ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ВЕСТНИК
ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

ОБЩЕНАУЧНЫЙ ПЕРИОДИЧЕСКИЙ ЖУРНАЛ

№ 293 Декабрь 2006

Серия «Информатика. Кибернетика. Математика»

Свидетельства о регистрации: бумажный вариант № 018694, электронный вариант № 018693
выданы Госкомпечати РФ 14 апреля 1999 г.
ISSN: печатный вариант – 1561-7793; электронный вариант – 1561-803X
от 20 апреля 1999 г. Международного центра ISSN (Париж)

СОДЕРЖАНИЕ

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

ТЕОРИЯ ВЕРОЯТНОСТИ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ИНФОРМАТИКА И ПРОГРАММИРОВАНИЕ

FEDERAL AGENCY OF EDUCATION

VESTNIC
TOMSK STATE UNIVERSITY

GENERAL SCIENTIFIC PERIODICAL

№ 293 December 2006

Series «Mathematics. Cybernetics. Informatics»

Certification of registration: printed version № 018694, electronic version № 018693
Issued by Russian Federation state committee for publishing and printing on April, 14, 1999.
ISSN: printed version – 1561-7793; electronic version – 1561-803X
on April, 20, 1999 by International centre ISSN (Paris)

CONTENTS

MATHEMATICAL MODELING

PROBABILITY THEORY AND MATHEMATICAL STATISTICS

INFORMATICS AND PROGRAMMING

УДК 519.237.8: 81'322.2

В.В. Поддубный, О.Г. Шевелев, Д.А. Бормашов

СРАВНЕНИЕ КАЧЕСТВА ПОДХОДОВ К КЛАСТЕРИЗАЦИИ ТЕКСТОВ НА ОСНОВЕ ГИПЕРГЕОМЕТРИЧЕСКОГО КРИТЕРИЯ

Работа поддержана грантом РФФИ 06-07-89320а

Рассматриваются подходы к кластеризации стилей текстов на основе гипергеометрического критерия при различных критических уровнях значимости, частотных признаках текстов и методах присоединения кластеров. На серии вычислительных экспериментов с использованием F-меры оценки качества кластеризации показано, что наилучшим является метод дальнего соседа, а наилучшими признаками оказываются частоты встречаемости биграмм и 500 самых распространенных слов.

Автоматическая кластеризация текстов используется при изучении схожести и различий текстов и групп текстов по каким-либо признакам, при поиске и сортировке электронных документов и в других актуальных на сегодняшний день задачах. В работе [1] был предложен метод кластеризации стилей текстов по частотным признакам на основе гипергеометрического критерия. Были рассмотрены две меры сходства стилей текстов для построения дендрограммы кластеризации – частота рассогласования и интегральная мера рассогласования. В качестве метода присоединения кластеров рассматривался метод дальнего соседа. В данной работе приводятся результаты дальнейших экспериментальных исследований методов кластеризации на основе гипергеометрического критерия. Исследуется качество кластеризации при использовании разных методов присоединения кластеров (дальний сосед, ближний сосед, медианный метод) при различных признаках (биграммы, 100 и 500 часто встречающихся слов, служебные слова Фоменко [3]) и при различных допустимых уровнях значимости критерия на примере кластеризации литературных текстов по авторскому стилю.

Кластеризация текстов на основе гипергеометрического критерия

Детали метода кластеризации на основе гипергеометрического критерия приведены в работе [1]. Этот метод используется для сравнения стилей двух текстов по одному частотному признаку. В ходе сравнения стилей по данному критерию проверяется нулевая гипотеза о том, что тексты имеют одинаковый стиль по данному признаку, против альтернативы – тексты различаются по стилю. Достигнутый уровень значимости критерия рассчитывается по формуле

$$p_0 = \sum_{x=\max(0,s-n_2)}^{\min(n_1,s)} \left\{ h(x \mid s, n_1, n_2) \le h(m_1 \mid s, n_1, n_2) \right\},$$

где $s = m_1 + m_2$, m_1 и m_2 – числа появления признака в первом и втором тексте, n_1 и n_2 – объемы текстов, $h(x \mid s, n_1, n_2) = C_{n_1}^{x} C_{n_2}^{s-x} / C_{n_1+n_2}^{s}$ – гипергеометрическое распределение, $x = \overline{\max(0, s-n_1), \min(n_2, s)}$. Статистикой критерия является наблюдаемое значение x, т.е. m_1. Решение в пользу альтернативы принимается при значении достигнутого уровня значимости p_0 критерия, меньшем или равном α (α – критическое, допустимое значение уровня значимости). При значении, большем α, оснований отвергнуть нулевую гипотезу нет.

Для проведения кластеризации набора из K текстов по L различным признакам предлагается следующее. Первоначально производится попарное сравнение всех текстов набора по гипергеометрическому критерию по каждому из L признаков. В результате таких сравнений получается K^2 достигнутых уровней значимости $\left\{ p_{0ij}, i, j = \overline{1, K} \right\}$ для каждого признака. Эти значения размещаются в L матрицах, из которых при фиксированном допустимом уровне значимости критерия делается соответственно L индикаторных матриц $y_{ij} = \left\{ \begin{matrix} 0, p_{0ij} > \alpha, \\ 1, p_{0ij} \le \alpha, \end{matrix} i, j = \overline{1, K} \right\}$, состоящих из нулей и единиц. Далее на основе меры расстояния строится матрица расстояний. В качестве меры расстояния между двумя текстами предлагается взять меру интегрального рассогласования, которая вычисляется следующим образом: $r_{ij} = \sum_{l=1}^{L} y_{ij}^{l}$. С использованием данной меры различие стилей пары текстов определяется суммой всех различий (единиц) по всем индикаторным матрицам. На основе матрицы расстояний производится иерархическая агломеративная кластеризация текстов: первоначально каждый текст является отдельным кластером, затем путем нахождения наиболее похожих по стилю текстов все тексты шаг за шагом объединяются в один кластер. Для проведения кластеризации необходимо выбрать метод присоединения кластеров. Отличие методов заключается в способе вычисления расстояния между кластерами. Рассмотрим среди этих методов три наиболее известных.

В методе дальнего соседа расстояние между кластерами вычисляется следующим образом:

$$c_{lk} = \arg\max_{t_l \in C_l, t_k \in C_k} \left(r\left(t_l, t_k\right) \right),$$

где $r(t_l, t_k)$– расстояние между текстом i, принадлежащим кластеру C_l, и k, принадлежащим кластеру C_k.

В методе ближнего соседа

$$c_{lk} = \arg\min_{t_l \in C_l, t_k \in C_k} \left(r\left(t_l, t_k\right) \right).$$

В медианном методе (другое название – UPGMA – Unweighted Pair-Group Method using arithmetic Averages [2]):

$$c_{lk} = \sum_{t_l \in C_l, t_k \in C_k} r(t_l, t_k) \Big/ \left(\mathrm{size}(C_l) \cdot \mathrm{size}(C_k)\right),$$

где $\mathrm{size}(C_l)$ и $\mathrm{size}(C_k)$ – число текстов в l-м и k-м кластерах соответственно.

В результате кластеризации получается дендрограмма (дерево) кластеризации. На дендрограмме отображается список текстов и структура объединения их в кластеры (рис. 1). Название текста формируется из имени его автора, названия произведения, к которому принадлежит текст, и номера текста. Линии дерева кластеризации показывают, какой текст и на каком уровне (расстоянии) присоединяется к тому или иному кластеру. Горизонтальный ряд чисел над деревом кластеризации показывает расстояния присоединения, которые равны значению интегральной меры рассогласования. Число признаков, по которым различаются объединяемые кластеры, – от 0 до L. Расстояния, на которых не произошло ни одного объединения, в целях экономии места пропускаются.

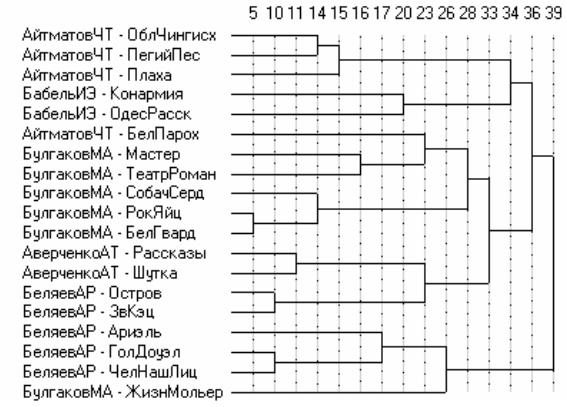

Рис. 1. Пример дендрограммы кластеризации

Оценка качества кластеризации

Оценку качества кластеризации в данной работе предлагается производить по F-мере [2], использующей априорное знание о том, каким образом должны группироваться тексты. Данная мера вычисляется следующим образом. Для каждого из кластеров $c \in C$ (на всех возможных уровнях кластеризации, включая первый – один текст, и последний – все тексты) вычисляется ряд частных мер $F_c^i = 2\Big/\left(1\big/p_c^i + 1\big/r_c^i\right)$, где p_c^i – точность определения априорно имеющегося кластера i (например, кластера текстов одного писателя) по отношению к фактически полученному кластеру c, а r_c^i – полнота определения кластера i по отношению к фактически полученному кластеру c. Причем $p_c^i = C_c^i \big/ N_c$, где C_c^i – число текстов априорного кластера i в фактическом кластере c, N_c – общее число текстов в кластере c. $r_c^i = C_c^i \big/ N_i$, где N_i – общее число текстов в априорном кластере i. Тогда $F = \sum_i \left(N_i \big/ N\right) \max\left(F_c^i\right)$, где N – общее число текстов, участвовавших в кластеризации. Данная величина лежит в интервале 0÷100% и позволяет оценить качество отдельной кластеризации (дендрограммы). Наибольшие частные меры F_c^i соответствуют фактически полученным кластерам, которые лучше всего представляют априорно имеющиеся кластеры.

ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Для оценки качества работы предложенного подхода рассмотрим ряд вычислительных экспериментов, проведенных при использовании различных методов присоединения кластеров, различных частотных признаков, уровней значимости и на разных наборах текстов. Проведем кластеризацию литературных текстов по авторскому стилю.

В качестве материала для исследований рассмотрим несколько множеств текстов известных классиков и современников. Так как известно, что качество кластеризации зависит от числа потенциальных кластеров, которые нужно найти, возьмем множества, состоящие из разного числа авторов. Кроме того, чтобы оценить, насколько сильно качество кластеризации зависит от состава текстов, возьмем также множества с текстами разных авторов, но при этом число авторов в этих множествах будет одинаковым.

Первое, самое большое, множество T_1 включает 133 текста 30 следующих авторов: Достоевский Ф.М., Гончаров И.А., Куприн А.И., Лесков Н.С., Салтыков-Щедрин М.Е., Толстой Л.Н., Тургенев И.С., Айтматов Ч.Т., Беляев А.Р., Булгаков М.А., Бунин И.А., Быков В.В., Горький М., Грин А.С., Ильф И. и Петров Е., Кассиль Л.А., Набоков В.В., Пантелеев Л., Шмелев И.С., Соболев Л.С., Сологуб Ф.К., Солженицын А.И., Толстой А.Н., Акунин Б., Довлатов С.Д., Лукьяненко С.В., Пелевин В.О., Поляков Ю.М., братья Стругацкие, Улицкая Л.Е.

Множества T_2–T_3 содержат тексты 10 авторов каждое. Авторы множества T_2 (XIX в.): Чехов А.П., Достоевский Ф.М., Гоголь Н.В., Гончаров И.А., Куприн А.И., Лесков Н.С., Пушкин А.С., Салтыков-Щедрин М.Е., Толстой Л.Н., Тургенев И.С. (48 текстов). Авторы множества T_3 (XX в.): Айтматов Ч.Т., Беляев А.Р., Булгаков М.А., Бунин И.А., Быков В.В., Горький М., Грин А.С., Ильф И. и Петров Е., Кассиль Л.А., Набоков В.В. (44 текста).

Множества T_4–T_6 содержат тексты 5 авторов каждое. Авторы множества T_4 (XIX в.): Чехов А.П., Достоевский Ф.М., Гоголь Н.В., Гончаров И.А., Лесков Н.С. (27 текстов). Авторы множества T_5 (XX в.): Айтматов Ч.Т., Аверченко А.Т., Бабель И.Е., Беляев А.Р., Булгаков М.А. (19 текстов). Авторы множества T_6 (современники): Акунин Б., Довлатов С.Д., Лукьяненко С.В., Пелевин В.О., Поляков Ю.М. (23 текста).

Рассмотрим следующие признаки стилей текстов: 1) частоты появления пар букв, исключая «ё», всего 1024 признака (биграммы); 2) частоты появления служебных слов, предложенных Фоменко в работе [3] (союзы, частицы, предлоги), всего 55 признаков; 3) частоты появления ста и 4) пятиста самых часто встречаемых слов по частотному словарю Шарова [4].

Проверим качество кластеризации при различных уровнях значимости критерия. Рассмотрим следующие критические уровни значимости α: 0,05; 0,001; 0,0001.

Первоначально рассмотрим качество кластеризации для множества T_1 (табл. 1).

Таблица 1

Качество кластеризации текстов из множества T_1

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
1	T_1	0,05	Служебные слова	Дальнего соседа	76,68
2	T_1	0,05	Служебные слова	Ближнего соседа	58,17
3	T_1	0,05	Служебные слова	Средней связи	78,16
4	**T_1**	**0,001**	**Служебные слова**	**Дальнего соседа**	**83,42**
5	T_1	0,001	Служебные слова	Ближнего соседа	49,75
6	T_1	0,001	Служебные слова	Средней связи	82,95
7	T_1	0,0001	Служебные слова	Дальнего соседа	81,75
8	T_1	0,0001	Служебные слова	Ближнего соседа	55,91
9	T_1	0,0001	Служебные слова	Средней связи	81,42
10	**T_1**	**0,05**	**Биграммы**	**Дальнего соседа**	**84,00**
11	T_1	0,05	Биграммы	Ближнего соседа	40,71
12	T_1	0,05	Биграммы	Средней связи	78,70
13	T_1	0,001	Биграммы	Дальнего соседа	82,27
14	T_1	0,001	Биграммы	Ближнего соседа	44,35
15	T_1	0,001	Биграммы	Средней связи	78,54
16	T_1	0,0001	Биграммы	Дальнего соседа	80,86
17	T_1	0,0001	Биграммы	Ближнего соседа	44,20
18	T_1	0,0001	Биграммы	Средней связи	76,27
19	T_1	0,05	100 слов	Дальнего соседа	76,62
20	T_1	0,05	100 слов	Ближнего соседа	53,43
21	T_1	0,05	100 слов	Средней связи	75,79
22	**T_1**	**0,001**	**100 слов**	**Дальнего соседа**	**81,85**
23	T_1	0,001	100 слов	Ближнего соседа	50,35
24	T_1	0,001	100 слов	Средней связи	80,84
25	T_1	0,0001	100 слов	Дальнего соседа	81,52
26	T_1	0,0001	100 слов	Ближнего соседа	51,57
27	T_1	0,0001	100 слов	Средней связи	80,93
28	T_1	0,05	500 слов	Дальнего соседа	82,92
29	T_1	0,05	500 слов	Ближнего соседа	52,49
30	T_1	0,05	500 слов	Средней связи	86,19
31	T_1	0,001	500 слов	Дальнего соседа	83,57
32	T_1	0,001	500 слов	Ближнего соседа	55,20
33	T_1	0,001	500 слов	Средней связи	83,00
34	**T_1**	**0,0001**	**500 слов**	**Дальнего соседа**	**87,94**
35	T_1	0,0001	500 слов	Ближнего соседа	56,20
36	T_1	0,0001	500 слов	Средней связи	85,87

Наилучший результат на множестве T_1 при кластеризации по служебным словам получен при α = 0,001 и методе дальнего соседа (83,42%). По биграммам наилучший результат получен при α = 0,05 и методе дальнего соседа (84%). По 100 частым словам – при α = 0,001 и методе дальнего соседа (81,85%). Самый лучший результат достигнут при кластеризации по 500 частым словам при α = 0,0001 и методе дальнего соседа (87,94%). Объединение кластеров по дальнему соседу практически во всех случаях (за исключением эксперимента № 30) дает лучшие по сравнению с другими методами результаты. Метод ближнего соседа работает всегда заметно хуже остальных двух методов. Уменьшение α с 0,05 до 0,001 почти всегда (за исключением биграмм) дает улучшение качества, дальнейшее уменьшение α до 0,0001, как правило (за исключением 500 частых слов), ухудшает результат.

Рассмотрим качество кластеризации для множества T_2 с меньшим числом текстов (табл. 2). Ввиду того что метод ближнего соседа, так же как и в предыдущем случае, всегда дает самые худшие результаты, в целях экономии места приведем цифры только по методу дальнего соседа и средней связи.

Таблица 2

Качество кластеризации текстов из множества T_2

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
37	T_2	0,05	Служебные слова	Дальнего соседа	76,74
38	**T_2**	**0,05**	**Служебные слова**	**Средней связи**	**82,71**
39	T_2	0,001	Служебные слова	Дальнего соседа	77,64
40	T_2	0,001	Служебные слова	Средней связи	72,67
41	T_2	0,0001	Служебные слова	Дальнего соседа	72,83
42	T_2	0,0001	Служебные слова	Средней связи	77,43

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
43	T_2	**0,05**	**Биграммы**	**Дальнего соседа**	**83,26**
44	T_2	0,05	Биграммы	Средней связи	78,56
45	T_2	0,001	Биграммы	Дальнего соседа	77,24
46	T_2	0,001	Биграммы	Средней связи	72,92
47	T_2	0,0001	Биграммы	Дальнего соседа	70,89
48	T_2	0,0001	Биграммы	Средней связи	64,54
49	T_2	0,05	100 слов	Дальнего соседа	75,66
50	T_2	0,05	100 слов	Средней связи	70,30
51	T_2	**0,001**	**100 слов**	**Дальнего соседа**	**80,15**
52	T_2	0,001	100 слов	Средней связи	78,06
53	T_2	0,0001	100 слов	Дальнего соседа	72,31
54	T_2	0,0001	100 слов	Средней связи	76,93
55	T_2	0,05	500 слов	Дальнего соседа	77,66
56	T_2	0,05	500 слов	Средней связи	75,13
57	T_2	**0,001**	**500 слов**	**Дальнего соседа**	**85,90**
58	T_2	0,001	500 слов	Средней связи	82,90
59	T_2	0,0001	500 слов	Дальнего соседа	81,74
60	T_2	0,0001	500 слов	Средней связи	78,58

Оценка качества по множеству T_2 показала, что наилучший результат опять же достигается при использовании в качестве признаков 500 самых часто встречающихся слов из частотного словаря Шарова (85,90%). Близкий результат дает использование биграмм (83,26%). Для трех из четырех наилучших результатов для каждой группы признаков лучшим методом объединения кластеров оказывается метод дальнего соседа. Уменьшение α с 0,05 до 0,001 для трех из четырех групп признаков на методе дальнего соседа дает улучшение качества кластеризации. Дальнейшее уменьшение α с 0,001 до 0,0001 для всех четырех групп признаков при использовании метода дальнего соседа ухудшает качество кластеризации.

Рассмотрим качество кластеризации для множества T_3 (табл. 3).

Таблица 3

Качество кластеризации текстов из множества T_3

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
61	T_3	0,05	Служебные слова	Дальнего соседа	82,25
62	T_3	0,05	Служебные слова	Средней связи	79,41
63	T_3	0,001	Служебные слова	Дальнего соседа	80,17
64	$\boldsymbol{T_3}$	**0,001**	**Служебные слова**	**Средней связи**	**90,43**
65	T_3	0,0001	Служебные слова	Дальнего соседа	86,53
66	T_3	0,0001	Служебные слова	Средней связи	80,69
67	$\boldsymbol{T_3}$	**0,05**	**Биграммы**	**Дальнего соседа**	**90,33**
68	T_3	0,05	Биграммы	Средней связи	87,26
69	T_3	0,001	Биграммы	Дальнего соседа	86,37
70	T_3	0,001	Биграммы	Средней связи	84,85
71	T_3	0,0001	Биграммы	Дальнего соседа	89,06
72	T_3	0,0001	Биграммы	Средней связи	83,36
73	T_3	0,05	100 слов	Дальнего соседа	81,23
74	T_3	0,05	100 слов	Средней связи	87,61
75	T_3	0,001	100 слов	Дальнего соседа	84,70
76	T_3	0,001	100 слов	Средней связи	88,42
77	$\boldsymbol{T_3}$	**0,0001**	**100 слов**	**Дальнего соседа**	**89,05**
78	T_3	0,0001	100 слов	Средней связи	86,59
79	T_3	0,05	500 слов	Дальнего соседа	87,49
80	$\boldsymbol{T_3}$	**0,05**	**500 слов**	**Средней связи**	**91,44**
81	T_3	0,001	500 слов	Дальнего соседа	88,50
82	T_3	0,001	500 слов	Средней связи	88,50
83	T_3	0,0001	500 слов	Дальнего соседа	91,34
84	T_3	0,0001	500 слов	Средней связи	88,50

Как и в предыдущих случаях, лучшие результаты кластеризации достигнуты с использованием 500 самых часто встречаемых слов (91,44%). Для трех из четырех групп признаков лучшие результаты были достигнуты с использованием метода дальнего соседа.

Использование 500 самых часто встречаемых слов вместо 100 во всех рассмотренных случаях (за исключением одного – эксперименты 20 и 29) при прочих равных параметрах дает улучшение качества кластеризации. В последующих экспериментах данная закономерность повторяется. Поэтому ниже приведем только результаты по трем группам признаков: служебные слова, биграммы, 500 самых часто встречаемых слов.

Для множества T_4, содержащего тексты 5 авторов, оценки качества кластеризации выглядят следующим образом (табл. 4).

Таблица 4

Качество кластеризации текстов из множества T_4

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
85	T_4	0,05	Служебные слова	Дальнего соседа	75,65
86	$\boldsymbol{T_4}$	**0,05**	**Служебные слова**	**Средней связи**	**79,44**
87	T_4	0,001	Служебные слова	Дальнего соседа	71,32
88	T_4	0,001	Служебные слова	Средней связи	74,41
89	T_4	0,0001	Служебные слова	Дальнего соседа	67,80
90	T_4	0,0001	Служебные слова	Средней связи	71,57
91	$\boldsymbol{T_4}$	**0,05**	**Биграммы**	**Дальнего соседа**	**86,15**
92	T_4	0,05	Биграммы	Средней связи	75,54
93	T_4	0,001	Биграммы	Дальнего соседа	74,98
94	T_4	0,001	Биграммы	Средней связи	53,48
95	T_4	0,0001	Биграммы	Дальнего соседа	74,98
96	T_4	0,0001	Биграммы	Средней связи	63,42
97	T_4	0,05	500 слов	Дальнего соседа	78,89
98	$\boldsymbol{T_4}$	**0,05**	**500 слов**	**Средней связи**	**82,19**
99	T_4	0,001	500 слов	Дальнего соседа	78,72
100	T_4	0,001	500 слов	Средней связи	74,66
101	T_4	0,0001	500 слов	Дальнего соседа	78,72
102	T_4	0,0001	500 слов	Средней связи	75,68

В отличие от предыдущих экспериментов, лучшие результаты при кластеризации множества T_4 дало использование биграмм. В двух из трех случаев самое высокое качество обеспечил метод средней связи. Для всех трех групп признаков лучшие результаты дало использование α = 0,05.

Рассмотрим результаты кластеризации T_5 (табл. 5).

Таблица 5

Качество кластеризации текстов из множества T_5

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
103	T_5	0,05	Служебные слова	Дальнего соседа	85,15
104	T_5	0,05	Служебные слова	Средней связи	94,56
105	T_5	0,001	Служебные слова	Дальнего соседа	97,13
106	$\boldsymbol{T_5}$	**0,001**	**Служебные слова**	**Средней связи**	**100,00**
107	T_5	0,0001	Служебные слова	Дальнего соседа	97,13
108	T_5	0,0001	Служебные слова	Средней связи	97,13
109	$\boldsymbol{T_5}$	**0,05**	**Биграммы**	**Дальнего соседа**	**93,68**
110	T_5	0,05	Биграммы	Средней связи	89,47
111	T_5	0,001	Биграммы	Дальнего соседа	89,47
112	T_5	0,001	Биграммы	Средней связи	92,71
113	T_5	0,0001	Биграммы	Дальнего соседа	93,68
114	T_5	0,0001	Биграммы	Средней связи	89,47
115	$\boldsymbol{T_5}$	**0,05**	**500 слов**	**Дальнего соседа**	**100,00**
116	T_5	0,05	500 слов	Средней связи	97,13
117	T_5	0,001	500 слов	Дальнего соседа	97,13
118	T_5	0,001	500 слов	Средней связи	97,13
119	T_5	0,0001	500 слов	Дальнего соседа	97,13
120	T_5	0,0001	500 слов	Средней связи	97,13

При кластеризации текстов из множества T_5 достигается качество в 100%. Такого результата удалось добиться как при использовании служебных слов, так и 500 самых часто встречающихся слов. Методы дальнего соседа и средней связи дали одинаковое число лучших результатов для разных наборов признаков. В трех из пяти лучших результатах использовался α = 0,05.

В заключение рассмотрим результаты кластеризации текстов из множества T_6 (табл. 6).

Таблица 6

Качество кластеризации текстов из множества T_6

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
121	T_6	0,05	Служебные слова	Дальнего соседа	97,39
122	T_6	0,05	Служебные слова	Средней связи	97,39
123	$\boldsymbol{T_6}$	**0,001**	**Служебные слова**	**Дальнего соседа**	**100,00**
124	T_6	0,001	Служебные слова	Средней связи	97,39
125	$\boldsymbol{T_6}$	**0,0001**	**Служебные слова**	**Дальнего соседа**	**100,00**
126	T_6	0,0001	Служебные слова	Средней связи	97,39

№ эксп.	Множество	α	Признаки	Метод объединения кластеров	Качество
127	$\boldsymbol{T_6}$	**0,05**	**Биграммы**	**Дальнего соседа**	**100,00**
128	T_6	0,05	Биграммы	Средней связи	94,98
129	T_6	0,001	Биграммы	Дальнего соседа	94,98
130	T_6	0,001	Биграммы	Средней связи	91,06
131	T_6	0,0001	Биграммы	Дальнего соседа	92,74
132	T_6	0,0001	Биграммы	Средней связи	88,04
133	$\boldsymbol{T_6}$	**0,05**	**500 слов**	**Дальнего соседа**	**100,00**
134	$\boldsymbol{T_6}$	**0,05**	**500 слов**	**Средней связи**	**100,00**
135	$\boldsymbol{T_6}$	**0,001**	**500 слов**	**Дальнего соседа**	**100,00**
136	$\boldsymbol{T_6}$	**0,001**	**500 слов**	**Средней связи**	**100,00**
137	$\boldsymbol{T_6}$	**0,0001**	**500 слов**	**Дальнего соседа**	**100,00**
138	$\boldsymbol{T_6}$	**0,0001**	**500 слов**	**Средней связи**	**100,00**

На множестве T_6 качество 100% достигается в 10 экспериментах. Причем при использовании частот 500 самых часто встречающихся слов стопроцентное качество обеспечивается при любом рассмотренном наборе параметров. Интересно заметить, что при кластеризации множества T_4, состоящего из текстов русских классиков XIX в., наилучший результат составляет 86,15%, в то время как для множества T_5, включающего произведения авторов начала XX в., качество в некоторых случаях уже достигает 100%, и наконец при кластеризации множества T_6, содержащего произведения современников, получение стопроцентного качества наименее зависит от значения параметров метода. Так как возможность хорошей кластеризации зависит от стабильности признаков стиля среди текстов одного автора, из полученных данных можно сделать предварительный вывод о том, что чем современней авторы текстов в наборе, тем больше шансов получить хорошую кластеризацию текстов.

Выводы

На основе проведенных экспериментов можно сделать следующие выводы:

1. Использование метода ближнего соседа для объединения кластеров текстов всегда уступает использованию метода дальнего соседа и метода средней связи.

2. Использование метода дальнего соседа заметно чаще дает результаты лучше, чем использование метода средней связи. Среди 31 лучшего результата 21 получен с использованием метода дальнего соседа.

3. При прочих равных параметрах наилучшие результаты кластеризации дает использование частот 500 самых часто встречаемых слов (5 из 6 рассмотренных множеств текстов).

4. Использование 100 самых часто встречаемых слов, независимо от параметров, дает более низкое качество кластеризации, чем использование 500 самых часто встречаемых слов.

5. Однозначной зависимости качества от уровня значимости гипергеометрического критерия не выявлено. Однако при использовании биграмм наилучшие результаты дает α = 0,05, при использовании служебных слов α = 0,05 или α = 0,001.

В итоге можно сказать, что гарантированно хороший и вероятно наиболее лучший результат можно получить при использовании в качестве признаков 500 самых часто встречаемых слов, в качестве метода объединения кластеров – метод дальнего соседа, в качестве уровня значимости критерия α = 0,01. Тем не менее, если имеется возможность проведения экспериментов с другими параметрами, имеет смысл проверить также метод средней связи и уровни значимости α = 0,05, α = 0,0001.

ЛИТЕРАТУРА

1. *Поддубный В.В., Шевелев О.Г.* Сравнение и кластерный анализ текстов по частотным признакам на основе гипергеометрического критерия // Квантитативная лингвистика: исследования и модели (КЛИМ – 2005): Материалы Всероссийской научной конференции. Новосибирск: Изд-во НГПУ, 2005. С. 205–217.
2. *Steinbach M., Karypis G., Kumar V.* A comparison of document clustering techniques // Proceeding of KDD Workshop on Text Mining. Boston, MA, August 2000.
3. *Фоменко В.П., Фоменко Т.Г.* Авторский инвариант русских литературных текстов / Предисловие А.Т. Фоменко // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2. М.: Изд-во МГУ, 1996. С. 768–820.
4. *Шаров С.А.* Частотный словарь [Электронный ресурс]. – Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный.

Статья представлена кафедрой прикладной информатики факультета информатики Томского государственного университета, поступила в научную редакцию «Кибернетика» 30 мая 2006 г.