УДК 004.93'1

Невдах М. М., аспирант (БГТУ); М. А. Зильберглейт, профессор (БГТУ)

РАЗРАБОТКА МЕТОДИКИ ДЛЯ ОЦЕНКИ ТРУДНОСТИ ТЕКСТОВ И ЕЕ ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

В статье рассмотрены основные этапы разработки метода автоматизированной оценки трудности учебных текстов. С помощью дискриминантного анализа выделены основные признаки, влияющие на усвоение учебного текста, и вычислены дискриминантные функции, на основе которых появляется возможность отнести каждый объект (текст), в том числе и неизвестный, к одной из известных групп (легкий — трудный). Полученные расчеты использованы для создания программного обеспечения, помогающего автоматизировать оценку понятности учебного материала для высшей школы.

In the article the basic development cycles of a method of the automated estimation of difficulty of educational texts are considered. By means of the discriminant analysis the basic signs influencing mastering of the educational text are allocated, and discriminant functions on which basis there is a possibility to carry each object (text) including the unknown, to one of known groups (easy — difficult) are calculated. The received calculations are used for creation of the software, helping to automate an estimation of clearness of a teaching material for the higher school.

Введение. 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 характеристику печатного материала, зависящую от его элементов внутри данного материала, которые влияют на успешность его усвоения определенной группой читателей.

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 и практически не связан со способностями читателей. 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.

Основная часть. 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 был решен ряд задач. Экспериментальным материалом послужили учебные издания для вузов по философии и экономической теории. Всего было отобрано 64 отрывка длиной 1800–2000 печатных знаков. Выбор данной величины обусловлен тем, что в [1] показано, что, начиная с объема в 1800 печатных знаков, статистические характеристики текста становятся относительно постоянными.

На первом этапе исследования были найдены объективные критерии, определяющие трудность восприятия текста отмеченной категорией читателей. С этой целью проведены эксперименты с помощью наиболее надежных из проанализированных методов: методики дополнения и балльных оценок. В качестве вспомогательного метода использовалась скорость чтения отрывка испытуемыми. Кроме того, впервые для оценки трудности понимания учебного материала для вузов был использован метод парных сравнений. В качестве испытуемых выступили 75 студентов Белорусского государственного технологического университета.

По данным проведенных экспериментов были найдены пять объективных критериев, определяющих трудность текста: процент правильно заполненных пропусков и относительное время работы с текстом (с использованием методики дополнения), средняя оценка трудности восприятия текста и относительное время работы с ним (с использованием балльных оценок), ранг текста [2].

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 получено разбиение текстов на группы по выделенным пяти показателям трудности.

Объективная трудность определялась путем анализа компонентов сложности текста. Для этого на втором этапе были выделены и вычислены значения 49 параметров учебных текстов по философии и экономической теории: длина текста в абзацах, длина текста в словах, длина текста в буквах, средняя длина

абзаца в фразах, средняя длина абзаца в словах и др. [3]. Использование такого большого числа характеристик для практических целей вызывает определенные трудности. В первую очередь это связано с тем, что данные параметры могут быть сильно коррелированны. С другой стороны, ничем не оправданное уменьшение числа переменных может привести к потере точности экспериментов.

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

При кластеризации исследуемых характеристик текста 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 измерения близости между кластерами использовались следующие методы: метод простого среднего, метод группового среднего, метод ближнего соседа, метод дальнего соседа, невзвешенный центроидный метод, взвешенный центроидный метод (медиана), метод Варда.

Количество кластеров по каждому алгоритму варьировалось от 3 до 10. 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 [4].

Далее для выделения групп связанных признаков был проведен факторный анализ, использованы следующие методы: а) метод главных факторов; б) центроидный метод; в) метод главных компонент. Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между факторами. Для более ясного представления о распределении переменных использовались диаграммы рассеяния [5].

При использовании метода корреляционных плеяд, исходя из определенного правила по корреляционной матрице признаков, был построен граф максимального корреляционного пути, который затем разбивался на подграфы, или плеяды [6].

С помощью метода вроцлавской таксономии получено нелинейное упорядочение изучаемых элементов текста. На основе матрицы расстояний между признаками был построен дендрит. Исходя из поставленной цели и анализа дендрита, определена максимальная величина расстояния между признаками. Исходный дендрит распался на семь групп взаимосвязанных признаков [7].

Применение метода многомерного шкалирования позволило представить расположение признаков в двумерном пространстве. Для визуального представления признаков на диаграмме использовались те же меры сходства, что и в кластерном анализе. Для каждой диаграммы был рассчитан коэффициент стресса, характеризующий отклонение результата от первоначальной модели. Чем ближе значение коэффициента к нулю, тем точнее матрица исходных расстояний согласуется с матрицей результирующих расстояний.

Сравнение результатов для учебных текстов по философии и экономической теории, полученных с помощью разных методов многомерного статистического анализа, позволил сделать следующий вывод: во многих случаях совпадают не только отдельные признаки в группах, но и сами группы. Из этого следует, что характеристики учебного текста для высшей школы по различным отраслям знаний целесообразно изучать в рамках единого информационного поля.

Прежде чем разработать решающее правило, из каждой полученной группы взаимосвязанных признаков следовало выделить по одному элементу, дающему наиболее полную информацию об изучаемом объекте, т. е. информативный признак. В данной работе для оценки информативности признаков в качестве информационной использовалась мера $J(1, 2)$ расхождения между статистическими распределениями 1 и 2, подробно изученная С. Кульбаком [8]. 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 j — номер диапазона признака x_i; i — номер признака; A_1 и A_2 — классы, которым может принадлежать рассматриваемый объект; $P(x_{ij}/A_1)$ 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 мы сократили число признаков до возможного минимума.

Для дальнейшего исследования характеристик текста и их влияния на понятность учебного материала использовался дискриминантный анализ, на основе которого было разработано решающее правило для отнесения учебных текстов по философии и экономической теории к группе легких или трудных [9].

На последнем этапе на основе разработанного решающего правила создана программа Readability analysis, предназначенная для автоматизации оценки трудности учебных текстов для студентов вузов. Программа написана на языке Delphi (алгоритм представлен на рис. 1) и включает в себя три подпрограммы: 1) «Расчет текстовых параметров» (рис. 2); 2) «Вычисление дискриминантных функций» (рис. 3); 3) «Вывод результатов» (рис. 4).

Первая подпрограмма проверяет объем текста, который должен превышать 1800 символов, и рассчитывает основные функции; вторая — вычисляет дискриминантные функции для текстов по философии и экономической теории; третья — сравнивает функции между собой и на этом основании выводит результаты относительно трудности текста.

Рис. 1. Укрупненная блок-схема программы Readability analysis

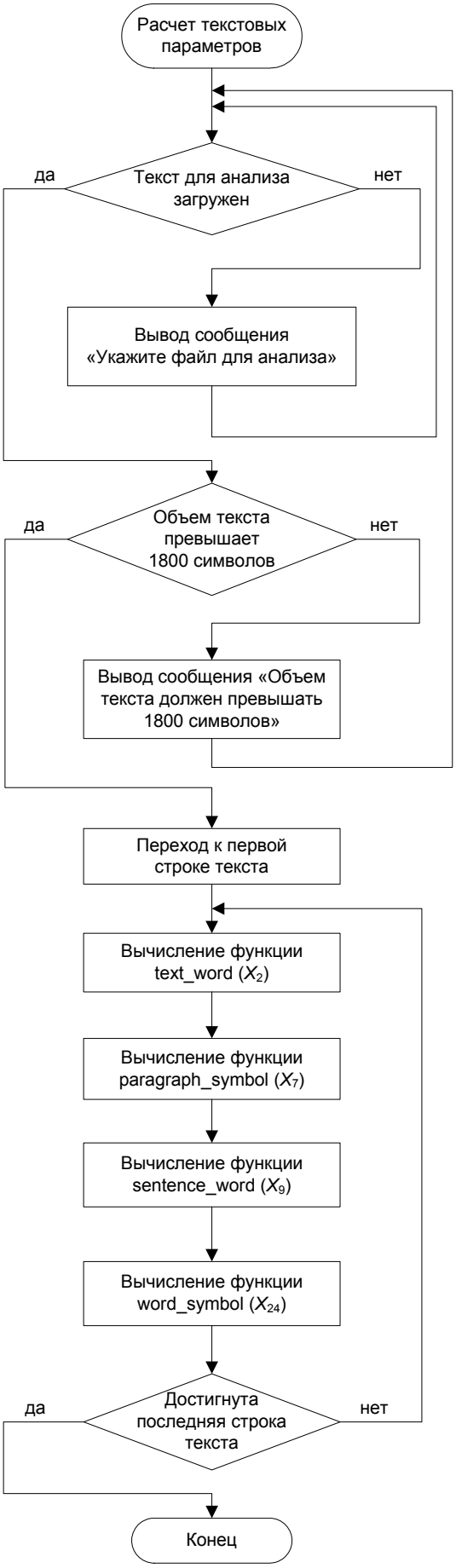

Рис. 2. Алгоритм подпрограммы «Расчет текстовых параметров»

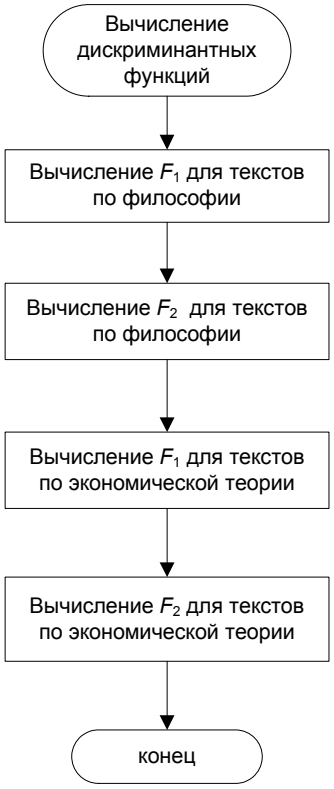

Рис. 3. Алгоритм подпрограммы «Вычисление дискриминантных функций»

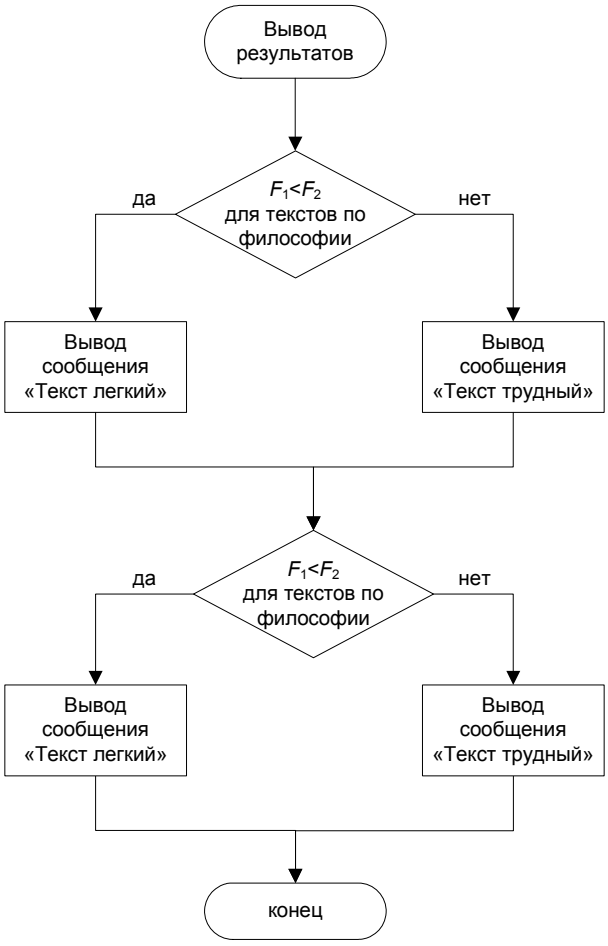

Рис. 4. Алгоритм подпрограммы «Вывод результатов»

Входные данные программы — текст на русском языке. Выходные данные представляют собой таблицу значений статистических характеристик текста и оценку уровня трудности текста для будущих читателей.

Оценка трудности учебных текстов с помощью созданной программы имеет следующие основные цели:

– определение количественных характеристик текста, влияющих на его трудность;

– запись статистических данных текста в отдельный файл с возможностью накапливания результатов по различным исследуемым текстам;

– автоматизированный анализ трудности текста в соответствии со способностями студентов.

Статистика количественных показателей текстов может накапливаться и в дальнейшем использоваться для сравнительного анализа характеристик текстов.

Программа выполняет вычисление характеристик текста, влияющих на уровень его трудности, и содержит следующие функции:

1. *Функция text_word* определяет длину текста в словах.

Следует отметить, что текстовый редактор MS Word после проверки правописания выдает статистику удобочитаемости, которая не всегда верно отражает реальные данные о количестве предложений и слов для текста на русском языке. Например, скобка, поставленная через пробел, считается как слово. Поэтому в программе Readability analysis определены символы, часто используемые для набора текста, но не являющиеся его словами: различные виды скобок ( ), [ ], { }, < >, звёздочки *, ***, чёрточки и др.

2. *Функция paragraph_symbol* определяет среднюю длину абзаца в печатных знаках. Разделителем абзацев является наличие двух и более пробелов в начале строки. Перед загрузкой текста в окно программы необходимо устранить лишние пробелы между словами, установить абзацный отступ, сохранить документ в формате .txt.

3. *Функция sentence_word* определяет среднюю длину предложения в словах. Как было сказано выше, MS Word выдает неверные результаты относительно количества предложений. Если строки текста отделены знаком абзаца, то каждая строка определяется как предложение. В программе Readability analysis в данной функции определены правила выявления конца предложения. Это знаки: точка (.), восклицательный знак (!), восклицательный знак и многоточие (!..), вопросительный знак (?), вопросительный знак и многоточие (?..), многоточие (…).

Следует отметить, что в учебных текстах (как и в любых других) часто встречаются ини-

циалы или другие сокращения, состоящие из одной буквы с точкой (например, К. Маркс, т. д., т. е.). В этом случае точка не является признаком конца предложения. Такие сочетания должны определяться как слово, но не увеличивать число предложений. В разработанной программе один символ с точкой не считается предложением.

4. *Функция word_ symbol* определяет среднюю длину слов в печатных знаках. Разделителем слов является пробел.

Все вычисленные количественные характеристики текста формируются в выходной массив и выводятся в виде списка в элемент Statistics окна программы. Результаты могут быть сохранены в отдельном текстовом файле.

Программа быстро справляется с обработкой больших массивов текстовой информации. Время оценки текста от нескольких секунд до двух минут.

Практическая значимость программы Readability analysis 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.

Заключение. 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 на основе его информационных характеристик, полученных в опоре на восприятие читателей.

Литература

1. 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.

2. Невдах, М. М. Разработка количественных методов оценки трудности восприятия учебного текста для высшей школы / М. М. Невдах // Труды БГТУ. Сер. IX, Издат. дело и полиграфия. — 2008. — Вып. XVI. — С. 87–90.

3. Невдах, М. М. Выделение текстовых характеристик, определяющих сложность учебного текста для высшей школы / М. М. Невдах, Ю. Ф. Шпаковский // Актуальныя праблемы мовазнаўства і лінгвадыдактыкі: матэрыялы Рэспубліканскай навуковай канферэнцыі (да 70-годдзя з дня нараджэння прафесара Г. М. Малажай), Брэст, 20–21 сакавіка 2008 г. — Брэст: БрДУ, 2008. — С. 202–204.

4. Невдах, М. М. Систематизация информационных характеристик учебного текста с использованием метода кластерного анализа / М. А. Зильберглейт, М. М. Невдах // Информатика. — 2008. — № 2. — С. 111–118.

5. Невдах, М. М. Систематизация информационных характеристик учебного текста методами факторного анализа / М. А. Зильберглейт, М. М. Невдах // Изв. вузов. Проблемы полиграфии и издательского дела. — 2008. — № 4. — С. 58–65.

6. Невдах, М. М. Анализ информационных характеристик учебных текстов с использованием эвристического метода корреляционных плеяд / М. М. Невдах // Электроника инфо. — 2008. — № 5. — С. 47–50.

7. Невдах, М. М. Упорядочение характеристик текста по философии методом вроцлавской таксономии / М. М. Невдах // Сб. докл. Международной научно-практической конференции студентов, аспирантов и молодых ученых (10–11 апреля): в 2 ч. — 2008. — Ч. I. — С. 165–168.

8. Гублер, Е. В. Вычислительные методы анализа и распознавания патологических процессов. — Л.: Медицина, 1978. — 296 с.

9. Невдах, М. М. Исследование информационных характеристик учебного текста методами многомерного статистического анализа / М. М. Невдах // Прикладная информатика. — 2008. — № 4. — С. 117–130.

Поступила 02.04.2010