ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 004.93'1

Невдах М. М., аспирант (БГТУ); **Зильберглейт М. А.**, доктор химических наук, профессор, заведующий кафедрой (БГТУ)

АВТОМАТИЗИРОВАННАЯ ОЦЕНКА ТРУДНОСТИ УЧЕБНЫХ ТЕКСТОВ

В статье приведены результаты исследования, посвященного автоматизированной оценке учебных текстов для высшей школы. Для достижения цели в работе на первом этапе были проведены эксперименты с использованием различных методик для получения объективных критериев относительно трудности текстов. На втором этапе были выделены и вычислены значения 49 параметров учебных текстов. Снижение признакового пространства осуществлялось методами многомерного статистического анализа. Для разработки решающего правила использовался дискриминантный анализ. На основе полученных дискриминантных функций создана программа Readability analysis, предназначенная для автоматизации оценки трудности учебных текстов для студентов вузов.

In article results of the research devoted to the automated estimation of educational texts for the higher school are resulted. For purpose achievement in work at the first stage experiments with use of various techniques for reception of objective criteria concerning difficulty of texts have been made. At the second stage values of 49 parameters of educational texts have been allocated and calculated. Decrease in quantity of parameters was carried out by methods of the multidimensional statistical analysis. For working out of a solving rule the discriminant analysis was used. On the basis of the received discriminant functions program Readability analysis, intended for automation of an estimation of difficulty of educational texts for students of high schools is created.

Введение. Проблема качества учебных изданий является одной из центральных в отечественном книгоиздании и привлекает к себе внимание широкого круга исследователей. От повышения качества учебной литературы будет зависеть совершенствование профессиональной подготовки специалистов. В настоящее время уровень учебного материала в основном зависит от профессионализма автора и редактора. Очевидно, что данная оценка не всегда является объективной. В связи с этим создание надежных и общепринятых методов автоматизированной проверки трудности учебного текста, ориентированной на потребности читателя, является крайне актуальной задачей.

Статистические методики анализа данных с поддержкой компьютерных технологий обладают огромным потенциалом в разрешении многих практических задач обработки текстовых массивов. Одной из областей анализа текстов с точки зрения его доступности для читателя является читабельность, под которой следует понимать некоторую характеристику печатного материала, зависящую от всех элементов внутри данного материала, которые влияют на успешность его усвоения определенной группой читателей. Мерой такого успешного усвоения является то, насколько средний читатель интересующей нас группы понимает исследуемый материал, в какой мере скорость, с которой он его читает, приближается к оптимальной и, наконец, какой интерес представляет данный материал для этого среднего читателя.

В настоящее время отсутствуют исследования в области читабельности с использованием современных информационных технологий и необходимого инструментария для классификации русскоязычных текстов по ряду областей знаний в зависимости от подготовленности читателя.

В связи с этим цель работы – автоматизированная оценка трудности учебных текстов по философии и экономической теории для высшей школы. Для решения цели можно выделить несколько задач:

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.

Основная часть. *Экспериментальным материалом* послужили учебные издания для ву-

зов по философии и экономической теории [1–12]. Всего было отобрано 48 отрывков длиной 1800–2000 печатных знаков. Выбор данной величины обусловлен тем, что в статье [13] показано: начиная с объема в 1800 печатных знаков, статистические характеристики текста становятся относительно постоянными.

Оценка трудности учебных текстов проводилась среди студентов старших курсов Белорусского государственного технологического университета. В основном эксперименте приняли участие 75 студентов.

На первом этапе были проведены эксперименты с использованием различных методик. С этой целью проанализированы основные методы определения трудности понимания текста: постановка вопросов к тексту, сводка основного содержания текста, методика дополнения, экспертные оценки трудности текста испытуемыми, составление плана или схем текста, угадывание текста по буквам, интонирование, пересказ, скорость чтения текста. В нашем исследовании использовались наиболее надежные методы: методика дополнения и метод балльных оценок. Кроме того, впервые для оценки трудности понимания учебного материала для вузов использовался метод парных сравнений.

Методика дополнения — это заполнение пропусков в тексте, в котором слова через определенный интервал заменены точками. Плюсы данной методики состоят в том, что пропускается всегда только одно слово, и слова пропускаются не по усмотрению исследователя, а по строгому правилу. В текстах на основе результатов предварительного эксперимента пропускалось каждое седьмое слово.

Суть *балльных оценок* трудности текста заключалась в следующем: после прочтения отрывка испытуемым предлагалось оценить его трудность по семибалльной шкале: 1 — сверхлегкий текст; 2 — очень легкий; 3 — легкий; 4 — текст со средней трудностью; 5 — трудный; 6 — очень трудный; 7 — сверхтрудный текст. Для того чтобы исключить поверхностное знакомство испытуемых с текстом и возможное искажение результатов при оценке его трудности, студентам перед суждением о трудности понимания текста по шкале предлагалось выписать несколько ключевых слов и выразить основное содержание отрывка одним предложением. При проведении методики дополнения и экспертных оценок фиксировалось также время работы с текстом.

Суть *метода парных сравнений* заключалась в том, что каждому испытуемому предлагался набор текстов, размещенных парами, и после прочтения студент должен был указать, какой из отрывков обладает заданным признаком (в нашем случае — какой отрывок воспринимается легче). Оценка каждого текста производилась путем сравнения с каждым другим текстом того же набора. Так как у нас в наборе имелось 24 отрывка по философии и столько же по экономической теории, следовательно, по одному предмету было составлено 276 пар. За один этап эксперимента студенту предъявлялось 8 пар текстов. Такое количество не вызывало утомления у испытуемого.

Обработка и анализ результатов экспериментов позволили выявить информацию относительно трудности восприятия учебного материала для вузов по философии и экономической теории. На основании полученных данных найдены пять объективных критериев, определяющих трудность текста: процент правильно заполненных пропусков (Y_1); относительное время работы с текстом (Y_2) — с использованием методики дополнения; средняя оценка трудности восприятия текста (Y_3); относительное время работы с текстом (Y_4) — с использованием балльных оценок; ранг текста (Y_5). Результаты экспериментов были сведены в таблицу, фрагмент которой представлен в табл. 1.

Таблица 1

Критерии трудности учебных текстов для высшей школы

Номер теста	Метод исследования				
	методика дополнения (процент правильно заполненных пропусков)	относительное время работы с текстом	экспертные оценки испытуемых	относительное время работы с текстом	метод парных сравнений (ранг)
1	78,82	38,55	4,21	0,017	6
2	77,11	30,74	3,51	0,014	22
3	71,02	39,29	3,31	0,020	12
4	53,69	35,44	3,63	0,024	14
5	55,19	52,86	4,28	0,022	2
6	65,82	35,52	4,03	0,020	11
…	…	…	…	…	…
48	55,46	28,40	4,57	0,022	3

Для каждого показателя была найдена середина диапазона всех полученных значений, в соответствии с которой производилось разбиение текстов на две группы (трудный — легкий текст). В итоге было получено разбиение текстов на группы по выделенным пяти показателям трудности. Фрагмент табл. представлен ниже (табл. 2).

Объективная трудность учебных текстов определялась путем анализа компонентов сложности текстов. Для этого ***на втором этапе*** были выделены и вычислены значения 49 параметров учебных текстов по философии и экономической теории: 1) длина текста в абзацах; 2) длина текста в словах и др.

Очевидно, что использование большого количества параметров текста является неэффективным по ряду причин: а) сильная взаимосвязанность признаков, что приводит к дублированию информации; б) неинформативность признаков, мало меняющихся при переходе от одного объекта к другому; в) возможность агрегирования по некоторым признакам. С другой стороны, ничем не оправданное уменьшение числа переменных может привести к потере точности экспериментов.

Для снижения числа признаков были использованы кластерный и факторный анализ, метод корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование.

Так как характеристики текста измерялись в различных единицах, то все данные были стандартизированы. Для этого использовалась нормализация, приводящая все переменные к стандартной z-шкале. Для анализа данных и проведения статистического анализа использован пакет SPSS.

Кластерный анализ представляет собой многомерную статистическую процедуру, выполняющую сбор данных о выборке объектов и упорядочивающую их в сравнительно однородные группы.

В этом исследовании при анализе данных в качестве критерия для определения подобия групп использовались следующие меры сходства: а) расстояние Евклида; б) квадрат расстояния Евклида; в) косинус угла; г) коэффициент корреляции; д) неравенство Чебышева; е) расстояние Минковского; ж) манхэттенское расстояние.

Для кластеризации выделенных характеристик текста использовались следующие основные алгоритмы метода кластерного анализа: метод простого среднего (межгрупповое связывание), метод группового среднего (внутригрупповое связывание), метод ближнего соседа (одиночное связывание), метод дальнего соседа (полное связывание), невзвешенный центроидный метод (центроидная кластеризация), взвешенный центроидный метод (центральное связывание), метод Варда. Количество кластеров по каждому алгоритму варьировалось от 3 до 10. После выбора всех соответствующих параметров получена информация о формировании кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик текста к тому или иному кластеру.

Выводимые результаты для наглядности были представлены и в виде дендрограмм, которые позволяют не только перейти к любому признаку на любом уровне кластеризации, но и дают возможность судить о том, каково расстояние между кластерами или признаками на каждом из уровней. Пример дендрограммы по центроидному методу на основе манхэттенского расстояния приведен на рис. 1.

Таблица 2

Разбиение текстов на группы в соответствии с субъективной оценкой трудности текста (1 — легкий текст, 0 — трудный текст)

Номер теста	Метод исследования				
	методика дополнения (процент правильно заполненных пропусков)	относительное время работы с текстом (по первому методу)	экспертные оценки испытуемых	относительное время работы с текстом (по второму методу)	метод парных сравнений
1	1	0	0	1	0
2	1	1	1	1	1
3	1	0	1	0	0
4	0	1	1	0	1
5	0	0	0	0	0
6	1	1	0	0	0
…	…	…	…	…	…
48	0	1	0	0	0

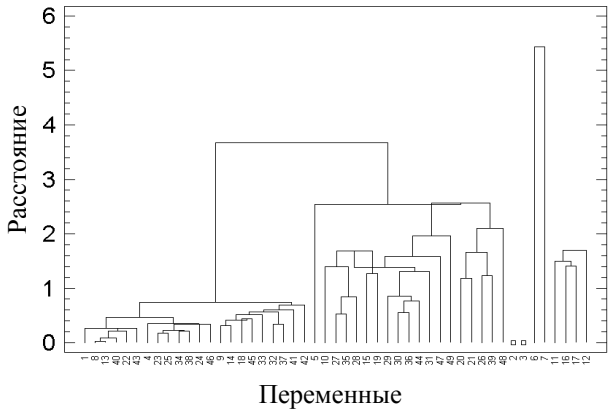

Рис. 1. Дендрограмма по центроидному методу на основе манхэттенского расстояния для 5 кластеров

В результате анализа данных о влиянии исследуемых характеристик текста с использованием всех известных алгоритмов и мер сходства были получены 784 дендрограммы, которые отражают кластеризацию переменных в условные группы.

В применении процедур кластерного анализа немаловажным аспектом является устойчивость структуры кластеров, отражающая реальную объективность классификации. В качестве методов проверки устойчивости могут быть использованы бутстреп-метод, предложенный Б. Эфроном в 1977 г., методы «складного ножа» и «скользящего контроля» [14]. 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 сводные таблицы, в которых четко прослеживаются особенности применения различных алгоритмов кластерного анализа, использующих разные меры сходства. Результаты формирования кластеров для текстов по философии согласуются практически по всем алгоритмам. Незначительно отличаются данные по методу Варда и центроидной кластеризации. Сравнение результатов с применением различных мер сходства показало, что наблюдаются заметные различия лишь в данных, полученных методами измерения близости, основанных на косинусах и корреляции векторов значений.

Для текстов по экономической теории результаты формирования кластеров по различным алгоритмам отличаются только по методу Варда. При использовании различных мер сходства наблюдаются заметные различия лишь в данных, полученных методами измерения близости, основанных на корреляции векторов значений и манхэттенском расстоянии.

Проведенный кластерный анализ для учебных текстов показал, что целесообразно выделить следующие группы признаков: философия — 1, 4, 8, 13, 18, 22–25, 33, 34, 38, 40, 43, 45, 46; 2; 3; 5; 6, 7; 9, 14, 19, 30–32, 36, 37, 41, 42, 44; 10, 15, 27–29, 35, 47, 49; 11, 12, 16, 17; 20, 21, 26, 39, 48 (9 групп); экономика — 1, 4, 8, 13, 22–25, 40, 42–44, 46; 2, 9, 14, 18; 3, 39, 45, 48; 5–7; 10–12, 16, 17; 15, 19–21; 26–30, 35–37; 31–34, 38, 41; 47, 49 (9 групп). Для последующей обработки достаточно пользоваться одним признаком из каждой группы.

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.

При изучении экспериментальных данных было установлено, что первые три фактора объясняют около 74% разброса дисперсии для текстов по философии, около 64% — для текстов по экономической теории.

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 и критерий «каменистой осыпи» Р. Кэтелла.

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

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между факторами (табл. 3–4).

Таблица 3

Распределение характеристик текстов по философии 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, 25–40	1, 5–14, 16, 17, 47, 48	18, 41, 42	22, 23, 25–40	1, 5, 14, 16, 17, 47, 48	18, 41, 42	22, 23, 25–40	1, 5, 14, 16, 17, 47, 48	18, 41, 42
Квартимакс	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42	19–23, 25–40	1, 5–7, 9–17, 47, 48	41	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42
Эквимакс	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42	19–23, 25–40	1, 5–7, 9–17, 47, 48	41	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42

Таблица 4

Распределение характеристик текстов по экономической теории 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 4	фактор 1	фактор 2	фактор 3	фактор 4	фактор 1	фактор 2	фактор 3	фактор 4
Варимакс	2, 22–25, 27–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
Квартимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
Эквимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7

Как видно из таблиц, факторы по всем методам вращения для текстов по философии и экономической теории практически идентичны. Для более ясного представления о распределении переменных использовались диаграммы рассеяния.

Результаты, полученные методом главных факторов, центроидным методом и методом главных компонент, позволяют выделить следующие группы признаков: философия — 1, 44, 48; 2, 3, 5–17, 24, 47, 49; 4, 46; 18–21, 41; 22, 23, 25–40; 42, 43 (6 групп); экономика — 1, 4, 8, 13, 22–25, 40, 42–44, 46; 2, 9, 14, 18; 3, 39, 45, 48; 5–7; 10–12, 16, 17; 15, 19–21; 26–38, 41; 47, 49 (8 групп).

Для снижения признакового пространства использовался и метод *корреляционных плеяд*. Выделение корреляционных плеяд осуществлялось следующим образом: упорядочивались признаки и рассматривались только те коэффициенты корреляции, которые соответствуют связям между элементами в упорядоченной системе. Упорядочение производилось на основании принципа максимального корреляционного пути. Для удобства построения графа были составлены упорядоченные корреляционные матрицы, фрагмент одной из которых представлен в табл. 5.

На основании упорядочения всех признаков были построены графы, которые представляют собой кратчайший незамкнутый путь. Пример графа для текстов по философии приведен на рис. 2. В графах соединены все исследуемые параметры текстов. Если задать определенное пороговое значение коэффициента корреляции (r_0), то полученные графы можно разбить на подграфы (плеяды), проводя разрыв между признаками со значением сопряженности, меньшим r_0.

Таблица 5

Упорядоченная корреляционная матрица исходных признаков текстов по философии

№ п/п	47	48	8	13	4	5	6	7	1	…	43
47	1	1,000	0,878	0,818	0,624	0,468	0,307	0,248	0,249	…	0,225
48		1	0,878	0,818	0,624	0,467	0,306	0,247	0,249	…	0,226
8			1	0,764	0,733	0,525	0,336	0,256	0,312	…	0,228
13				1	0,530	0,474	0,360	0,291	0,263	…	0,174
4					1	0,796	0,650	0,561	0,694	…	0,346
5						1	0,953	0,913	0,808	…	0,308
6							1	0,981	0,831	…	0,273
7								1	0,762	…	0,173
1									1	…	0,402
…										…	…
43											1

С использованием прямого (z) и обратного (z^{-1}) преобразования Фишера был определен коэффициент r_0 для заданного объема выборки. Исходя из поставленной цели и анализа корреляционной матрицы исследуемых характеристик текстов, задали пороговый коэффициент корреляции r = 0,9, что позволило выявить наиболее связанные друг с другом признаки. Исходный граф распался на пять подграфов для текстов по философии и шесть подграфов — для текстов по экономической теории. Признаки, не вошедшие в выделенные группы, требуют дальнейшего исследования. Использование метода корреляционных плеяд позволило выделить следующие группы близких параметров текста: философия — 1; 2; 3; 4; 5-7; 8; 9–12, 14–21; 22, 23, 25–38; 24; 39, 40; 41; 42; 43; 44; 45; 46; 47, 48; 49 (18 групп); экономическая теория — 1; 2; 3; 4; 5–7; 8; 9, 11, 12, 14, 16–18, 20, 21; 10; 13; 15, 19; 22–25, 27–33, 35–37; 26; 34; 38; 39, 40; 41; 42; 43; 44; 45; 46; 47, 48; 49 (23 группы).

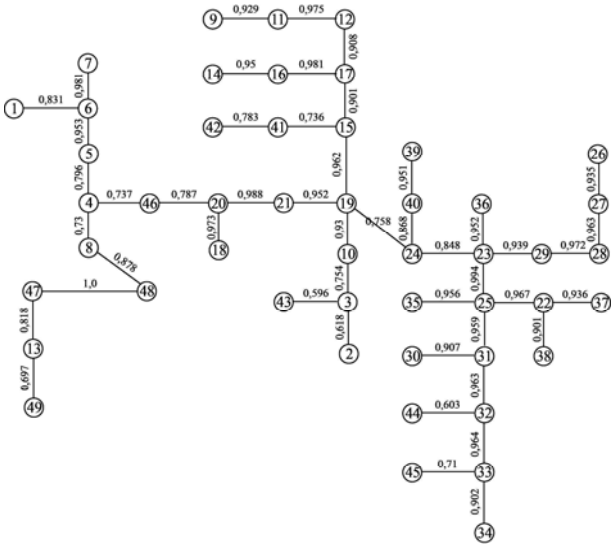

Рис. 2. Граф максимального корреляционного пути (для текстов по философии)

С помощью метода *вроцлавской таксономии* было получено нелинейное упорядочение изучаемых элементов текста. С целью построения дендрита вычислены матрицы расстояний (на основе расстояния Евклида) между изучаемыми характеристиками учебных текстов.

Далее из составленных матриц расстояний между признаками были выбраны единицы с близкими значениями. В результате для текстов по философии получены следующие пары признаков с близкими значениями: 1–22, 2–12, 3–7, 4–25, 5–39, 6–7, 7–6, 8–13, 9–14, 10–15, 11–12, 12–11, 13–8, 14–9, 15–10, 16–17, 17–16, 18–46, 19–30, 20–21, 21–20, 22–1, 23–25, 24–46, 25–23, 26–39, 27–35, 28–35, 29–36, 30–19, 31–37, 32–37, 33–34, 34–38, 35–27, 36–44, 37–32, 38–34, 39–26, 40–13, 41–31, 42–45, 43–22, 44–36, 45–18, 46–18, 47–10, 48–26, 49–15. Далее были найдены пары с общим признаком, которые затем объединялись друг с другом. Например, пары 43–22 и 22–1 образовали цепочку 43–22–1. В результате было получено 16 отдельных конструкций, называемые скоплениями 1-го порядка.

Полученные скопления не удовлетворяют основному условию дендрита, а именно они не связаны в единое целое. Для достижения этой цели было выбрано наименьшее расстояние между единицами, входящими в различные скопления 1-го порядка. В результате получены скопления 2-го порядка. Объединение признаков в скопления 3-го, 4-го, n-го порядков происходило до тех пор, пока любые две точки исследуемого множества параметров не оказались связанными друг с другом.

Исходя из поставленной цели и анализа дендрита, определили максимальную величину расстояния между признаками, равную 0,08 для текстов по философии, и 0,15 — для текстов по экономической теории. В первом случае исходный дендрит распался на семь наиболее связанных друг с другом групп признаков, во втором — на пять.

Использование метода вроцлавской таксономии позволило выделить следующие группы признаков: философия — 1, 4, 8, 13, 18, 22, 23, 25, 33, 34, 38, 40, 43, 45, 46; 2; 3; 5; 6; 7; 9, 14; 10; 11, 12; 15; 16, 17; 19; 20, 21; 24; 26; 27, 35; 28; 29; 30; 31, 32, 37; 36; 39; 41; 42; 44; 47; 48; 49 (28 групп); экономическая теория — 1, 4, 8, 9, 13, 14, 18, 22–25, 34, 38, 40–43, 45, 46; 2; 3; 5; 6; 7; 10; 11; 12; 15, 19; 16; 17; 20; 21; 26, 39; 27, 28, 35; 29–33, 36, 37; 44; 47; 48; 49 (21 группа).

Основная задача *многомерного шкалирования* заключалась в преобразовании исходной матрицы 49×49 в гораздо более простую матрицу 49×2 и визуальном ее представлении.

После расположения точек в заданном пространстве для всей модели в многомерном шкалировании были вычислены стресс и коэффициент R^2. Наилучшей моделью для текстов по философии (stress = 0,210, R^2 = 0,856) стала модель, полученная с использованием меры сходства, основанной на неравенстве Чебышева; по экономической теории (stress = = 0,230; R^2 = 0,763) — модель, полученная с использованием квадрата расстояния Евклида. На их основе были получены следующие группы признаков: по философии — 1; 2, 4–14, 47, 49; 3, 15–18; 19–21, 45; 22, 23, 25–38; 24; 39, 40; 41; 42, 43; 44, 48; 46 (11 групп); по экономической теории — 1, 39, 40, 46; 2; 3, 4, 9–12, 14, 16, 17, 47, 49; 5–7, 18; 8, 13,42, 44; 15, 19–21, 43; 22–38, 41, 45, 48 (7 групп).

Для дальнейшего изучения характеристик текста важнейшей задачей является выделение наиболее информативного признака из каждой полученной группы. В данной работе для оценки информативности признаков в качестве информационной использовалась мера $J(1, 2)$ расхождения между статистическими распределениями 1 и 2. Для дискретных распределений эта мера вычисляется по формуле

$$J(x_i / A_1, x_i / A_2) = \sum_j J(x_i / A_1, x_i / A_2) =$$

$$= \sum_j \lg \frac{P(x_{ij} / A_1)}{P(x_{ij} / A_2)} \left[P(x_{ij} / A_1) - P(x_{ij} / A_2) \right],$$

где A_1 и A_2 — классы, которым может принадлежать рассматриваемый объект; j — номер диапазона признака x_i; i — номер признака; $P(x_{ij} / A_1)$ и $P(x_{ij} / A_2)$ — вероятность попадания объекта, принадлежащего к A_1 или к A_2, в диапазон j признака x_i.

По данной формуле были вычислены информационные меры каждого из 49 признаков, а затем отобраны те из них, которые обладают наибольшей информативностью среди признаков своей группы. В результате число признаков было сокращено до возможного минимума.

Для дальнейшего исследования характеристик текста и их влияния на понятность учебного материала ***на третьем этапе*** использовался дискриминантный анализ. При анализе дискриминантных функций в учет принимались только функции, у которых процент точности классификации – максимальный, а количество переменных при этом минимальное.

В текстах по философии чаще других фигурировали следующие признаки (в порядке убывания): 3 (длина текста в буквах), 24 (средняя длина слов в печатных знаках), 9 (средняя длина предложения в словах), 40 (средняя частота повторения слова), 1 (длина текста в абзацах), 43 (процент конкретных существительных), 48 (процент простых предложений); *в текстах по экономической теории*: 39 (процент неповторяющихся слов), 47 (процент сложных предложений), 48 (процент простых предложений), 15 (средняя длина самостоятельного предложения в слогах), 42 (процент повторяющихся существительных), 41 (процент неповторяющихся существительных).

Выделенные признаки позволяют сделать важный вывод относительно факторов трудности текста. Они связаны, прежде всего, с объемом текста (признак 3), длиной слов и предложений (признаки 9, 15, 24), со сложностью организации текста (признаки 1, 47, 48), богатством словаря и абстрактностью изложения материала (признаки 39, 40–43).

Анализ результатов дискриминантного анализа показал, что для учебных текстов по философии наилучшими являются следующие дискриминантные функции:

$$F_1 = -53{,}06 + 15{,}10X_{24} + 0{,}83X_9 - 0{,}01X_3;$$

$$F_2 = -42{,}72 + 8{,}66X_{24} + 0{,}55X_9 + 0{,}01X_3.$$

Точность классификации при данном наборе дискриминантных переменных составляет 91,6% (22 из 24 правильных предсказаний).

Для учебных текстов по экономической теории наилучшими являются следующие дискриминантные функции:

$$F_1 = -92{,}96 + 2{,}62X_{39} - 0{,}03X_{47};$$

$$F_2 = -111{,}93 + 2{,}96X_{39} - 0{,}20X_{47}.$$

Точность классификации при данном наборе дискриминантных переменных составляет 87,5% (21 из 24 правильных предсказаний).

Для автоматизированной оценки рукописи на основе полученных функций создана программа Readability analysis, предназначенная для автоматизации оценки трудности учебных текстов для студентов вузов.

Практическая значимость программы связана с тем, что она может быть использована в редакционно-издательской деятельности при подготовке учебной литературы для высшей школы. Анализ трудности текста на стадии его подготовки и дальнейшее усовершенствование материала позволят привести уровень сложности учебного текста в соответствие со способностями читателей.

Выводы. Результаты исследования дают возможность продолжить автоматизацию редакционно-издательского процесса. Полная или частичная замена человека специализированной системой позволит добиться не только невозможного для человека быстродействия, но и необходимого качества изданий благодаря объективной оценке трудности текста исходя из его информационных характеристик, полученных на основе восприятия читателей.

Литература

1. Волчек, Е. З. Философия: учеб. пособие с хрестоматийными извлечениями / Е. З. Волчек. — Минск: Экоперспектива, 2003. — 544 с.

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

3. Философия: учеб. пособие для студентов высших учебных заведений / В. С. Степин [и др.]. — Минск: РИВШ, 2006. — 624 с.

4. Философия: учеб. пособие для студентов высших учебных заведений / Ю. А. Харин [и др.]. — Минск: ТетраСистемс, 2006. — 448 с.

5. Сажина, М. А. Основы экономической теории: учеб. пособие для неэкономических специальностей вузов / М. А. Сажина, Г. Г. Чибриков. — М.: Экономика, 1995. — 368 с.

6. Экономическая теория: учебник / Н. И. Базылев [и др.]. — Минск: Экоперспектива, 1997. — 368 с.

7. Экономическая теория: учеб. для студентов вузов / под ред. В. Д. Камаева. — М.: ВЛАДОС, 2001. — 640 с.

8. Экономическая теория: учеб. пособие / Л. Н. Давыденко [и др.]. — Минск: Вышэйшая школа, 2002. — 366 с.

9. Кажуро, Н. Я. Основы экономической теории / Н. Я. Кажуро. — Минск: ФАУинформ, 2001. — 672 с.

10. Экономическая теория: учеб. пособие / В. Л. Клюня [и др.]. — Минск: ТетраСистемс, 2001. — 400 с.

11. Курс экономической теории: учеб. пособие / под общ. ред. М. Н. Чепурина, Е. А. Киселёвой. — Киров, 1994. — 624 с.

12. Экономическая теория: учебник / В. И. Антипина [и др.]. — М.: ТК Велби; Проспект, 2002. — 576 с.

13. 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.

14. Количественные методы в исторических исследованиях / под ред. И. Д. Ковальченко. — М.: Высшая школа, 1984. — 384 с.

Поступила 01.04.2011