ИССЛЕДОВАНИЕ ФОРМАЛЬНЫХ ХАРАКТЕРИСТИК УЧЕБНЫХ ТЕКСТОВ МЕТОДАМИ ФАКТОРНОГО АНАЛИЗА

Ю. Ф. Шпаковский, М. М. Невдах

Белорусский государственный технологический университет
Минск, Республика Беларусь
E-mail: yury_s@tut.b, nevdax@tut.by

В статье методами факторного анализа (главных факторов и главных компонент, центроидным методом) проведена систематизация 49-ти информационных характеристик текста, влияющих на усвоение учебного материала. Обработка полученных результатов позволила выделить семь условных групп близких параметров текста. Полученные расчеты будут использованы для построения решающего правила разбиения.

Ключевые слова: факторный анализ, характеристики текста, систематизация, корреляционная матрица, существенные факторы.

В редакционно-издательской практике оценка трудности текста для определенной категории читателей выносится 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 информационных технологий 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.

Одним из этапов создания автоматической системы является измерение параметров, характеризующих определенный объект. Текст можно представить как объект, характеризующийся многомерным вектором, состоящим из различного рода переменных.

В качестве экспериментального материала использовались 32 отрывка из учебных изданий по философии и экономической теории для вузов [1—8]. Объем одной выборки составил 1800—2000 печатных знаков.

В качестве переменных было выбрано 49 признаков 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.

Использование большого количества параметров текста является неэффективным по ряду причин [10, с. 516]: а) сильная взаимосвязанность признаков, что приводит к дублированию информации; б) неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); в) возможность агрегирования (простого или «взвешенного» суммирования) по некоторым признакам.

В связи с этим представляется целесообразным с помощью методов многомерного статистического анализа перейти от p исходных показателей анализируемого материала к существенно меньшему числу наиболее информативных переменных (p').

В проведенных ранее исследованиях была проведена группировка признаков с использованием кластерного анализа, методов корреляционных плеяд и вроцлавской таксономии, многомерного шкалирования. Цель данной работы — исследование параметров текста методами факторного анализа.

Снижение размерности набора переменных в методах факторного анализа базируется в основном на взаимной коррелированности исходных признаков. В связи с этим первый этап исследования заключался в вычислении корреляционной матрицы. фрагмент которой представлен в табл. 1.

Таблица 1

Корреляционная матрица исходных признаков

№ п/п	1	2	3	4	5	6	7	8	9	...	49
1	1,00	-0,31	0,05	-0,79	-0,86	-0,88	-0,83	-0,34	-0,47	...	-0,52
2	-0,31	1,00	0,40	0,50	0,53	0,41	0,43	0,46	0,50	...	0,42
3	0,05	0,40	1,00	-0,14	0,02	0,16	0,28	-0,20	0,04	...	-0,10
4	-0,79	0,50	-0,14	1,00	0,90	0,82	0,76	0,68	0,51	...	0,40
5	-0,86	0,53	0,02	0,90	1,00	0,96	0,93	0,54	0,65	...	0,54
6	-0,88	0,41	0,16	0,82	0,96	1,00	0,98	0,38	0,57	...	0,47
...	...	...	...	...	...	...	...	...	...	...	...
49	-0,52	0,42	-0,10	0,40	0,54	0,47	0,40	0,38	0,60	...	1,00

При изучении экспериментальных данных было установлено, что первые три фактора объясняют около 64% разброса дисперсии (табл. 2.).

Так как факторный анализ является методом сокращения числа переменных, то возникает вопрос, какие из факторов следует оставить для дальнейшей обработки. Однако установить заранее назначение каждого фактора не всегда представляется

возможным, поэтому для начала были использованы формальные критерии: критерий Кайзера [11] и критерий «каменистой осыпи» Р. Кэтелла [12].

Таблица 2

Объясненная дисперсия исследуемых параметров текста

Метод	Фактор	Исходные собственные значения		
		Собственные значения	Процент дисперсии	Кумулятивный процент
Метод главных факторов	1	17,75060	36,22572	36,22572
	2	9,81164	20,02376	56,24948
	3	4,02424	8,21273	64,46222
	4	3,06513	6,25537	70,71759
Центроидный метод	1	17,44565	35,60337	35,60337
	2	9,63674	19,66681	55,27019
	3	4,23599	8,64487	63,91505
	4	3,17027	6,46995	70,38500
Метод главных компонент	1	17,92808	36,58793	36,58793
	2	9,99074	20,38927	56,97720
	3	4,29961	8,77471	65,75191
	4	3,28236	6,69869	72,45060

Так как факторный анализ является методом сокращения числа переменных, то возникает вопрос, какие из факторов следует оставить для дальнейшей обработки. Исследователи рекомендуют руководствоваться здравым смыслом и оставлять только те факторы, которые имеют понятную или логическую интерпретацию. Однако установить заранее назначение каждого фактора не всегда представляется возможным, поэтому для начала были использованы формальные критерии: критерий Кайзера [11] и критерий «каменистой осыпи» Р. Кэтелла [12].

На основании первого критерия, предложенного Кайзером в 1960 году, для дальнейшего анализа необходимо сохранить те факторы, чьи собственные значения превышают единицу. В данном случае следует оставить восемь факторов для всех методов факторного анализа. Критерий «каменистой осыпи» является графическим методом. Для выделения факторов используется график их собственных значений (рис. 1).

По утверждению Р. Кэттела следует найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Анализ графиков для всех методов показал, что целесообразно оставить от 3—4 фактора.

Следует отметить, что первый критерий, как правило, сохраняет слишком много факторов, в то время как второй — слишком мало, поэтому решение об оптимальном количестве факторов можно принять только после их вращения и интерпретации.

Целью вращения факторов является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам. Нагрузка (значение лежит в пределах от –1 до 1) отражает

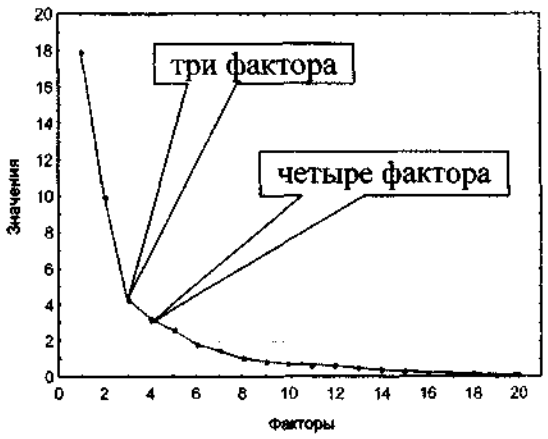

Рис. 1. – График собственных значений для метода главных компонент

связь между переменной и фактором. В работе использовались ортогональные методы вращения: варимакс, квартимакс и эквимакс. В результате были получены матрицы нагрузок для переменных. Фрагмент представлен в табл. 3.

Таблица 3

Факторные веса при анализе 49-ти информационных характеристик текста с использованием метода главных компонент и вращением эквимакс

Параметры текста	Фактор 1	Фактор 2	Фактор 3
1. Длина текста в абзацах	0,16201	-0,58391	0,318794
2. Длина текста в словах	-0,62186	0,41246	0,337997
3. Длина текста в буквах	0,36217	0,28747	0,372348
4. Средняя длина абзаца в фразах	-0,53533	0,45706	-0,374788
5. Средняя длина абзаца в словах	-0,35950	0,69265	-0,200807
6. Средняя длина абзаца в буквах	-0,12685	0,71192	-0,226083
7. Средняя длина абзаца в печатных знаках	-0,07227	0,69957	-0,141158
8. Средняя длина предложения в фразах	-0,59543	0,51458	-0,277823
9. Средняя длина предложения в словах	-0,29474	0,87561	0,056910
10. Средняя длина предложения в слогах	0,04078	0,89114	0,016074
...	...	...	...
49. Процент придаточных предложений среди фраз	-0,31478	0,58613	0,105082

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между четырьмя факторами (табл. 4).

Таблица 4

Распределение характеристик текста с использованием различных методов факторного анализа и методов вращения

Метод вращения	Метод факторного анализа								
	метод главных факторов			центроидный метод			метод главных компонент		
	фактор 1	фактор 2	фактор 3	фактор 1	фактор 2	фактор 3	фактор 1	фактор 2	фактор 3
варимакс	22, 23, 25–38	5, 6, 9–12, 14, 16	15, 18, 19, 21	22, 23, 25–38	4–7, 9–12, 14	15, 18–21	22, 23, 25–38	5–7, 9–12, 14, 16	15, 19
квартимакс	22, 23, 25–38	9–12, 14, 16	15, 19	22, 23, 25–38	5–7, 9–12, 14	15, 18–21	20, 22, 23, 25–38	6, 9–12, 14, 16, 17	19
эквимакс	22, 23, 25–38	9–12, 14, 16	15, 19	22, 23, 25–38	5–7, 9–12, 14	15, 18–21	20, 22, 23, 25–38	6, 9–12, 14, 16, 17	19

Как видно из таблицы, факторы по всем методам вращения практически идентичны. Сравнение данных, полученных ранее с помощью кластерного анализа показало, что результаты не совпадают.

Для более ясного представления о распределении переменных использовались диаграммы рассеяния. Для трех факторов диаграммы изображены в трехмерном пространстве (рис. 2).

Результаты, полученные методом главных факторов, центроидным методом и методом главных компонент, позволяют выделить семь условных групп близких параметров текста.

Первая группа. Признаки 1, 4, 8, 13, 18, 22, 23, 25, 40, 43 и 46.

Вторая группа. Признаки 2, 9, 14, 24, 41, 47 и 49.

Третья группа. Признаки 3, 26—38.

Четвертая группа. Признаки 5—7.

Пятая группа. Признаки 10—12, 16 и 17.

Шестая группа. Признаки 15, 19—21.

Седьмая группа. Признаки 39, 42, 44, 45 и 48.

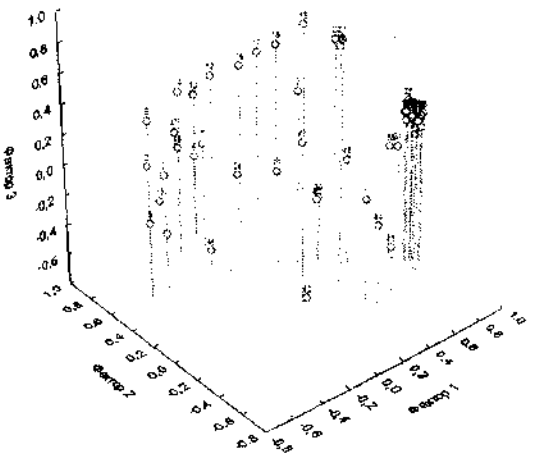

Рис. 2. – Диаграмма рассеяния признаков для метода главных

Выводы. Проведенный анализ позволил определить достаточное количество признаков для дальнейшего исследования данных о влиянии параметров текста на его читабельность. Для последующей обработки достаточно пользоваться одним признаком из каждой группы, например, длина текста в абзацах, длина текста в словах, длина текста в буквах. Полученные данные использовались для построения решающего правила, т.е. методики отнесения объекта к какому-либо классу.

ЛИТЕРАТУРА

1. *Волчек, Е. З.* Философия: учеб. пособие с хрестоматийными извлечениями / Е. З. Волчек. Мн.: Интер-пресссервис, Экоперспектива, 2003. 544 с.

2. *Спиркин, А. Г.* Философия: учебник для студентов высших учебных заведений / А. Г. Спиркин. 2-е изд. М.: Гардарики, 2004. 736 с.

3. Философия: учебное пособие для студентов высших учебных заведений / В. С. Степин [и др.]; под общ. ред. Я. С. Яскевич. Мн.: РИВШ, 2006. 624 с.

4. Философия: учебное пособие для студентов высших учебных заведений / Ю. А. Харин [и др.]; под общ. ред. Ю. А. Харина. Мн.: ТетраСистемс, 2006. 448 с.

5. *Сажина, М. А.* Основы экономической теории: учебное пособие для неэкономических специальностей вузов / М. А. Сажина, Г. Г. Чибриков; отв. ред. и рук. авт. коллектива П. В. Савченко. М.: Экономика, 1995.

6. Экономическая теория: учебник / Н. И. Базылев, А. В. Бондарь, С. П. Гурко и др.; под общ. ред. Н. И. Базылева, С. П. Гурко. Мн.: Экоперспектива, 1997.

7. Экономическая теория: учебник для студентов вузов / Под ред. В. Д. Камаева. 6-е изд., перераб. и доп. М.: ВЛАДОС, 2001.

8. Экономическая теория: учебное пособие / Л. Н. Давыденко, А. И. Базылева, А. А. Дичковский и др.; под общ. ред. Л. Н. Давыденко. Мн.: Вышэйшая школа, 2002.

10. *Айвазян, С. А.* Прикладная статистика и основы эконометрики: учебник для вузов / С. А. Айвазян, В. С. Мхитарян. М.: ЮНИТИ, 1998. 1022 с.

11. *Kaiser, H. F.* The application of electronic computers to factor analysis / H. F. Kaiser // Educational and Psychological Measurement. 1960. № 20. P. 141–151.

12. *Cattell, R. B.* The scree test for the number of factors / R. B. Cattell // Multivariate Behavioral Research. 1966. № 1. P. 245–276.