ИНФОРМАТИКА

2008 **апрель-июнь** **№ 2**

АНАЛИЗ И ЗАЩИТА ДАННЫХ

УДК 519.237.8-37

М.М. Невдах, М.А. Зильберглейт

СИСТЕМАТИЗАЦИЯ ИНФОРМАЦИОННЫХ ХАРАКТЕРИСТИК УЧЕБНОГО ТЕКСТА С ИСПОЛЬЗОВАНИЕМ МЕТОДА КЛАСТЕРНОГО АНАЛИЗА

Проводится кластеризация 49 информационных характеристик учебного текста по экономической теории с использованием пакета SPSS. Для анализа данных в качестве критерия близости групп используются такие меры сходства, как расстояние Евклида, квадрат расстояния Евклида, косинус угла, коэффициент корреляции, неравенство Чебышева, расстояние Минковского, манхэттенское расстояние. В результате кластеризации все признаки разбиваются на девять условных групп. При дальнейшей обработке достаточно использовать один признак из каждой группы.

Введение

В редакционно-издательской практике оценка трудности текста для определенной категории читателей производится редактором на основе детального анализа рукописи. Очевидно, что данная оценка, от которой зависит качество подготовки издания, основана на квалификации редактора и его профессиональном опыте. Развитие идей кибернетики, в частности научной дисциплины «Распознавание образов», позволяет поставить вопрос о внедрении в редакторскую подготовку изданий автоматизированных систем, выполняющих информационные, логические, аналитические и другие задачи, решение которых до сих пор связывают иногда с деятельностью живого мозга. Полная или частичная замена человека (корректора, редактора) сложной специализированной системой позволяет добиться не только невозможного для человека быстродействия, но и необходимого качества изданий благодаря объективной оценке трудности текста на основе его информационных характеристик.

Отметим, что в процессе анализа рукописи редактор, опираясь на множество факторов, принимает ограниченное число решений: а) рукопись пригодна к публикации; б) рукопись не пригодна к публикации; в) рукопись пригодна после некоторой доработки. Таким образом, автоматизация ряда процессов предполагает конструирование автоматических устройств, способных реагировать на изменяющиеся характеристики различных объектов (текстов разных жанров) определенным количеством удовлетворительных для человека реакций.

Как известно, распознавание образов (объектов, сигналов, ситуаций, явлений или процессов) – задача идентификации объекта или определения каких-либо свойств по его изображению или аудиозаписи и другим характеристикам. В нашем случае стоит задача определения понятности (читабельности) текста для будущих читателей, решение которой включает несколько этапов:

1) нахождение и реализация методов для определения трудности понимания различных текстов данной группой лиц;

2) выбор формальных характеристик текста (и только тех, которые поддаются точному измерению);

3) создание автоматизированной системы, которая бы на основе ответов испытуемых, полученных экспериментальным путем, предсказывала понятность текста для будущих читателей.

Таким образом, базой для решения задачи отнесения объектов к тому или иному классу и построения конкретного алгоритма являются не только измерения параметров, характеризующих определенный объект, но и ответы (реакции) испытуемых, по которым можно судить о степени трудности текста.

Исходя из предложенной процедуры на первом этапе были проведены эксперименты с использованием методики дополнения, экспертных оценок трудности текста и метода парных сравнений, обработка и анализ результатов которых позволили выявить необходимую инфор-

мацию относительно трудности восприятия текста. В качестве экспериментального материала использовались учебные тексты по экономической теории для высшей школы [1–4]. Объем выборки составил 1800–2000 печатных знаков. Выбор данной величины обусловлен тем, что, начиная с объема в 1800 печатных знаков, статистические характеристики текста становятся относительно постоянными [5].

Следующим этапом исследования является изучение информационных характеристик учебного текста. Проблемам, связанным с исследованием влияния информационных характеристик текста на его читабельность, посвящен ряд работ [6–9], в которых основное внимание уделяется небольшому числу параметров текста, включающих обычно среднюю длину предложения в словах, предложения в слогах, предложения в буквах, предложения в печатных знаках, слов в слогах, слов в буквах, слов в печатных знаках, а также процент числа конкретных/абстрактных существительных, повторяющихся существительных, прилагательных, сложных и простых предложений, слов, превышающих определенную длину, глаголов и др. Только в отдельных работах [10, 11] число исследуемых характеристик текста, влияющих на его читабельность, превышает 100 признаков. Систематических исследований, посвященных изучению влияния значительного числа параметров текста на его усвоение, до настоящего времени не проводилось.

Текст можно представить как объект, характеризующийся многомерным вектором, состоящим из различного рода переменных. В связи с этим он может быть исследован с помощью методов многомерного статистического анализа.

В настоящей работе изучено 49 признаков учебных текстов по экономической теории. Очевидно, что использование такого большого числа характеристик для практических целей невозможно. В первую очередь это связано с тем, что данные параметры могут быть сильно коррелированы. С другой стороны, ничем не оправданное уменьшение числа переменных может привести к потере точности экспериментов. Таким образом, цель данной работы – кластеризация выбранных 49 признаков текста.

Кластеризация информационных характеристик текста

Для анализа данных и проведения статистического анализа был использован пакет SPSS.

Кластерный анализ представляет собой «многомерную статистическую процедуру, выполняющую сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающую объекты в сравнительно однородные группы» [12].

Для анализа были выделены следующие 49 параметров текста:

1) длина текста в абзацах;
2) длина текста в словах;
3) длина текста в буквах;
4) средняя длина абзаца в фразах;
5) средняя длина абзаца в словах;
6) средняя длина абзаца в буквах;
7) средняя длина абзаца в печатных знаках;
8) средняя длина предложения в фразах;
9) средняя длина предложения в словах;
10) средняя длина предложения в слогах;
11) средняя длина предложения в буквах;
12) средняя длина предложения в печатных знаках;
13) средняя длина самостоятельного предложения в фразах;
14) средняя длина самостоятельного предложения в словах;
15) средняя длина самостоятельного предложения в слогах;
16) средняя длина самостоятельного предложения в буквах;
17) средняя длина самостоятельного предложения в печатных знаках;
18) средняя длина фразы в словах;
19) средняя длина фразы в слогах;
20) средняя длина фразы в буквах;

21) средняя длина фразы в печатных знаках;
22) средняя длина слов в слогах;
23) средняя длина слов в буквах;
24) средняя длина слов в печатных знаках;
25) средняя длина слов по Деверу;
26) процент слов длиной в 5 букв и больше;
27) процент слов длиной в 6 букв и больше;
28) процент слов длиной в 7 букв и больше;
29) процент слов длиной в 8 букв и больше;
30) процент слов длиной в 9 букв и больше;
31) процент слов длиной в 10 букв и больше;
32) процент слов длиной в 11 букв и больше;
33) процент слов длиной в 12 букв и больше;
34) процент слов длиной в 13 букв и больше;
35) процент слов в 3 слога и больше;
36) процент слов в 4 слога и больше;
37) процент слов в 5 слогов и больше;
38) процент слов в 6 слогов и больше;
39) процент неповторяющихся слов;
40) средняя частота повторения слова;
41) процент неповторяющихся существительных;
42) процент повторяющихся существительных;
43) процент конкретных существительных;
44) процент абстрактных существительных;
45) процент прилагательных;
46) процент глаголов;
47) процент сложных предложений;
48) процент простых предложений;
49) процент придаточных предложений среди фраз.

Следует сделать несколько уточнений. Под термином «фраза» в данной статье понимается отрезок текста, в котором содержится одна предикативная связь. Исходя из этого к фразе относятся простое предложение, части сложносочиненного предложения, главное и придаточное предложения в сложноподчиненном. Самостоятельным предложением считаются простые предложения, части сложносочиненного предложения и сложноподчиненное в целом. Впервые такую единицу текста использовал Р. Флеш в [7]. Средняя длина слов по Деверу рассчитывалась делением общего количества знаков с пробелами на число знаков без пробелов.

Так как характеристики текста измерялись в различных единицах (табл. 1), то все данные были стандартизированы. Для этого использовалась нормализация, приводящая все переменные к стандартной *z*-шкале (табл. 2).

Таблица 1

Значения признаков исследуемых текстов

Номер признака	Номер теста															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	6	9	5	7	4	6	4	5	6	5	2	4	4	4	6	6
2	215	229	263	218	220	200	251	250	204	258	251	218	260	225	238	199
3	1473	1572	1710	1552	1810	1551	1707	1589	1570	1623	1766	1711	1699	1596	1654	1454
4	3,5	3	5	2,9	4,8	2,8	5,3	5,8	3	2,6	9	4,8	6,5	3,5	3,2	3
5	35,8	25,4	52,6	31,1	55	33,3	62,8	50	34	51,6	125,5	54,5	65	56,3	39,7	33,2
6	245,5	174,7	342	221,7	452,5	258,5	426,8	317,8	261,7	324,6	883	427,8	424,8	399	275,7	242,3
7	289,7	204,2	406,2	260,4	565,8	297,8	501,3	379	301,5	385,8	1037,5	490,5	505,3	466,8	326	285,7
8	1,91	1,5	1,39	1,43	1,27	1,21	1,5	1,61	1,2	1	1,1	1,1	1,2	1,1	1,1	1
9	19,5	12,7	14,6	15,6	14,7	14,3	17,9	13,9	13,6	19,8	17,9	12,8	12,3	20,2	12,9	12,6
10	55,4	36,9	41,2	54,7	50,7	50,6	60,9	36,8	45,8	54,3	55,4	42,7	34,5	61,2	36,7	40,3
11	133,9	87,3	95	110,9	120,7	110,8	121,9	88,3	104,7	125,1	110,4	100,6	77,2	122,8	97,3	80,8

Продолжение табл. 1

Номер признака	Номер теста															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
12	158	102,1	112,8	130,2	150,9	127,6	143,2	105,3	120,6	148,4	129,7	115,4	91,9	143,6	115,1	95,2
13	1,3	1,4	1,3	1,2	1,3	1,3	1,4	1,5	1	1	1	1,1	1,04	1,1	1,1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
49	23,81	18,52	20	15	5,26	11,76	19,05	17,24	5,6	69,2	33,3	26,3	30,8	28,6	15,8	11,1

Таблица 2

Стандартизированные переменные исследуемых текстов

Номер признака	Номер теста															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-0,002	-0,001	-0,001	-0,001	-0,001	-0,002	-0,001	-0,001	-0,002	-0,002	-0,001	-0,001	-0,001	-0,002	-0,001	-0,002
2	0,003	0,003	0,003	0,003	0,002	0,002	0,002	0,003	0,002	0,003	0,001	0,002	0,003	0,002	0,003	0,003
3	0,031	0,030	0,026	0,030	0,023	0,029	0,025	0,028	0,029	0,027	0,018	0,025	0,025	0,027	0,028	0,031
4	-0,002	-0,001	-0,001	-0,002	-0,001	-0,002	-0,001	-0,001	-0,002	-0,002	-0,001	-0,001	-0,001	-0,002	-0,001	-0,002
5	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,000	-0,001	-0,001	-0,001	0,000	-0,001	0,000	-0,001	-0,001	-0,001
6	0,004	0,002	0,004	0,003	0,005	0,004	0,005	0,004	0,004	0,004	0,008	0,005	0,005	0,005	0,003	0,004
7	0,005	0,003	0,005	0,004	0,006	0,004	0,006	0,006	0,004	0,005	0,010	0,006	0,006	0,007	0,004	0,005
8	-0,002	-0,001	-0,001	-0,002	-0,001	-0,002	-0,001	-0,001	-0,002	-0,002	-0,001	-0,001	-0,001	-0,002	-0,001	-0,002
9	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001
10	-0,001	-0,001	-0,001	0,000	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001
11	0,001	0,000	0,000	0,001	0,000	0,001	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,001	0,000	0,000
12	0,002	0,001	0,000	0,001	0,001	0,001	0,001	0,000	0,001	0,001	0,000	0,000	0,000	0,001	0,001	0,000
13	-0,002	-0,001	-0,001	-0,002	-0,001	-0,002	-0,001	-0,001	-0,002	-0,002	-0,001	-0,001	-0,001	-0,002	-0,001	-0,002
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
49	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	-0,002	0,000	-0,001	-0,001	-0,001	-0,001	-0,001	-0,002

В большинстве алгоритмов многомерной классификации используется понятие, которое носит название «мера сходства» или «мера подобия» между объектами. Существуют следующие коэффициенты сходства [12]: коэффициент корреляции, меры расстояния, коэффициенты ассоциативности, вероятностные коэффициенты сходства.

Для анализа данных в качестве критерия для определения подобия групп использовались следующие меры сходства:

– расстояние Евклида;
– квадрат расстояния Евклида;
– косинус угла;
– коэффициент корреляции;
– неравенство Чебышева;
– расстояние Минковского;
– манхэттенское расстояние.

Для кластеризации приведенных информационных характеристик текста использовались следующие основные алгоритмы метода кластерного анализа: межгрупповое связывание, внутригрупповое связывание, одиночное связывание, полное связывание, центроидная кластеризация, центральное связывание, метод Варда. Количество кластеров по каждому алгоритму варьировалось от 3 до 10 (табл. 3). После выбора всех соответствующих параметров была получена необходимая информация по формированию кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик текста к тому или иному кластеру.

Выводимые результаты для наглядности можно представить и в виде дендрограммы, которая позволяет не только перейти к любому признаку на любом уровне кластеризации, но и дает возможность судить о том, каково расстояние между кластерами или признаками на каждом из уровней (рисунок).

Таблица 3

Кластеризация на примере использования алгоритма «Метод Варда», основанного на расстоянии Евклида

Номер признака	Количество кластеров							
	10	9	8	7	6	5	4	3
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	1
3	3	3	3	3	3	3	3	2
4	1	1	1	1	1	1	1	1
5	4	4	4	4	4	4	2	1
6	5	5	5	5	5	5	4	3
7	6	6	6	5	5	5	4	3
8	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1
10	4	4	4	4	4	4	2	1
11	7	7	7	6	6	2	2	1
12	7	7	7	6	6	2	2	1
13	1	1	1	1	1	1	1	1
…	…	…		…	…	…	…	…
49	10	9	4	4	4	4	2	1

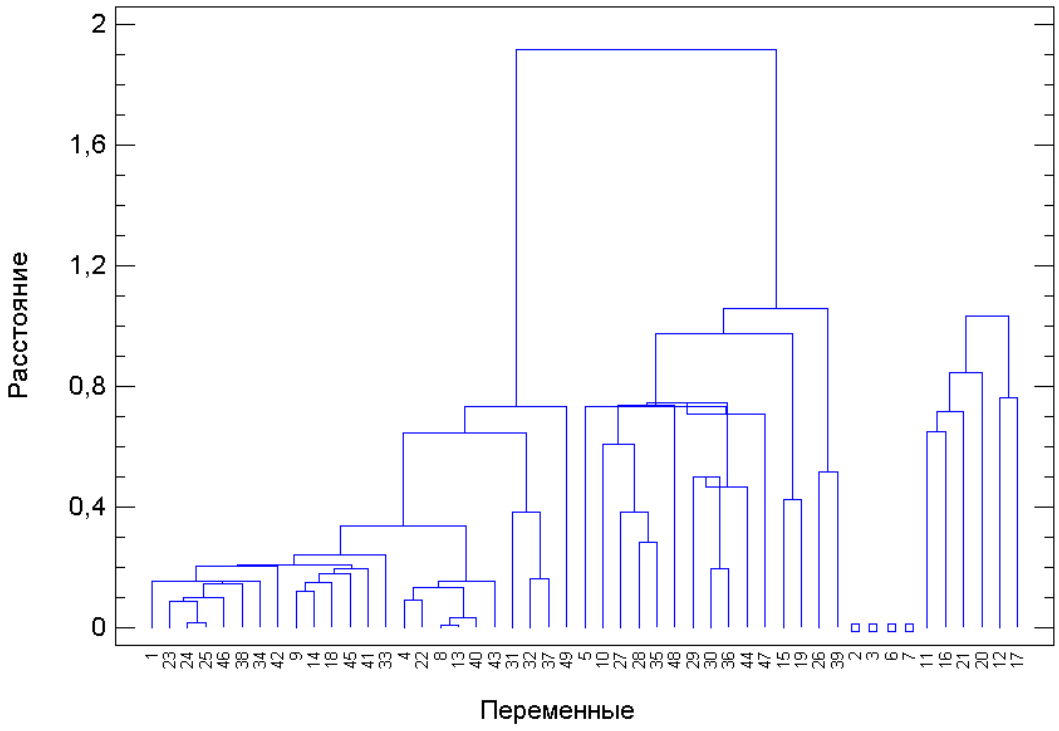

Рис. Дендрограмма по центроидному методу на основе манхэттенского расстояния для шести кластеров

В результате анализа данных о влиянии исследуемых характеристик текста с использованием всех известных алгоритмов и мер сходства были получены 392 дендрограммы, которые отражают кластеризацию переменных в условные группы.

Известно, что на результаты кластерного анализа сильное влияние оказывает как используемая мера сходства, так и алгоритм кластеризации. Поэтому в применении процедур кластерного анализа немаловажным аспектом является устойчивость структуры кластеров, отражающая реальную объективность классификации. Одним из наиболее простых и эффективных способов проверки устойчивости результатов является метод сравнения результатов, полученных для различных алгоритмов кластеризации, который и использовался в данной работе. Для

этого все данные для наглядности были объединены в семь сводных таблиц. В них четко прослеживаются особенности применения различных алгоритмов кластерного анализа, использующих разные меры сходства. Результаты формирования кластеров согласуются практически по всем алгоритмам. Незначительно отличаются данные по методу Варда. Сравнение результатов с применением различных мер сходства показало, что наблюдаются заметные различия лишь в данных, полученных методами измерения близости, которые основаны на корреляции векторов значений и манхэттенском расстоянии.

Проведенный анализ показал, что целесообразно выделить следующие девять условных групп:

1. Признаки 1, 4, 8, 13, 22–25, 40, 42–44, 46 – длина текста в абзацах, средняя длина абзаца в фразах, средняя длина предложения в фразах, средняя длина самостоятельного предложения в фразах, средняя длина слов в слогах, средняя длина слов в буквах, средняя длина слов в печатных знаках, средняя длина слов по Деверу, средняя частота повторения слова, процент повторяющихся существительных, процент конкретных существительных, процент абстрактных существительных, процент глаголов.

2. Признаки 2, 9, 14, 18 – длина текста в словах, средняя длина предложения в словах, средняя длина самостоятельного предложения в словах, средняя длина фразы в словах.

3. Признаки 3, 39, 45, 48 – длина текста в буквах, процент неповторяющихся слов, процент прилагательных и процент простых предложений.

4. Признаки 5–7 – средняя длина абзаца в словах, средняя длина абзаца в буквах, средняя длина абзаца в печатных знаках.

5. Признаки 10–12, 16, 17 – средняя длина предложения в слогах, средняя длина предложения в буквах, средняя длина предложения в печатных знаках, средняя длина самостоятельного предложения в буквах, средняя длина самостоятельного предложения в печатных знаках.

6. Признаки 15, 19–21 – средняя длина самостоятельного предложения в слогах, средняя длина фразы в слогах, средняя длина фразы в буквах, средняя длина фразы в печатных знаках.

7. Признаки 26–30, 35–37 – процент слов длиной в 5 букв и больше, процент слов длиной в 6 букв и больше, процент слов длиной в 7 букв и больше, процент слов длиной в 8 букв и больше, процент слов длиной в 9 букв и больше, процент слов в 3 слога и больше, процент слов в 4 слога и больше, процент слов в 5 слогов и больше.

8. Признаки 31–34, 38, 41 – процент слов длиной в 10 букв и больше, процент слов длиной в 11 букв и больше, процент слов длиной в 12 букв и больше, процент слов длиной в 13 букв и больше, процент слов в 6 слогов и больше, процент неповторяющихся существительных.

9. Признаки 47, 49 – процент сложных предложений, процент придаточных предложений среди фраз.

Следует отметить, что впервые в исследованиях по читабельности с использованием метода кластерного анализа снижена размерность признакового пространства учебного текста. Анализ кластеризации показал, что выделенные группы связаны, прежде всего, с такими параметрами, как длина слов и предложений. Многие исследования показали, что более длинные слова являются более информативными, хотя и менее знакомыми [13, с. 203; 14]. Поэтому длина слова является хорошим показателем сложности текста, который входит в большинство формул читабельности. Длина предложения является также одним из наиболее часто встречающихся факторов трудности текста. Это можно объяснить тем, что сущность понимания заключается в осознании связей между словами и отражаемыми ими предметами и явлениями действительности. Более длинное предложение предполагает осознание большего количества связей. Кроме того, для образования связи между словами человек использует кратковременную память, которая имеет предел [15]. Это накладывает ограничение на размер предложений. Слишком длинные предложения понимаются читателями хуже, чем короткие.

При использовании кластерного анализа важной является оценка качества кластеризации. Известно много алгоритмов оптимизации кластерных решений. Например, в [16] автором дан обзор 45 функционалов качества. Все это свидетельствует об отсутствии универсального критерия оптимизации кластеризации. Поэтому целесообразной является проверка согласованности решения с выводами, сделанными с помощью других методов многомерной статистики. Сравнение условных групп, выделенных с помощью факторного анализа и метода корреляци-

онных плеяд, показало, что наиболее устойчивыми, а следовательно, тесно связанными являются следующие признаки: 5–7; 11, 12; 16, 17; 15, 19–21; 26–30; 31–34; 35–37.

Сравнение результатов для учебного текста по философии и экономической теории, полученных с помощью различных методов многомерного статистического анализа, позволяет сделать следующий вывод: во многих случаях совпадают не только отдельные признаки в группах (например, признаки 6, 7; 10–12; 16, 17 и др.), но и сами группы (например, группа признаков 5–7; 15, 19–21 и др.). Из этого следует, что характеристики учебного текста для высшей школы по различным отраслям знаний целесообразно изучать в рамках единого информационного поля.

Заключение

Проведенный анализ позволил определить достаточное количество кластеров для дальнейшего исследования данных о влиянии параметров текста на его читабельность. Научная и практическая значимость полученных результатов заключается в том, что, во-первых, из дальнейшей обработки данных будут исключены характеристики текста, которые сильно коррелированы между собой; во-вторых, уменьшение числа признаков обосновано использованием многомерного статистического анализа. Эти факторы должны повлиять на точность конечного результата. Таким образом, для последующей обработки достаточно пользоваться одним признаком из каждой группы, например средней частотой повторения слова, средней длиной предложения в словах, процентом неповторяющихся слов, средней длиной абзаца в словах, средней длиной предложения в слогах, средней длиной самостоятельного предложения в слогах, процентом слов длиной в пять букв и больше, процентом слов в шесть слогов и больше и процентом сложных предложений. В будущем полученные данные будут использованы для построения решающего правила, т. е. методики отнесения объекта к какому-либо классу.

Список литературы

1. Экономическая теория: учебное пособие / Л. Н. Давыденко [и др.]. – Минск: Вышэйшая школа, 2002. – 366 с.

2. Экономическая теория: учебник / Н.И. Базылев [и др.]; под общ. ред. Н.И. Базылева, С.П. Гурко. – Минск: Экоперспектива, 1997. – 368 с.

3. Экономическая теория: учебник для студентов вузов / Под ред. В. Д. Камаева. – М.: ВЛАДОС, 2001. – 640 с.

4. Сажина, М.А. Основы экономической теории: учебное пособие для неэкономических специальностей вузов / М.А. Сажина, Г.Г. Чибриков. – М.: Экономика, 1995. – 367 с.

5. Косова, М.М. Описательная статистика учебных текстов по физике / М.М. Косова, М.А. Зильберглейт // Труды БГТУ. Сер. VI. Физ.-мат. науки и информатика. – 2006. – Вып. XIV. – С. 167–170.

6. Chall, J.S. Readability: an appraisal of research and application / J.S. Chall // Bureau of education research monographs. – Columbus: Ohio State University Press, 1958. – № 34. – P. 58–68.

7. Flesch, R. The art of readable writing / R. Flesch. – New York: Harper, 1949. – 149 p.

8. Fry, E.B. The readability graph validated at primary levels / E.B. Fry // The reading teacher. – 1969. – № 22. – P. 534–538.

9. Paul, T. Guided Independent Reading / T. Paul. – Madison: School Renaissance Institute, 2003. – 74 p.

10. Микк, Я.А. Методика разработки формул читабельности / Я.А. Микк // Советская педагогика и школа. – Тарту, 1974. – Вып. 9. – С. 78–163.

11. Gray, W.S. What makes a book readable / W.S. Gray. – Chicago: Chicago University Press, 1935.

12. Дубнов, П.Ю. Обработка статистической информации с помощью SPSS / П.Ю. Дубнов. – М.: АСТ, 2004. – 221 с.

13. Пиотровский, Р.Г. Текст, машина, человек / Р.Г. Пиотровский. – Л.: Наука, 1975. – 327 с.

14. Flesch, R. The Art of Plain Talk / R. Flesch. – New-York: Harper and Brothers Publishers, 1946. – 210 p.

15. Миллер, Дж. Магическое число плюс или минус два. О некоторых пределах нашей способности перерабатывать информацию / Дж. Миллер // Инженерная психология. – М., 1964. – С. 192–225.

16. Мандель, И.Д. Кластерный анализ / И.Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.

Поступила 26.02.08

Белорусский государственный
технологический университет,
Минск, ул. Свердлова, 13а
e-mail: mazi@mail.ru,
nevdax@tut.by

M.M. Neudakh, M.A. Zilbergleit

SYSTEMATIZATION OF INFORMATION CHARACTERISTICS OF EDUCATIONAL TEXTS USING CLUSTER ANALYSIS

Clustering of 49 text information characteristics using a statistical package SPSS is described. The following group similarity measures were used: Euclidean distance, squared Euclidean distance, cosine of angle, correlation coefficient, Chebychev distance, city block distance, Minkowski distance. As a result of clusterization all attributes were partitioned into nine conditional groups. It is sufficient to use one attribute from each group for further processing.