БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

УДК 303.732.4

РЫЖАНКОВА
Анастасия Сергеевна

Методика поддержки принятия решений при обработке и оценке текстовых учебных материалов

Автореферат диссертации на соискание ученой степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка информации

Минск, 2015

Научная работа выполнена в Учреждении образования
«Белорусский государственный технологический университет»

Научный руководитель	**Зильберглейт Марк Аронович** доктор химических наук, профессор, УО «Белорусский государственный технологический университет»
Официальные оппоненты:	**Старовойтов Валерий Васильевич** доктор технических наук, профессор, главный научный сотрудник, Объединенный институт проблем информатики НАН Беларуси
	Герман Олег Витольдович кандидат технических наук, доцент, Белорусский государственный университет информатики и радиоэлектроники
Оппонирующая организация	Республиканское унитарное предприятие «Криптотех» Гознака

Защита состоится «29» апреля 2015 г. в 15^{00} час. на заседании совета по защите диссертаций К 02.05.01 при Белорусском национальном техническом университете по адресу: 220013, г. Минск, пр. Независимости, 65 в зале заседаний Ученого Совета, e-mail: gurski2010@gmail.ru, тел. 8(017)2939564.

С диссертацией можно ознакомиться в библиотеке Белорусского национального технического университета.

Автореферат разослан «27» марта 2015 г.

Ученый секретарь
Совета по защите диссертаций К 02.05.01,
кандидат технических наук, доцент Н. Н. Гурский

КРАТКОЕ ВВЕДЕНИЕ

Движущим фактором трансформации рынка продукции и услуг стала масштабная компьютеризация производства и развитие информационных сетей, обеспечивающих высокоскоростную автоматизированную обработку данных и минимизацию количества ошибок. Ряд разработок автоматизированного анализа текста активно применяется и в сфере редакционно-издательского дела: сканирование и распознавание информации, атрибутивный анализ и проверка текстов на подлинность, проверка орфографии и пунктуации, обработка терминологических данных, перевод текста, работа с pdf-файлом и др.

Внедрение компьютерных технологий на допечатной стадии подготовки издания явило собой необратимый процесс трансформации всего производства. Помимо ярко выраженных достоинств такой модернизации, четко проявляются ее недостатки, связанные со спецификой книжной продукции. Издание представляет собой продукт духовной и творческой деятельности, именно поэтому для его комплексной оценки необходим анализ не только полиграфического исполнения, но и внутреннего содержания. Оценка последнего носит часто субъективный характер и может быть дана только рецензентами либо специалистами той отрасли, проблемы которой изложены в издании.

Текст перестали воспринимать как художественный продукт, теперь в нем видят образование, подчиненное математическим законам и управляемое пользователем через запросы. Зная основные характеристики текста, можно определить возрастные градации читателя и объем достаточной для восприятия информации.

Компьютерные технологии позволяют сформировать признаковое пространство текстового материала, определить его статистическую структуру, выполнить подсчет необходимых показателей. Формально-логическая структура текста является ориентиром для развития исследований по удобочитаемости и эффективности, что помогает определить направление работы редакции и отказаться от текстов, не отвечающих требованиям содержания.

Анализ результатов предыдущих исследований, используемых для обработки текстовой информации, позволил сделать вывод о том, что большая часть разработок предназначена для проверки орфографии, лемматизации, автоматического перевода, группировки родственных данных, семантического анализа и пр.

В этих работах практически отсутствуют исследования, связанные с разработкой методики оценки удобочитаемости учебного текстового материала, в них не проводится анализ на устойчивость классификационных правил к помехам и ошибкам.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В настоящее время работа над учебным текстовым материалом носит субъективный характер. По каким бы принципам не выполнялась оценка текста, каким бы требованиям он не соответствовал, редактор и рецензент оценивают его с позиций собственного восприятия читабельности и эффективности.

Развитие современных информационных технологий позволяет оценить текст иными методами, а также провести объективный и независимый анализ учебного материала. Для этого необходимо изучать его не как продукт художественной деятельности человека, а как статистическую структуру, подчиненную законам построения и функционирования. Объединение различных подходов к обработке учебного материала позволит достигнуть поставленной цели в автоматизированной оценке качества учебного текста и поддержке принятия решений при его обработке.

Связь работы с научными программами и темами. Диссертационное исследование выполнено в соответствии с планом научно-исследовательских работ, проводимых кафедрой редакционно-издательских технологий Белорусского государственного технологического университета, среди которых: ГБ 42-11 «Системный анализ количественных и качественных характеристик текста» (2011−2015 г.) (участие); ГБ 13-037 «Исследование формальных характеристик текста методами многомерного статистического анализа» (руководитель).

Тема диссертационного исследования соответствует «Перечню приоритетных направлений научно-технической деятельности в Республике Беларусь» (Постановление Совета Министров Республики Беларусь от 19 апреля 2010 г. № 585 «Об утверждении перечня приоритетных научных исследований Республики Беларусь на 2011−2015 гг»); подпункту 5.1. «Методы математического и компьютерного моделирования, компьютерные технологии и интеллектуальные системы поддержки принятия решений»; 5.4. «Математические и интеллектуальные методы, информационные технологии и системы распознавания и обработки образов, сигналов, речи и мультимедийной информации».

Цель и задачи исследования. Целью диссертационной работы является разработка методики оценки и поддержки принятия решений при анализе текстовых учебных материалов.

Для достижения цели были сформулированы и решены следующие задачи:

− установить минимальный объем текстового фрагмента, на основе анализа которого можно получить достоверные информационные характеристики, описывающие статистическую структуру издания в целом;

− оценить качество текстовых учебных материалов на примере литературы по специальности «Издательское дело»;

− сформировать факторное пространство, описывающее статистическую структуру всего издания в целом;

− на основе применения методов распознавания образов с учителем и без него определить метод классификации текстовых учебных материалов по специальности «Издательское дело» на предмет их трудности и удобочитаемости, соответствующий результатам экспериментальной части работы;

− сформулировать устойчивые решающие правила разделения объектов в виде классификационных функций;

− разработать программное средство поддержки принятия решения при оценке текстовых учебных материалов на примере специальности «Издательское дело»;

− оценить влияние полиграфических характеристик издания на качество восприятия текстового учебного материала и установить их взаимосвязь.

Научная новизна. Впервые обоснован выбор объема текстового фрагмента в размере 1800−2000 символов, позволяющего описать статистическую структуру текста; установлено, что формальная структура текста состоит из трех классов; впервые установлен метод классификации текстовых учебных материалов на предмет их трудности и удобочитаемости и предложен метод дискриминантного анализа для формальной диагностики текста; впервые получено устойчивое решение классификации текстовых учебных материалов на основе преобразования статистических параметров текста при помощи логарифмической функции; впервые разработано программное средство «MAZI» − инструмент оценки и принятия решений при анализе текстовых учебных материалов на предмет их трудности и удобочитаемости.

Положения, выносимые на защиту:

1. Методом дисперсионного анализа установлено, что объем текстового фрагмента в размере 1800−2000 символов является достаточным (представительным) для оценки статистической структуры текста.

2. Формальную структуру текста можно представить в виде трех классов, в число которых входит класс, характеризующий общую длину предложения в той или иной системе измерения; класс, характеризующий дифференциальную структуру текста; класс, характеризующий оценку отдельных слов в различных единицах измерения.

3. Метод классификации текстовых учебных материалов на предмет их трудности и удобочитаемости установлен на основе применения методов обучения с учителем и без учителя, что позволило предложить метод дискриминантного анализа для формальной диагностики текста.

4. Устойчивое решающее правило классификации текстовых учебных материалов получено на основе преобразования статистических параметров текста при помощи логарифмической функции.

5. Методика и программное средство «MAZI» (Математический анализатор заданий издательства) являются инструментом оценки и поддержки принятия решений при анализе текстовых учебных материалов.

Личный вклад соискателя ученой степени. Все изложенные в диссертации результаты получены автором лично на основе описанных в работе экспериментальных методов и методов многомерного статистического анализа. Вклад соавторов заключается в научном руководстве при постановке задач, анализе и описании результатов. Автор принимал непосредственное участие в получении, обработке, интерпретации экспериментальных данных, в написании и подготовке научных публикаций, в разработке алгоритма и его программной реализации, во внедрении результатов исследования в редакционно-издательскую практику.

Апробация диссертации и информация об использовании ее результатов. Основные положения, методика и результаты исследований обсуждались и получили положительную оценку на различных международных и республиканских научно-практических конференциях: VI Машеровские чтения: Международная научно-практическая конференция студентов, аспирантов и молодых ученых (27–28 сентября 2012 г., г. Витебск); 76, 77, 78 научно-технические конференции профессорско-преподавательского состава, научных сотрудников и аспирантов. Секция «Издательское дело и полиграфия» (2012 г., 2013 г., 2014 г., г. Минск); Управление информационными ресурсами: Международная научно-практическая конференция (21 ноября 2012 г., Академия управления при Президенте Республики Беларусь); I Республиканская научно-практическая конференция молодых аналитиков «Повестка 2015» (26–27 сентября 2013 г., СОК «Бригантина», Республика Беларусь); 10-я Международная научно-практическая конференция на тему «Информационные технологии в образовании, консультационной деятельности и сельскохозяйственном производстве» (24–25 апреля 2013 г., г. Новочеркасск, Российская Федерация); Информационные технологии и системы 2013 (ITS 2013): Международная научная конференция (БГУИР, г. Минск, 23 октября 2013 г.); Международная научно-практическая конференция «Молодежь XXI века: образование, наука, инновации» (4 декабря 2014 г., г. Витебск).

Опубликование результатов диссертации. Результаты диссертационной работы опубликованы в 6 статьях, из них 3 рецензируемые научные периодические издания (~ 1 авторский лист) соответствуют пункту 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь; а также отражены в 4 публикациях сборников материалов научных конференций; тезисах 5 докладов на конференциях.

Структура и объем диссертации. Диссертационная работа состоит из введения, общей характеристики работы, четырех глав, заключения, списка литературы и приложений. Общий объем диссертации составляет 169 с., в том числе 133 с. текста, 60 таблиц, 25 рисунков, 5 приложений. В приложении приведены экспериментальные материалы и исходные коды разработанного программного средства. Результаты исследований подтверждены справками внедрения в издательствах республики и регистрацией программного продукта в Национальном центре интеллектуальной собственности Республики Беларусь.

ОСНОВНАЯ ЧАСТЬ

Первая глава содержит аналитический обзор литературных источников по теме диссертационного исследования. В нем приведены работы, посвященные изучению статистической структуры текста с позиций формально-логического подхода.

Система обработки текстов включает электронные словари, орфокорректоры, поисковые системы, системы машинного перевода, автоматизированного дистрибутивно-статистического анализа и др. Среди их особенностей можно отметить: скорость и точность обработки, реферирование, индексирование, навигацию, статистический частотный анализ, автоматическую классификацию, кластеризацию и смысловой поиск.

Анализ имеющихся на IT-рынке компьютерных разработок, а также доступных в сети Интернет приложений позволил выявить программы, программные модули, системы и компоненты, направленные на анализ и обработку текстового материала. Среди них: Light Reader, Readability analysis, WordStat, Лингвоанализатор, Система Пропись, WordTabulator, WordStat, Concordance 3.3, Склонятель, Морфологический анализатор, Худломер, Система Полиглот, TextAnalyst 2.0, Рабочее место лингвиста и др. Среди специализированных средств обработки данных выделены инструменты реализации контент-анализа и пакеты статистической обработки: STADIA, SYSTAT, STATGRAPHICS, STATA, STATISTIKA, JMR, NCSS, PRISM, Статэксперт, Эвриста, R, EpiInfo, PSPP, SOFA и др.

Отдельное внимание в аналитическом обзоре уделено вопросу оценки удобочитаемости и эффективности восприятия учебного текстового материала. Отмечена необходимость измерения, анализа и корректировки трудности его понимания. Решить данные задачи позволяют формулы, полученные для определения удобочитаемости текста. В аналитическом обзоре представлены классические формулы читабельности (Флеша, Флеша − Кинкейда, Ганнинга, SMOG, Дейла − Чолл, Пауэрса – Самнера − Кера, FORCAST, Спеша, Колемана − Лиау, автоматический индекс читабельности, формула письма; два графических метода (графики Фрая и Рэйгора)), а также формулы, выведенные для отдельных языков, целей и возрастных групп.

В обзоре рассмотрены и диагностические методы оценки учебного материала: метод семантического дифференциала; дополнения; заканчивания предложения; косвенного исследования семантики; парных сравнений; градуальное шкалирование; опросник; подчеркивание ключевых слов; «компрессия» текста; построение шкал понимания и др.

Анализ литературных источников показал, что работы по изучению учебного материала чаще всего выполняются в педагогическом направлении. Формулы читабельности рассчитаны в большинстве случаев на школьные издания, построены на анализе иноязычных текстов, позволяют установить возрастную группу читателей, уровень их подго-

товки, стиль текста, но не определяют, насколько планируемое издание отвечает требованиям читателей, насколько оно интересно и актуально. По результатам анализа не выявлено также программных средств, оценивающих уровень подготовки учебного материала на основе применения устойчивых решений.

Вторая глава содержит методическую часть диссертационного исследования. В ней рассмотрены характеристики объектов анализа – текстовые учебные материалы изданий по специальности «Издательское дело»; установлен достоверный для исследования статистических характеристик объем текстового фрагмента; описаны методы, этапы и результаты опроса; сформированы обучающие выборки, а также проведен обзор методов многомерного статистического анализа.

По результатам дисперсионного анализа 378 фрагментов объемом 500−2000 символов установлено, что объем текстовой информации, при котором ее статистические показатели находятся на относительно однородном уровне, составляет 1800−2000 символов. *F*-проверка результатов на равноточность и совместимость при использовании текстовых фрагментов объемом 30 000 символов и более, анализ показателей квадратичного отклонения разностей s_δ и максимальной погрешности этой разности ε подтвердили сформулированные выводы.

Для получения оценок, характеризующих трудность учебных текстовых материалов с позиций обучающихся, в работе использованы три метода опроса: метод балльных оценок (МБО), метод дополнений (МД), метод парных сравнений (МПС). Количество респондентов, принявших участие в опросе, составило 735 человек. Экспериментальная часть исследования позволила определить качество текстового материала с позиций обучающихся, а также установить уровень его восприятия и эффективность прочтения.

Для реализации опроса по МБО предложена 9-балльная оценочная шкала. Средний балл учебного текстового материала по результатам опроса составил 5,03 балла (базовый уровень восприятия). Количественным показателем уровня восприятия текстового фрагмента по МД определено интегральное значение суммы. Среднее количество удаленных слов в каждом из текстовых фрагментов составило 50−60 единиц, количество заполненных и представленных к анализу бланков – 764 единицы, ячеек – 40 535 единиц. Парное сравнение объектов выборки позволило установить текстовый фрагмент, который обучающиеся относят к группе наиболее доступных к восприятию учебных материалов. Количественным показателем уровня восприятия по данному методу опроса является количество упоминаний текстового фрагмента.

По результатам эксперимента были установлены пороговые значения и сформулированы обучающие выборки, согласно которым по МБО: 69 объектов являются сложными, 32 – легкими; по МД: 85 – легкими, 16 – сложными; по МПС: 79 – легкими, 22 – сложными.

Для анализа и обработки данных были использованы методы: кластерного, регрессионного, дискриминантного, факторного, дисперсионного анализа; классификация с помощью деревьев решений и искусственных нейронных сетей; анализ данных методом эталонов; ближайших соседей; главных компонент; кратчайшего незамкнутого пути, а также применения меры *l*.

Третья глава посвящена изучению статистической структуры факторного пространства текстовых учебных материалов. Были выделены и определены 14 существенных параметров: N_1 − средняя длина слов в слогах; N_2 − средняя длина слов в буквах; N_3 − средняя длина слов по Деверу; N_4 − средняя длина слов в 3 слога и более; N_5 − средняя длина слов в 4 слога и более; N_6 − средняя длина слов в 5 слогов и более; N_7 − средняя длина слов в 6 слогов и более; N_8 − средняя длина слов в 7 слогов и более; N_9 − процент односложных слов; N_{10} − средняя длина предложения в словах; N_{11} − средняя длина предложения в слогах; N_{12} − процент чисел от общего количества слов; N_{13} – отношение показателя N_4 к N_7; $N_{14} – N_5$ к N_7.

Для описания статистической структуры факторного пространства методом корреляционных плеяд сформированы три графа. Методом факторного анализа и методом главных компонент выделены 3 фактора, объясняющие около 96% и 84,5% дисперсии соответственно, а также сформированы 5 классов факторного пространства. Методами кластерного анализа выделены 5 кластеров, классифицирующих издания выборки. Методом кратчайшего незамкнутого пути построены деревья, отражающие структуру факторного пространства.

Установлено, что статистическая структура текстовых фрагментов имеет различную форму, однако можно выделить общие факторы для всех методов. Они объединяются в классы: 1-й класс описывает длину предложения (N_{10}; N_{11}); 2-й класс − дифференциальную структуру текста (N_{13}; N_{14}); 3-й класс характеризует длину слов в различных единицах измерения (N_1−N_9). Так как в таких группах отсутствует возможность выбора наиболее представительных факторов, простейшую статистическую структуру объектов можно описать в виде отдельных представителей каждой из них, например: N_{10}; N_{13}; N_1.

Четвертая глава посвящена исследованию методов распознавания объектов с использованием обучающей выборки и без нее. К методам распознавания без использования обучающей выборки относятся: кластерный анализ, факторный анализ, метод корреляционных плеяд и метод главных компонент.

Результаты распознавания, полученные методом кластерного анализа, не совпадают с результатами эксперимента.

Метод корреляционных плеяд не может быть применен для разделения объектов по классам, так как большое количество обрабатываемых данных и наличие минимальных значений коэффициентов парной

корреляции $r \leq 0,7$ препятствует построению плеяд и не позволяют объединить рассматриваемые объекты в классы.

В результате выполнения операций факторного анализа получены значения веса каждого фактора. Установлено, что первые 3 фактора объясняют 84% дисперсии. После проведения дополнительной операции кластеризации были сформированы 2 класса объектов. Установлено, что 100 объектов выборки относятся к классу «легкий уровень восприятия текстовой информации», 1 объект – к классу «сложный уровень восприятия текстовой информации». Таким образом, метод не позволяет провести разделение, соответствующее задачам исследования. Такие же результаты были получены и при анализе данных методом главных компонент.

Следовательно, методы распознавания объектов без учителя не пригодны для достижения поставленной цели исследования.

Данная глава исследования содержит также анализ, посвященный распознаванию объектов на основе применения обучающих выборок.

В результате выполнения набора операций по построению линейного уравнения множественной регрессии для каждого из методов опроса, а также связи зависимых переменных с несколькими независимыми были получены величины оценок коэффициентов регрессии, стандартные ошибки коэффициентов, *t*-статистики, величины скорректированного коэффициента детерминации и др.

Анализ *F*-критерия Фишера для уравнения регрессии, составленного по МБО, показал, что при уровне значимости $\alpha = 0,05$ $F_{табл.} < F_{факт.}$, следовательно, уравнение регрессии статистически значимо, надежно; по МД уравнение регрессии статистически незначимо, ненадежно ($F_{табл.} > F_{факт.}$); по МПС уравнение регрессии также статистически незначимо, ненадежно ($F_{табл.} > F_{факт.}$). Значения критерия Дарбина − Уотсона, которые по трем методам менее 2,5, характеризуют отсутствие автокорреляции (таблица 1).

Таблица 1. – Результаты регрессионного анализа

Показатель	МБО	МД	МПС
R-квадрат , %	38,25	18,36	21,35
R-квадрат нормированный (*d.f.*), %	28,20	5,07	8,55
Стандартная ошибка	0,45	8,55	23,65
Среднее отклонение	0,32	6,48	18,50
Критерий Дарбина – Уотсона	1,86	1,71	1,77

Значения коэффициента детерминации *R*-квадрат и *R*-квадрат нормированный (*d.f.*) позволили сделать вывод о неадекватности модели: по трем методам опроса значение *R*-квадрат не превысило 50%. Таким образом, метод не может быть использован для построения модели оценки текстовых учебных материалов.

В модели искусственной нейронной сети выделено 4 уровня. Первый слой состоит из 14 входных переменных (статистические парамет-

ры текста), второй − содержит нейрон для каждой из 101 переменной выборки, третий – суммирует информацию из каждого класса и относит ее к выходному слою, указанному в обучающей выборке. Неизвестные классы классифицированы нахождением звена исходного слоя, который имел наибольшую вероятность или большую скорость. Обработка данных методом искусственной нейронной сети показала относительно невысокие результаты: процент верно классифицированных объектов по МБО составляет 49,50%; по МД – 79,21%; по МПС – 73,27%.

Распознавание объектов по методу эталонов выполнено по следующему правилу: объект относится к тому классу, расстояние от которого до эталона минимально. Результаты распознавания данным методом показали, что доля верно установленных принадлежностей к одному из классов не превышает 70%, соответственно, использовать метод для классификации объектов не целесообразно (таблица 2).

Таблица 2. – Результаты анализа методом эталонов

Метод	Обучающая выборка, объектов	Распознаны верно, объектов	Результат, %
МБО	1 класс – 69; 2 класс – 32	1 класс – 37; 2 класс – 21	53,60/65,63
МД	1 класс – 85; 2 класс – 16	1 класс – 54; 2 класс – 11	63,50/68,75
МПС	1 класс – 79; 2 класс – 22	1 класс – 50; 2 класс – 13	63,30/59,09

Для распознавания принадлежности объектов к классу методом трех ближайших соседей за основу взяты матрицы расстояний, полученные в результате обработки данных методом кластерного анализа. Для определения минимальных значений в каждом из классов построена сфера наименьшего радиуса с центром в точке так, что в нее попали не менее трех точек выборки. Лучше всего программа определяет объекты, относящиеся к первому классу; при отнесении ко второму классу наблюдается полное нераспознавание (МД). Также высок показатель нераспознанных объектов (таблица 3). Таким образом, метод трех ближайших соседей не может быть использован для распознавания объектов, так как характеризуется низкими показателями достоверности.

Таблица 3. – Результаты анализа методом трех ближайших соседей

Метод	Обучающая выборка, объектов	Распознаны верно, объектов	Результат, %
МБО	1 класс – 69; 2 класс – 32	1 класс – 52; 2 класс – 4	75,4/12,5
МД	1 класс – 85; 2 класс – 16	1 класс – 82; 2 класс – 0	96,5/0
МПС	1 класс – 79; 2 класс – 22	1 класс – 69; 2 класс – 5	87,3/22,7

Метод пяти ближайших соседей также не может быть использован для распознавания, так как характеризуется низким показателем достоверности результатов при распознавании объектов второго класса сложности (таблица 4). Количество нераспознанных объектов по МБО составляет 36; по МД – 17; по МПС – 25.

Таблица 4. – Результаты анализа методом пяти ближайших соседей

Метод	Обучающая выборка, объектов	Распознаны верно, объектов	Результат, %
МБО	1 класс – 69; 2 класс – 32	1 класс – 60; 2 класс – 5	87/15,6
МД	1 класс – 85; 2 класс – 16	1 класс – 84; 2 класс – 0	98,8/0
МПС	1 класс – 79; 2 класс – 22	1 класс – 73; 2 класс – 3	92,4/13,6

Результаты анализа на основе использования меры l характеризуются низким показателем правдоподобия. Сформулированные решающие правила позволили разделить объекты на два класса, однако процент распознанных верно объектов по трем методам не превысил 50%.

Для решения задачи исследования был применен метод построения деревьев решений (рисунок 1, таблица 5). В качестве алгоритма построения использован метод *CART*, в основе которого лежит индекс *Gini*. Результаты построения деревьев решений показали, что наибольшее влияния на классификацию объектов оказывают факторы: N_{10}, N_6, N_4 (МБО); N_3, N_{14}, N_7 (МД); N_6, N_1, N_4 (МПС).

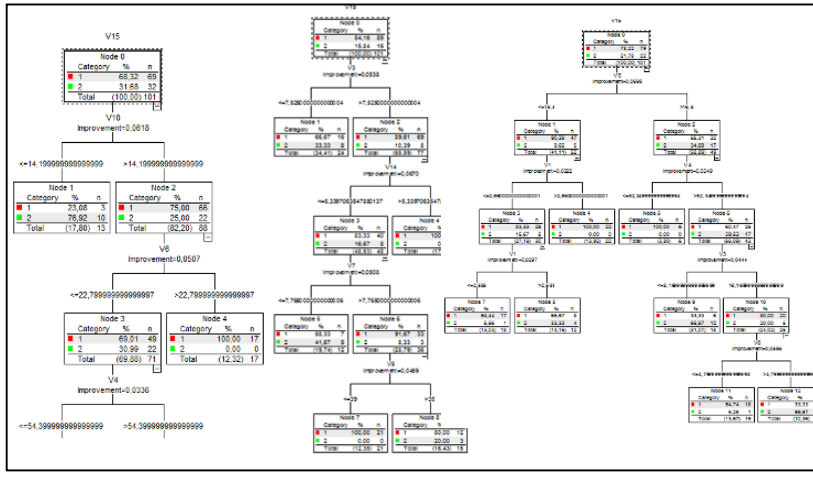

Рисунок 1. – Графические представления деревьев решений (фрагмент)

Таблица 5. – Результаты анализа методом деревьев решений

Класс	МБО (69/32)		Класс	МД (85/16)		Класс	МПС (79/22)	
	1	2		1	2		1	2
1	41	0	1	50	0	1	63	2
2	28	32	2	38	16	2	16	20
Всего, %	59,42	100	Всего, %	58,82	100	Всего, %	79,75	90,90

Таким образом, метод деревьев решений также не решает задачу классификации.

Для анализа данных методом дискриминантного анализа была построена линейная модель первого порядка n=1, однако она не дала результатов, отвечающих задачам исследования. С целью улучшения качества классификации был осуществлен переход к модели более высокого порядка. Результаты для уравнения 5-го порядка представлены в таблице 6.

Таблица 6. – Результаты дискриминантного анализа при n=5

Показатель	МБО	МД	МПС
Собственное значение	2,066	2,159	3,195
Каноническая корреляция	0,821	0,827	0,873
Коэффициент Уилк − Лямбда	0,326	0,317	0,238
Значение хи-квадрат	71,719	73,611	91,773

Установлено, что метод дискриминантного анализа характеризуется наивысшим показателем точности распознавания и классификации объектов по заранее установленным классам (таблица 7): МБО – 97,03%; МД – 97,03%; МПС – 98,02 %.

Таблица 7. – Сводная таблица результатов дискриминации

Класс	МБО			МД			МПС		
	Кол-во объектов	1	2	Кол-во объектов	1	2	Кол-во объектов	1	2
1	69	68	1	85	82	3	79	78	1
2	32	2	30	16	0	16	22	1	21

Выполнением этапа валидации на 16 новых объектах доказано, что результаты классификации практически полностью совпадают с результатами, полученными по обучающим выборкам (МБО – 96%, МД – 97%, МПС – 97%). Следовательно, классификационные функции, полученные методом дискриминантного анализа, могут быть определены как решающие правила для оценки изданий и принятия решения при отнесении их к классу «легкий уровень восприятия текстовой информации» либо «сложный уровень восприятия текстовой информации».

Известно, что методы распознавания образов достаточно «чувствительны» к изменению объема выборки. Исследований, посвященных данному вопросу по отношению к изучаемой тематике, нами не выявлено. В связи с этим для оценки устойчивости найденных решений была исследована чувствительность модели к исходным данным. В качестве такой меры нами был выбран коэффициент вариации, рассчитанный для каждого из методов преобразования как мера относительного разброса случайной величины. Формулировка выводов проведена на основе анализа трех выборок различного объема.

Для исследования изменения чувствительности к форме представления исходных данных изучены преобразования, основанные на использовании степенной, логарифмической, квадратичной функций, а

также трансформации по методу Бокса − Кокса и др. Точность классификации объектов рассчитана для МБО, МД, МПС.

При преобразовании данных методом линейного дискриминантного анализа были получены следующие результаты. Точность классификации объектов первой выборки для МБО – 84%, МД – 86%, МПС – 80%, объектов второй выборки: МБО – 86,27%, МД – 82,35%, МПС – 82,35%, объектов третьей выборки: МБО – 76,24%, МД – 83,17%, МПС – 81,19%. Среднее значение – 82,17 (МБО), 83,84 (МД), 81,18 (МПС); дисперсия – 27,66 (МБО), 3,67 (МД), 1,38 (МПС).

Одним из способов преобразования был перевод значений статистических параметров текстов в порядковую шкалу. Для каждого из факторов определены минимальные и максимальные значения и выделены границы 10 интервалов. Результаты такого преобразования: точность классификации объектов первой выборки для МБО – 80%, МД – 80%, МПС – 78%, объектов второй выборки: МБО – 72,55%, МД – 80,39%, МПС – 84,31%, объектов третьей выборки: МБО – 72,28%, МД – 89,11%, МПС – 79,21%. Среднее значение – 74,94 (МБО), 83,17 (МД), 80,51 (МПС); дисперсия – 19,20 (МБО), 26,53 (МД), 11,22 (МПС).

Метод преобразования, основанный на вычислении среднего значения каждого из выделенных интервалов, показал следующие результаты: точность классификации объектов первой выборки для МБО – 84%, МД – 94%, МПС – 80%, объектов второй выборки: МБО – 78,43%, МД – 90,20%, МПС – 84,31%, объектов третьей выборки: МБО – 72,28%, МД – 89,11%, МПС – 79,21%. Среднее значение – 78,24 (МБО), 91,10 (МД), 81,17 (МПС); дисперсия – 34,37 (МБО), 6,59 (МД), 7,54 (МПС).

Было применено преобразование, основанное на использовании логарифмической функции. Результаты: точность классификации объектов первой выборки для МБО – 86%, МД – 92%, МПС – 84%, объектов второй выборки: МБО – 84,31%, МД – 88,24%, МПС – 84,31%, объектов третьей выборки: МБО – 78,22%, МД – 85,15%, МПС – 80,2%. Среднее значение – 82,84 (МБО), 88,46 (МД), 82,84 (МПС); дисперсия – 16,75 (МБО), 11,77 (МД), 5,24 (МПС).

В результате преобразования данных методом извлечения из значений статистических параметров корня квадратного установлено: точность классификации объектов первой выборки для МБО – 90%, МД – 90%, МПС – 88%, объектов второй выборки: МБО – 86,27%, МД – 86,27%, МПС – 82,35%, объектов третьей выборки: МБО – 76,24%, МД – 83,17%, МПС – 79,21%. Среднее значение – 84,17 (МБО), 86,48 (МД), 83,19 (МПС); дисперсия – 50,64 (МБО), 11,70 (МД), 19,84 (МПС).

Одним из наиболее распространенных способов преобразования является трансформация методом Бокса − Кокса. Результаты по данному методу: точность классификации объектов первой выборки для МБО – 88%, МД – 94%, МПС – 90%, объектов второй выборки: МБО – 88,24%, МД – 88,24%, МПС – 84,31%, объектов третьей выборки: МБО – 78,22%,

МД – 84,16%, МПС – 83,17%. Среднее значение – 84,82 (МБО), 88,80 (МД), 85,83 (МПС); дисперсия – 32,68 (МБО), 24,44 (МД), 13,39 (МПС).

В результате выполнения преобразований на основе применения степенной функции установлено: точность классификации объектов первой выборки для МБО – 92%, МД – 92%, МПС – 80%, объектов второй выборки: МБО – 90,2%, МД – 84,31%, МПС – 80,39%, объектов третьей выборки: МБО – 78,22%, МД – 86,14%, МПС – 78,22%. Среднее значение – 86,81 (МБО), 87,48 (МД), 79,54 (МПС); дисперсия – 56,11 (МБО), 16,14 (МД), 1,34 (МПС).

В результате использования квадратичной функции для трансформации данных были получены результаты: точность классификации объектов первой выборки для МБО – 82%, МД – 88%, МПС – 88%, объектов второй выборки: МБО – 88,24%, МД – 88,24%, МПС – 80,39%, объектов третьей выборки: МБО – 77,23%, МД – 85,15%, МПС – 83,17%. Среднее значение – 82,49 (МБО), 87,13 (МД), 83,85 (МПС); дисперсия – 30,49 (МБО), 2,95 (МД), 14,83 (МПС).

В таблице 8 приведены результаты распознавания объектов при использовании дискриминантных функций 2-го, 3-го и 4-го порядков. В таблице 9 – статистические показатели анализа.

Таблица 8. – Результаты распознавания объектов

Выборка	Точность классификации, %								
	2-й порядок			3-й порядок			4-й порядок		
	МБО	МД	МПС	МБО	МД	МПС	МБО	МД	МПС
Первая	98	96	92	100	100	98	82	86	78
Вторая	90,2	96,08	92,16	98,04	100	98,04	45,1	82,35	21,57
Третья	78,22	86,14	87,13	83,17	90,1	91,09	87,13	95,05	94,06

Таблица 9. – Статистические показатели анализа

Показатель	2-й порядок			3-й порядок			4-й порядок		
	МБО	МД	МПС	МБО	МД	МПС	МБО	МД	МПС
Среднее	88,81	92,74	90,43	93,74	96,7	95,71	71,41	87,80	64,54
Дисперсия	99,27	32,67	8,17	84,70	32,67	16,01	525,74	42,75	1449,51
Среднеквадратичное отклонение	9,96	5,72	2,86	9,20	5,72	4,00	22,93	6,54	38,07

Как было указано ранее, в качестве меры устойчивости нами был определен коэффициент вариации, рассчитанный для каждого из представленных способов преобразования (таблица 10).

Таблица 10. – Значения коэффициентов вариации

Метод преобразования	МБО, %	МД, %	МПС, %
Линейный дискриминантный анализ	6,40	2,28	1,45
Перевод в порядковую шкалу	5,85	6,19	4,16

Окончание таблицы 10

Метод преобразования	МБО, %	МД, %	МПС, %
Среднее значение интервала	7,49	2,82	3,38
Логарифмическая функция	4,94	3,88	2,76
Извлечение корня квадратного	8,45	3,95	5,35
Метод Бокса − Кокса	6,74	5,57	4,26
Использование степенной функции	8,63	4,59	1,45
Использование квадратичной функции	6,69	1,97	4,59
Дискриминантный анализ 2-го порядка	11,22	6,16	3,16
Дискриминантный анализ 3-го порядка	9,82	5,91	4,18
Дискриминантный анализ 4-го порядка	32,11	7,45	58,99

Коэффициенты вариации для дискриминантного уравнения пятого порядка составляют для МБО – 35,77%, для МД – 64,28%, для МПС – 59,67%. На рисунках 2, 3, 4 представлены графические результаты выполненных этапов исследования. Ось абсцисс содержит значения коэффициентов вариации, ось ординат – средние значения процентного соотношения совпадений расчетов с результатами экспериментальной части.

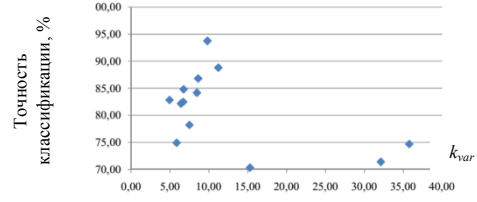

Рисунок 2. – Графические результаты анализа для МБО

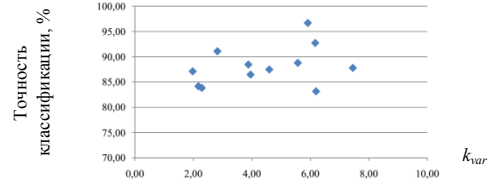

Рисунок 3. – Графические результаты анализа для МД

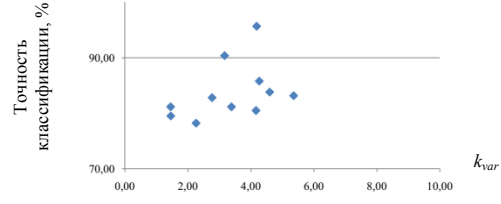

Рисунок 4. – Графические результаты анализа для МПС

В результате поиска устойчивой к помехам системы было определено, что задачам исследования отвечают преобразования двух типов, основанных на использовании логарифмической и степенной функций. Однако наименьшие значения коэффициентов вариации наблюдаются при использовании преобразования с помощью десятичного логарифма.

Полученные результаты легли в основу методики оценки качества и алгоритма программного средства «MAZI», предназначенного для принятия решений при анализе учебных текстовых материалов на предмет их трудности и удобочитаемости. Основная экранная форма разработанного программного средства представлена на рисунке 5, блок-схема – на рисунке 6.

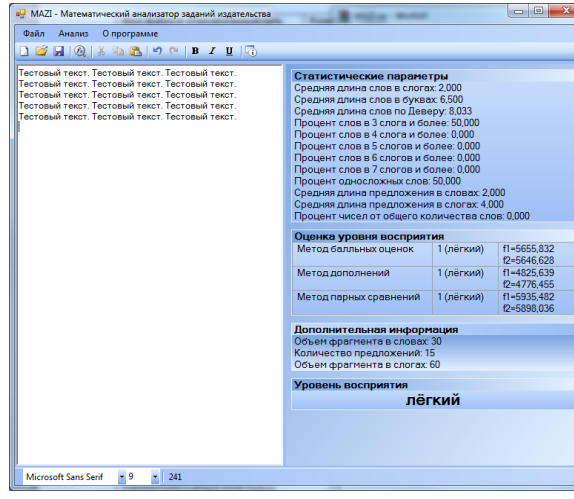

Рисунок 5. – Экранная форма программного средства «MAZI»

Программа не выдает рекомендации по усовершенствованию текста рукописи и не содержит критические значения каждого из показателей. Основное назначение «MAZI» – **дать оценку текстовому материалу** на предмет его читабельности и сложности **с позиций обучающихся, то есть принять решение при анализе сложности текста**. Результат работы программы может не совпадать с мнением редакции и автора, оценка не носит субъективный характер и получена только на основе математических вычислений.

Все предварительные расчеты выполняются в программе интерактивно, результаты выводятся в рабочем окне программы, что делает ее интерфейс легким и доступным для пользователя. Таким образом, программное средство «MAZI» выполняет поставленные в рамках диссертационного исследования задачи.

Разработанное программное средство является первым инструментом принятия решений при оценке учебных текстовых материалов, основанным на использовании устойчивых классификационных функций.

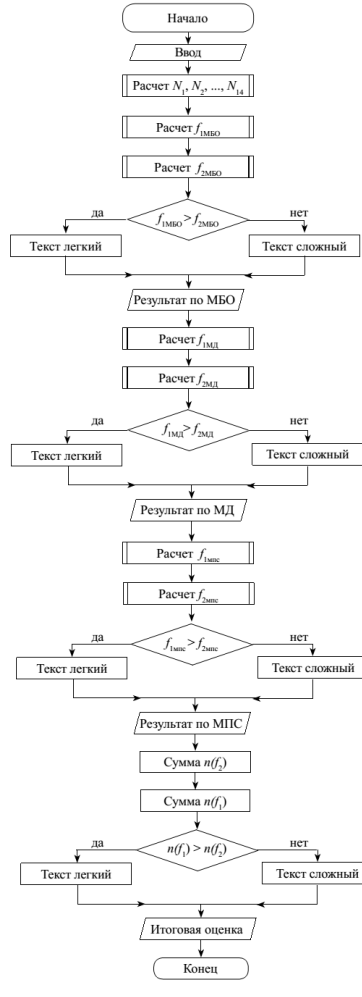

Рисунок 6. – Алгоритм анализа текста, реализованный в программном средстве «MAZI»

Для установления зависимости между издательско-полиграфическим исполнением книги и уровнем восприятия учебного материала проведена классификация изданий методом кластерного, регрессионного, дискриминантного, факторного анализа, методом деревьев решений и методом главных компонент. Установлено, что издательско-полиграфическое исполнение книги не оказывает влияния на формулировку оценки относительно трудности текста при постановке обучающимся задачи анализа учебного материала. Все методы показали отрицательные результаты.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Методом дисперсионного анализа установлено, что для получения достоверных статистических характеристик текстов следует использовать текстовые фрагменты объемом не менее 1800−2000 символов. Результаты, полученные при использовании текстового фрагмента данного объема, равноточны результатам, полученным при изучении статистических характеристик текста объемом до 30 000 символов [1−А, 7−А, 11−А, 2−А].

2. Для оценки уровня восприятия учебной литературы предложено использование метода балльных оценок, метода дополнений и метода парных сравнений, которые позволяют оценить уровни восприятия учебных материалов с позиций обучающихся. Метод балльных оценок отражает общую оценку издания, помогает установить степень интереса обучающихся и определить общий уровень восприятия учебного материала; оценка носит поверхностный характер и не раскрывает сложности внутренней структуры издания. Метод дополнений является методом, который раскрывает логичность изложения материала, показывает наличие в тексте скважистостей и отсутствие связей. Способ и вариативность заполнения пропущенных единиц являются основными показателями правильности подачи учебного материала и его архитектоники. Метод парных сравнений является закрепляющим методом, который позволяет подтвердить либо опровергнуть результаты двух первых методов эксперимента. Показана высокая корреляция результатов, полученных данными методами [5−А, 12−А, 13−А, 14−А].

3. Впервые изучена структура факторного пространства учебных текстовых материалов. Выделены статистические характеристики, оказывающие влияние на классификацию объектов.

Статистическая структура текстовых фрагментов имеет различную форму, однако путем свертки изучаемых факторов выделены общие факторы для всех используемых методов. Они объединяются в классы:

− 1 класс: средняя длина предложения в словах; средняя длина предложения в слогах;

− 2 класс: отношение показателя «Средняя длина слов в 3 слога и более» к показателю «Средняя длина слов в 6 слогов и более»; отношение показателя «Средняя длина слов в 4 слога и более» к показателю «Средняя длина слов в 6 слогов и более».

− 3 класс: средняя длина слов в слогах; средняя длина слов в буквах; средняя длина слов по Деверу; средняя длина слов в 3 слога и более; средняя длина слов в 4 слога и более; средняя длина слов в 5 слогов и более; средняя длина слов в 6 слогов и более; средняя длина слов в 7 слогов и более; процент односложных слов; процент чисел от общего количества слов.

Данная группировка может быть использована для описания структуры изученных объектов. Представить упрощенную структуру изучаемых объектов можно в виде отдельных представителей каждой из групп. Например: средняя длина предложения в словах; отношение показателя «Средняя длина слов в 3 слога и более» к показателю «Средняя длина слов в 6 слогов и более»; средняя длина слов в слогах [3−А].

4. Показано, что только часть методов распознавания – обучения с учителем − пригодна для достижения поставленных в рамках диссертационного исследования задач. К этим методам относится метод дискриминантного анализа с логарифмическим преобразованием исходных данных, что позволяет получить как наиболее высокие, так и наиболее устойчивые результаты. Для разделения на классы – легкий/сложный – можно использовать классификационные функции, полученные методом дискриминантного анализа. Точность классификации текстов составляет: метод балльных оценок – 97,03%; метод дополнений – 97,03%; метод парных сравнений – 98,02% [4−А, 8−А, 9−А, 10−А].

5. Разработанная методика и программное средство «MAZI» (Математический анализатор заданий издательства) позволяют проводить независимую оценку текстов с позиций обучающихся, основанную на анализе соответствующих статистических данных [6−А, 15−А].

6. Впервые изучена взаимосвязь полиграфических характеристик изданий с восприятием учебного материала. Отвергнуто предположение о наличии прямой зависимости между оценкой обучающимися трудности самого текста и редакционно-техническим исполнением книги. Следовательно, при работе с текстовым материалом учащиеся оценивают структуру самого текста, а не его оформление [1−А].

Рекомендации по практическому использованию результатов

1. Программное средство «MAZI» (Математический анализатор заданий издательства) заслуживает внимания с практической точки зрения и может быть применено на предприятиях: Редакционно-издательский центр ГУО «Республиканский институт высшей школы»; «Издательский центр Государственного института управления и социальных технологий БГУ»; «Издательский центр Белорусского государственного университета».

Программное средство зарегистрировано в Национальном центре интеллектуальной собственности Республики Беларусь 08.07.2014 г., Свидетельство № 679.

2. Программное средство предназначено для оценки и принятия решений при анализе учебных текстовых материалов по специальности «Издательское дело» на предмет их трудности и удобочитаемости.

3. Перспективы развития данного научного направления заключаются в том, что данная методика может быть использована для оценки учебных текстовых материалов иных технических специальностей.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Статьи в изданиях согласно перечню ВАК

1. Оценка возможности использования морфологического анализа текста для выделения стилей / Зильберглейт М. А., Малюкевич А. С. // Труды БГТУ: научный журнал. – Минск, 2012. – № 9. – С. 93–98. – ISSN 1683-0377.

2. Статистический анализ текстов учебных изданий по издательскому делу / Зильберглейт М. А., Малюкевич А. С. – Электроника-инфо: научно-практический журнал для специалистов. – Минск, 2013 г. − № 1. – С. 25–28.

3. Снижение размерности факторного пространства при изучении статистических характеристик текста / Малюкевич А. С., Зильберглейт М. А. // Труды БГТУ: научный журнал. – Минск, 2013. − № 8: Издательское дело и полиграфия. – ISSN 1683-0377. – С. 67–71.

4. Применение методов распознавания образов для оценки качества учебных текстов / Зильберглейт М. А., Малюкевич А. С. // Электроника-инфо: научно-практический журнал для специалистов. – Минск, 2013. – № 10. – С. 51–55.

5. Сравнительный анализ методов опроса и компьютерного анализа данных для изучения восприятия текстов студентами высших учебных заведений / А. С. Малюкевич, М. А. Зильберглейт // Известия ГГУ им. Ф. Скорины. – № 6 (81) Естественные науки. – ISSN 1609-9672. – Минск, 2013. – С. 134–138.

6. Рыжанкова, А. С. «Математический анализатор заданий издательства» как программное средство допечатной оценки издательского оригинала // Труды БГТУ. 2014. − № 9: Издательское дело и полиграфия. − С. 78–82.

Материалы конференций

7. Установление взаимозависимости между параметрами «количество символов» и «статистические параметры текста» методом дисперсионного анализа / Малюкевич А. С. // Машеровские чтения: материалы международной научно-практической конференции студентов, аспирантов и молодых ученых. Витебск, 27−28 сентября 2012 г. / Витебский государственный университет; редкол.: А. П. Солодков (гл. ред.) [и др.]. – Витебск: УО «ВГУ имени П. М. Машерова», 2012. – 534 с. – ISBN 978-985-517-362-6. – С. 521.

8. Анализ текстов учебных изданий по издательскому делу методом дискриминантного анализа / Малюкевич А. С. // Управление информационными ресурсами: мат-лы IX Межд. науч.-практ. конф.; Минск, 21 ноября 2012 г. / Академия управления при Президенте Республики Беларусь. – … – Минск: Академия управления при Президенте Республики Беларусь, 2012. – 295 с.

9. Анализ текстов учебных изданий методами распознавания образов / Малюкевич А. С. // Информационные технологии в образовании и консультационной деятельности, сельскохозяйственном производстве: материалы Международной науч.-произв. конф.; Новочеркасск, 24−25 апреля 2013 г. − Дон ГАУ, КПКА. – Новочеркасск, 2013. – 180 с. – ISBN 978-5-98252-204-7. − С 96−99.

10. Ryzhankova, N. Stability of dicisive rules in the analysis of educational texts // The Youth of 21st Century: Education, Science, Innovations: materials of the International Conference for Students, Postgraduates and Young Scientists; Vitebsk, December 4, 2014 / Vitebsk State University; Editorial board.: I. M. Prischepa (editor and chief) [and others]. – Vitebsk: VSU named after P. M. Masherov, 2014. – P. 12−13.

Тезисы докладов

11. Оценка возможности использования морфологического анализатора для выделения стилей / А. С. Малюкевич, М. А. Зильберглейт // Издательское дело и полиграфия : тезисы 76-й науч.-технич. конференции профессорско-преподавательского состава, научных сотрудников и аспирантов. Минск, 13−20 февраля 2012 г. [Электронный ресурс] / отв. за издание И. М. Жарский, УО «БГТУ». – Минск: БГТУ, 2012. – 45 с. – Деп. в ГУ «БелИСА» 25.04.2012 № Д201223. – С. 30.

12. Применение метода распознавания образов для оценки качества учебных текстов / Малюкевич А. С., Зильберглейт М. А. – Издательское дело и полиграфия: тезисы 77-й научно-технической конференции профессорско-преподавательского состава, научных сотрудников и аспирантов. Минск 4-9 февраля 2013 г.

13. Информатизация в сфере образования / Малюкевич А. С.// Беларусь глазами молодых аналитиков [Электронный ресурс]: тез. Респ. науч.-практ. конф. «Повестка−2015» (в рамках реализации проекта «Умные сети»). – Минск: Изд. центр БГУ, 2013. – Режим доступа: http: // www.elip.bsu.by, ограниченный. – ISBN 978-985-553-161-6. − С. 121−123.

14. Использование различных методов исследования для оценки уровня восприятия текстовой информации / Малюкевич А. С. // Информационные технологии и системы 2013 (ITS 2013) / редкол.: Л. Ю. Шилин [и др.]. – Минск: БГУИР, 2013. – 352 с.

15. Сравнительный анализ методов опроса и статистической обработки текста при оценке уровня восприятия учебной информации / Малюкевич А. С. // Издательское дело и полиграфия: тезисы 78-й научно-технической конференции профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием). Минск 3–13 февраля 2014 г. [Электронный ресурс] / отв. за издание И. М. Жарский; УО БГТУ. – Минск: БГТУ, 2014. – 50 с.

РЭЗЮМЭ

РЫЖАНКОВА Настасся Сяргееўна

МЕТОДЫКА ПАДТРЫМКІ ПРЫНЯЦЦЯ РАШЭННЯЎ ПРЫ АПРАЦОЎЦЫ І АЦЭНЦЫ ТЭКСТАВЫХ ВУЧЭБНЫХ МАТЭРЫЯЛАЎ

Ключавыя словы: устойлівасць, вучэбныя тэкставыя матэрыялы, цяжкасць і эфектыўнасць, інфармацыйныя характарыстыкі тэксту, мнагамерны статыстычны аналіз, распазнаванне вобразаў, падтрымка прыняцця рашэнняў.

Мэта даследавання: аналіз методыкі ацэнкі вучэбных тэкставых матэрыялаў і пошук сродкаў падтрымкі прынняцця рашэнняў пры іх апрацоўцы.

Метады даследавання: метад бальных ацэнак, методыка дапаўнення, метад парных параўнанняў; кластарны, рэгрэсійны, дыскрымінантны, фактарны, дысперсійны аналіз і метад галоўных кампанентаў; класіфікацыя пры дапамозе дрэваў рашэнняў і штучных нейронных сетак; метады эталонаў, бліжэйшых суседзяў, меры *l*, карацейшага незамкнутага шляху і карэляцыйных плеяд.

Атрыманыя вынікі і іх навізна. Абгрунтаваны выбар аб'ёму тэкставага фрагменту ў памеры 1800−2000 сімвалаў, які дазваляе апісаць інфармацыйную структуру тэксту; фармальная структура тэксту складаецца з трох класаў, у якія ўключаны клас, адказны за агульную даўжыню сказаў у той ці іншай сістэме вымярэння; клас, адказны за дыферэнцыяльную структуру тэксту; клас, адказны за ацэнку асобных слоў у розных адзінках вымярэння. Упершыню ўстаноўлены аптымальны метад класіфікацыі вучэбных тэкставых матэрыялаў на прадмет іх цяжкасці і чытэльнасці на аснове выкарыстання метадаў навучэння з настаўнікам і без яго, што дазволіла прапанаваць метад дыскрымінантнага аналізу для фармальнай дыягностыкі тэксту. Найбольш устойлівае рашэнне класіфікацыі вучэбных тэкставых матэрыялаў атрымана на аснове пераапрацоўкі статыстычных параметраў тэксту пры дапамозе лагарыфмічнай функцыі. Праграмны сродак «MAZI» з'яўляецца інструментам ацэнкі і прыняцця рашэнняў пры апрацоўцы вучэбных тэкставых матэрыялаў на прадмет іх цяжкасці і чытэльнасці.

Рэкамендацыі па выкарыстанні: праграмны сродак «MAZI» укаранёны на прадпрыемствах: РВЦ ДУА «Рэспубліканскі інстытут вышэйшай школы», «Выдавецкі цэнтр ДЗІУСТ БДУ», «Выдавецкі цэнтр БДУ». Можа быць выкарыстаны выдавецтвамі, а таксама аўтарамі для ацэнкі вучэбнага матэрыялу.

Галіна выкарыстання: выдавецтвы, аўтарскія калектывы.

РЕЗЮМЕ

РЫЖАНКОВА Анастасия Сергеевна

МЕТОДИКА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ОБРАБОТКЕ И ОЦЕНКЕ ТЕКСТОВЫХ УЧЕБНЫХ МАТЕРИАЛОВ

Ключевые слова: устойчивость, учебные текстовые материалы, сложность и эффективность, статистические характеристики текста, многомерный статистический анализ, распознавание образов, принятие решений.

Цель исследования: анализ методики оценки учебных текстовых материалов и поиск средств поддержки принятия решений при их обработке.

Методы исследования: метод балльных оценок, метод дополнений, метод парных сравнений; кластерный, регрессионный, дискриминантный, факторный, дисперсионный анализ и метод главных компонент; классификация на основе деревьев решений и искусственных нейронных сетей; методы эталонов, ближайших соседей, меры l, кратчайшего незамкнутого пути и корреляционных плеяд.

Полученные результаты и их новизна. Обоснован выбор объема текстового фрагмента в размере 1800−2000 символов, который позволяет описать статистическую структуру текста; формальная структура текста состоит из трех классов, в которые входит класс, отвечающий за общую длину предложения в той или иной системе измерения; класс, отвечающий за дифференциальную структуру текста; класс, отвечающий за оценку отдельных слов в различных единицах измерения. Впервые установлен оптимальный метод классификации учебных текстовых материалов на предмет их трудности и удобочитаемости на основе использования методов обучения с учителем и без него, что позволило предложить метод дискриминантного анализа для формальной диагностики текста. Наиболее устойчивое решение классификации учебных текстовых материалов получено на основе преобразования статистических параметров текста при помощи логарифмической функции. Программное средство «MAZI» является инструментом оценки и принятия решений при обработке текстовых учебных материалов на предмет их трудности и удобочитаемости.

Рекомендации по использованию. Программное средство «MAZI» внедрено на предприятиях: РИЦ ГУО «Республиканский институт высшей школы», «Издательский центр ГИУСТ БГУ», «Издательский центр БГУ».

Сфера применения: издательства, авторские коллективы.

SUMMARY

RYZHANKOVA Nastassia Sergeevna

METHOD OF SUPPORT DECISION-MAKING IN PROCESSING OF ASSESSMENT EDUCATIONAL TEXT MATERIALS

Keywords: stability, educational texts, complexity and efficiency, information text characteristics, multivariate statistical analysis, pattern recognition, decision-making.

Objective: Analysis methodology for assessing training materials and text search instruments of decision support during their processing.

Methods: scores, additions technique, the method of paired comparisons; clustering, regression, discriminant, factor, analysis of variance and principal components method; classification based on decision trees and artificial neural networks; methods of measurement standards nearest neighbors, measures *l*, the shortest of the open path and correlation of the pleiades.

Obtained results and their novelty. The choice of the amount of text in the amount of 1800−2000 characters, which allows us to describe the information structure of the text; the formal structure of the text is divided into three classes, which include a class that is responsible for the overall length of sentence in a particular system of measurement; class responsible for the differential structure of the text; class responsible for the evaluation of individual words in different units. First established the optimal method of classification learning of text materials for their difficulties and readability through the use of teaching methods with the teacher and without that allowed us to propose a method of discriminant analysis for a formal diagnosis text. The most sustainable solution to the classification of educational text material obtained through changes in the statistical parameters of the text using the logarithmic function. Software tool «MAZI» a tool for evaluating and decision-making in the processing of learning of text materials on premet their difficulties and readability.

Recommendations for application. Software tool «MAZI» implemented in enterprises: «National Institute for Higher Education», «Publishing Center SIMST BSU», «Publishing Center BSU».

Scope: publishers, authors collectives.

Научное издание

Рыжанкова Анастасия Сергеевна

МЕТОДИКА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ОБРАБОТКЕ И ОЦЕНКЕ ТЕКСТОВЫХ УЧЕБНЫХ МАТЕРИАЛОВ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук
по специальности 05.13.01 – Системный анализ,
управление и обработка информации

Ответственный за выпуск А. С. Рыжанкова

Подписано в печать 23.03.2015. Формат 60×84 1/16. Бумага офсетная.
Гарнитура Times New Roman. Печать офсетная.
Усл.-печ. л. 1,5. Уч.-изд. л. 1,5.
Тираж 60 экз. Заказ __

Издатель и полиграфическое исполнение:
УО «Белорусский государственный технологический университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/227 от 20.03.2014.
ЛП № 02330/12 от 30.12.2013.
Ул. Свердлова, 13а, 220006, г. Минск.