2. В каждом предложении выделяются пары текстовых единиц с общим словом, связанные определительным, аппозитивным или атрибутивным отношением. Эти пары объединяются в именные группы.

3. В каждом предложении выделяются пары текстовых единиц, связанные предикатным отношением. Главная текстовая единица относится к носителю p_{ij} отношения r_{ij}. Для зависимой единицы устанавливается ее участие в именной группе.

4. В каждом предложении выделяются пары текстовых единиц, связанные обстоятельственным отношением. Для главной текстовой единицы устанавливается ее участие в предикатном отношении r_{ij}. Зависимая текстовая единица относится к атрибуту a_{ij} отношения r_{ij}, если она относится к категории «наречие». В противном случае ее участие устанавливается в именной группе, а семантический падеж v_p на основании грамматической характеристики $f_m^{(k)}$.

5. На выходе алгоритма формируется набор пар отношений $r_{ij}=<u_i, u_j, p_{ij}, a_{ij}, v_p>$, характеризующих весь контент.

Выводы:

1. Предложено формальное описание синтаксической модели предложения и семантической структуры контента.

2. Разработан алгоритм перехода от синтаксической модели к семантической.

3. Показана взаимосвязь между синтаксическими характеристиками единиц предложения и структурными единицами предложенного смыслового описания контента.

СПИСОК ЛИТЕРАТУРЫ

1. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста. // Информационные технологии. 2002. № 7. С. 30-34.

2. Гладкий А.В., Мельчук И.А. Грамматики деревьев I. Опыт формализации преобразований синтаксических структур естественного языка. Сб. Инф. вопросы семиотики, лингвистики и автоматического перевода. М.: Наука, 1971.

3. Гладкий А.В. Формальные грамматики и языки. М.: Наука, 1973.

Сергей Александрович БАРДАСОВ —
доцент кафедры экономики
и управления собственностью
Тюменского государственного университета,
кандидат физико-математических наук

УДК 519.224.2

ПРИМЕНЕНИЕ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАЙКА ДЛЯ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНОГО ЧИСЛА ГРУПП ВАРИАЦИОННОГО РЯДА

АННОТАЦИЯ. Информационный критерий Акайка применяется для определения оптимального числа групп при построении гистограммы. Рассматриваются случаи интервалов равной длины и равной вероятности. Для распределения Гаусса получены формулы, позволяющие рассчитать число групп по заданному объему выборки.

The Akaike information criterion is applied to definition of optimum number of classes when histogram is under construction. The cases of intervals of equal length and equal probability are considered. For Gaussian distribution the formulas allowing to calculate number of classes on the given sample size are received.

Обычно для оценки числа групп распределений близких к распределению Гаусса применяют формулу Стерджесса [1]:

$$m=1+\log_2 n,$$

которая не имеет достаточного обоснования, хотя имеются и другие более обоснованные формулы [2, 3]. В данной работе для оценки числа групп применяется информационный критерий Акайка.

Пусть имеется n значений случайной величины x, ранжированных в порядке возрастания, $x_1 \le x_2 \le \ldots \le x_n$. Построим интервальный вариационный ряд, то есть образуем m групп. Определим для полученного ряда функцию максимального правдоподобия, зависящую от числа групп m, по формуле

$$L(m)=L\left(x_1, x_2, \ldots, x_n; m\right)=\left(\frac{f_1}{n\,\lambda_1}\right)^{f_1}\times\left(\frac{f_2}{n\,\lambda_2}\right)^{f_2}\times\ldots\times\left(\frac{f_m}{n\,\lambda_m}\right)^{f_m}, \qquad (1)$$

где f_i — количество значений признака в i-ой группе $\left(\sum_{i=1}^{m} f_i = n,\ f_i \ge 1\right)$,

λ_i — длина i-го группового интервала $\left(\lambda_i > 0,\ \sum_{i=1}^{m}\lambda_i = \Lambda = x_n - x_1\right)$.

Таким образом, если какое-либо значение признака оказалось в i-ой груп

то плотность вероятности в точке x полагается равной $\left(\frac{f_i}{n\,\lambda_i}\right)$. При определени

функции (1) полагаем, что при построении ряда не должно быть пустых групп и групповых интервалов, равных нулю.

В случае равных интервалов имеем:

$$L(m)=\left(\frac{f_1}{n\lambda}\right)^{f_1}\times\left(\frac{f_2}{n\lambda}\right)^{f_2}\times\ldots\times\left(\frac{f_m}{n\lambda}\right)^{f_m}=\frac{f_1^{f_1}\times f_2^{f_2}\times\ldots\times f_m^{f_m}}{n^n\lambda^n}=$$

$$=\frac{f_1^{f_1}\times f_2^{f_2}\times\ldots\times f_m^{f_m}\times m^n}{n^n\Lambda^n},\quad \lambda=\frac{\Lambda}{m}.$$

Натуральный логарифм функции правдоподобия равен

$$ln(L(m))=\sum_{i=1}^{m} f_i \ln(f_i)+n\ln(m)-n\ln(n)-n\ln(\Lambda),$$

$$ln(L(m))=\sum_{i=1}^{m} f_i \ln(f_i)+n\ln(m)+const, \qquad (2)$$

где *const* — слагаемые, не зависящие от числа групп.

Согласно информационному критерию Акайка, значение параметра m должно быть таким, чтобы выражение

$$-2\ln(L(m))+2m \tag{3}$$

приняло минимальное значение. Использование коэффициента 2 в формуле (3) является общепринятым.

Не принимая в расчет величины, не зависящие от числа групп m, подставим (2) в (3), поменяем знак и разделим на 2. Получится, что в случае равных интервалов оптимальное число групп $\hat{m}$ равно:

$$\hat{m}=\arg\max_{m}\left(\sum_{i=1}^{m} f_i \ln(f_i)+n\ln(m)-m\right) \tag{4}$$

Рассмотрим случай равновероятных (равночастотных) интервалов:

$$L(m)=\left(\frac{f}{n\lambda_1}\right)^f\times\left(\frac{f}{n\lambda_2}\right)^f\times\ldots\times\left(\frac{f}{n\lambda_m}\right)^f=\frac{f^n}{n^n(\lambda_1\times\lambda_2\times\ldots\times\lambda_m)^f}=$$

$$=\frac{1}{m^n(\lambda_1\times\lambda_2\times\ldots\times\lambda_m)^{\frac{n}{m}}},\quad f=\frac{n}{m}.$$

Логарифмируя, получим

$$\ln(L(m))=-\frac{n}{m}\sum_{i=1}^{m}\ln(\lambda_i)-n\ln(m). \tag{5}$$

Тогда, согласно критерию Акайка (3), оптимальное число групп равно:

$$\hat{m}=\arg\min_{m}\left(\frac{n}{m}\sum_{i=1}^{m}\ln(\lambda_i)+n\ln(m)+m\right). \tag{6}$$

Соотношения (4,6) пригодны для оценки числа групп различных распределений. Однако в общем случае поиск оптимального значения достаточно трудоемкий процесс. Поэтому целесообразно получить явные формулы для расчета числа групп по заданному объему выборки.

Пусть случайная величина подчиняется распределению Гаусса. Рассмотрим два варианта:

1) длины групповых интервалов равны;

2) частоты групповых интервалов равны.

Вариант 1. Рассмотрим стандартное нормальное распределение. Для проведения оценки вместо бесконечных пределов интегрирования будем полагать, что все значения случайной величины находятся в промежутке $[-3, 3]$, который разделим на m равных частей. Тогда число значений, оказавшихся в i-ой группе, будет равно

$$f_i=\frac{n\alpha}{\sqrt{2\pi}}\int_{-3+6(i-1)/m}^{-3+6i/m}\exp\left(-0{,}5t^2\right)dt$$

где $\alpha\approx 1{,}002707105$ — нормировочный множитель, введенный из-за конечности промежутка распределения.

Для каждого заданного количества данных n, которое варьировалось от 100 до 1000000, определялось число групп m, при котором выражение (4) принимало максимальное значение. Результаты представлены на рис. 1 и в табл. 1.

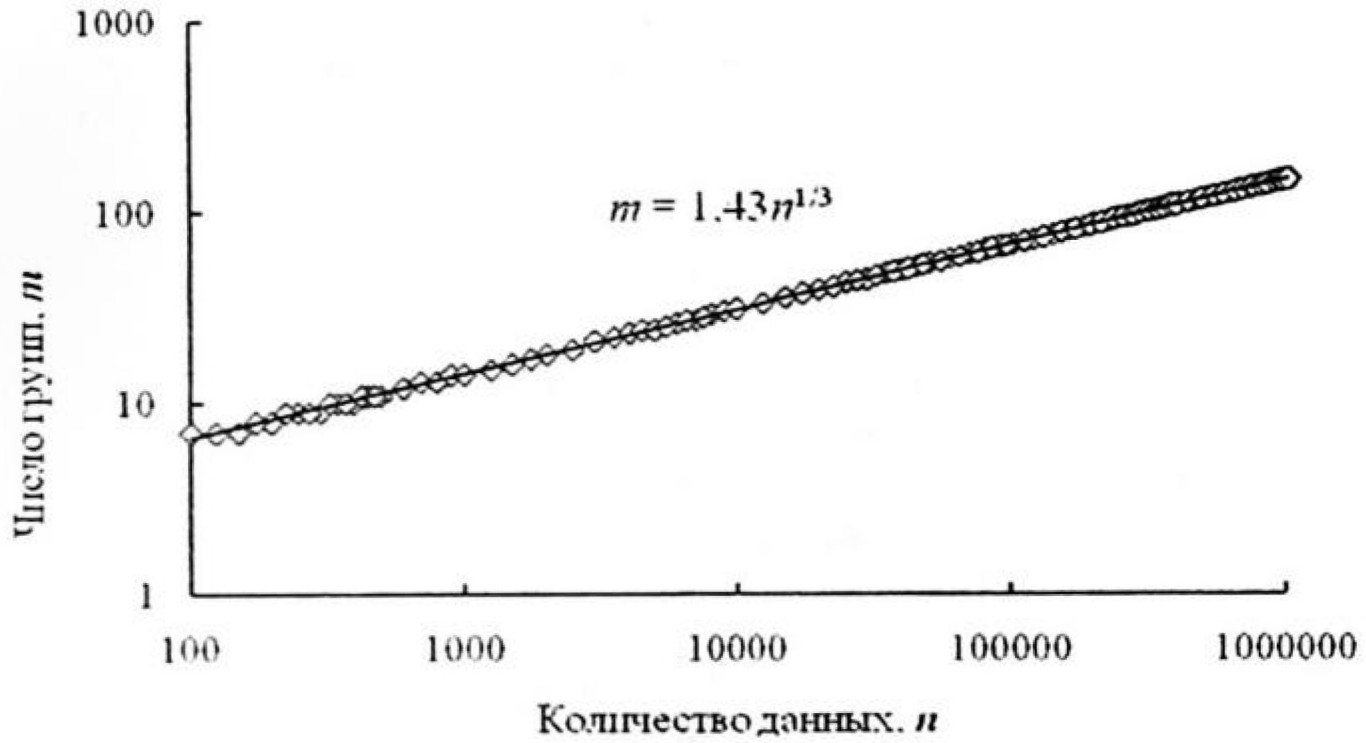

Рис. 1. Зависимость числа групп m от количества данных n в случае равных интервалов

Согласно рис. 1, в логарифмических шкалах мы имеем практически прямую линию, что соответствует степенной зависимости числа групп от количества данных. Некоторое отклонение точек от прямой связано с дискретностью значений обеих величин и со слагаемыми меньшего порядка малости. Регрессионный анализ показывает, что показатель степени очень близок к значению $^1/_3$. Кроме того, данное значение было получено в работах [2, 3]. Полагая, что показатель степени равен $^1/_3$, оценив параметр b_0 в зависимости $m=b_0n^{1/3}$ методом регрессионного анализа, получим:

$$m=b_0n^{1/3}.$$

Сплошная линия на рис. 1 соответствует зависимости (7).

Таблица 1

Оптимальное число групп m при заданном количестве данных n для интервалов равной длины

n	m	n	m	n	m	n	m	n	m	n	m
100	7	1250	15	22500	40	110000	68	320000	98	725000	128
125	7	1500	16	25000	42	120000	70	330000	99	725000	129
150	7	1750	17	27500	43	130000	72	340000	100	750000	130
175	8	2000	18	30000	44	140000	74	350000	101	775000	131
200	8	2500	19	32500	46	150000	76	360000	102	800000	132
225	9	3000	21	35000	47	160000	78	370000	103	800000	133
250	9	3500	22	37500	48	170000	79	380000	104	825000	134
275	9	4000	23	40000	49	180000	81	390000	104	850000	135
300	9	4500	24	42500	50	190000	82	400000	105	850000	136
325	10	5000	24	45000	51	200000	84	425000	107	875000	136
350	10	5500	25	47500	52	210000	85	450000	110	875000	137
375	10	6000	26	50000	53	220000	86	475000	111	900000	138
400	10	6500	27	55000	54	230000	88	475000	112	925000	139
425	11	7000	27	60000	56	240000	89	500000	113	925000	140
450	11	7500	28	65000	57	250000	90	525000	115	950000	140
475	11	8000	29	70000	59	260000	91	550000	117	950000	141
500	11	9000	30	75000	60	270000	92	575000	119	975000	142

n	m	n	m	n	m	n	m	n	m	n	m
600	12	10000	31	80000	62	280000	93	600000	121	1000000	142
700	13	12500	33	85000	63	280000	94	625000	122	1000000	143
800	13	15000	35	90000	64	290000	95	650000	124		
900	14	17500	37	95000	65	300000	96	675000	125		
1000	14	20000	39	100000	66	310000	97	700000	127		

Дальнейшие расчеты показали, что для распределения Гаусса независимо от величины стандартного отклонения оптимальное по критерию Акайке число групп необходимо рассчитывать по формуле (7). Для данного распределения практически все данные находятся в интервале длиной 6σ. Тогда для длины группового интервала получим оценку:

$$h = \frac{6\sigma}{m} = 4{,}2\,\sigma\, n^{-1/3}. \qquad (8)$$

Вариант 2. В случае интервалов равной вероятности для стандартного распределения Гаусса были получены результаты, представленные в табл. 2 и на рис. 2.

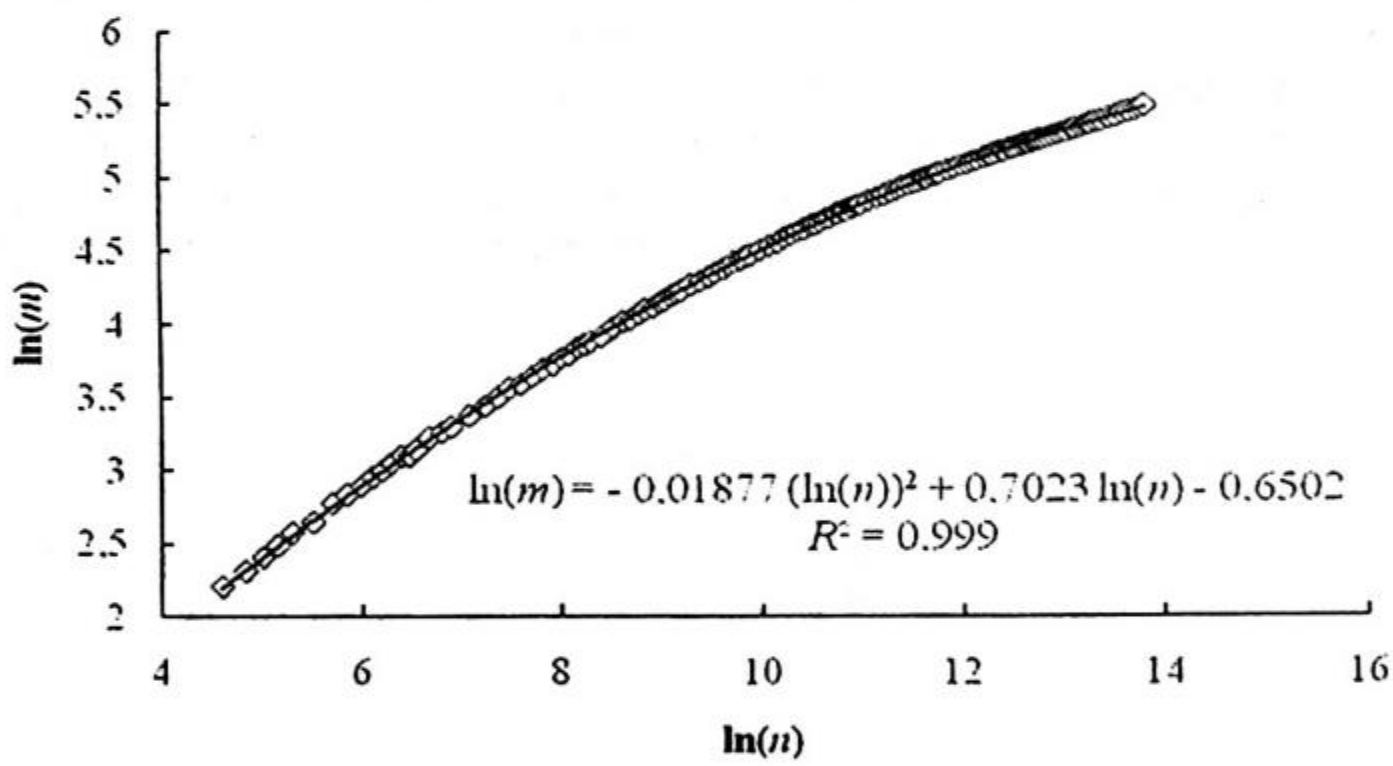

Рис. 2. Зависимость числа групп m от количества данных n в случае равных частот

Вероятность попадания значений признака во все интервалы, кроме двух крайних, была равна $1/m$. Длина крайних интервалов корректировалась для того, чтобы все интервалы уложились в отрезок $[-3; 3]$.

Согласно рис. 2, в случае равновероятностных интервалов число групп можно оценить по формуле:

$$m = \exp(-0{,}6502 + 0{,}7023 ln(n)) - 0{,}01877 ln^2(n)$$

или

$$n = 0{,}5219 n^{0.7023 - 0.01877 ln(n)}. \qquad (9)$$

Сплошная линия на рис. 2 соответствует формуле (9).

Таким образом, критерии (4), (6) позволили получить оценки (7), (9) для расчета оптимального числа групп величины, подчиняющейся распределению Гаусса.

Таблица 2

Оптимальное число групп m при заданном количестве данных n для интервалов равной частоты

n	m	n	m	n	m	n	m	n	m	n	m
100	9	2000	36	12000	73	48000	115	150000	157	380000	194
125	10	2250	38	13000	75	50000	117	160000	160	400000	196
150	11	2500	40	14000	77	52000	118	170000	162	420000	198
175	12	2750	41	15000	79	54000	119	180000	164	460000	202
200	13	3000	43	16000	81	56000	121	190000	166	480000	204
250	14	3250	44	17000	83	58000	122	200000	168	500000	206
300	16	3500	46	18000	84	60000	123	210000	170	520000	207
350	17	3750	47	19000	86	65000	126	220000	172	540000	209
400	18	4000	48	20000	87	70000	129	230000	174	560000	211
450	19	4500	50	22000	90	75000	131	240000	176	580000	212
500	20	5000	53	24000	93	80000	134	250000	177	600000	214
550	21	5500	55	26000	95	85000	136	260000	179	650000	217
600	22	6000	56	28000	98	90000	138	270000	180	700000	221
650	22	6500	58	30000	100	95000	140	280000	182	750000	224
700	23	7000	60	32000	102	100000	142	290000	183	800000	228
800	25	7500	61	34000	104	105000	144	300000	185	850000	231
900	26	8000	63	36000	106	110000	145	310000	186	900000	234
1000	27	8500	64	38000	107	115000	147	320000	187	950000	237
1200	29	9000	66	40000	109	120000	149	330000	188	1000000	240
1400	31	9500	67	42000	111	125000	150	340000	190	1000000	241
1600	33	10000	68	44000	112	130000	152	350000	191		
1800	35	11000	71	46000	114	140000	155	360000	192		

СПИСОК ЛИТЕРАТУРЫ

1. Sturges H. The choice of a class-interval // Journal of the American Statistical Association. 1926. V. 21. No 153. P. 65-66.

2. Scott D.W. On optimal and data-based histograms // Biometrika. 1979. 66. P. 605-610.

3. Freedman D., Diaconis P. On this histogram as a density estimator: L2 theory // Zeitschrift Wahrscheinlichkeitstheorie verwandte Gebiete. 1981. 57. P. 453-476.