，胡学钢，李　凯. CR 一种逆向的关联规则挖掘算法[J]. 微电子学与计算机,2004,21(9):83－86.