ФОРМУЛИРОВКА РЕШАЮЩИХ ПРАВИЛ ОЦЕНКИ ТЕКСТОВОЙ УЧЕБНОЙ ИНФОРМАЦИИ МЕТОДОМ СТАТИСТИЧЕСКОГО АНАЛИЗА

А. С. Рыжанкова
Кафедра редакционно-издательских технологий, Белорусский государственный технологический университет
Минск, Республика Беларусь
E-mail: asmalyk@rambler.ru

В данной работе рассматривается возможность применения методов статистической обработки данных при анализе текстовых учебных материалов по специальной дисциплине для построения устойчивых решающих правил классификации материала на "сложный/легкий". В качестве обучающей выборки использованы результаты опроса, обработанные и преобразованные в классы в зависимости от специфики проведения каждого из них. Результаты исследования представляют собой программное средство, построенное на сформулированном алгоритме обработки текстовой информации

Введение

Современное развитие информационных технологий позволяет проводить качественный анализ текстов всех уровней сложности: распознавание, атрибуция, проверка орфографии, перевод, обработка терминологии, построение конкордансов и т.д. Одним из весьма перспективных и актуальных направлений является анализ читабельности учебных текстов. Это объясняется спецификой их построения, наличием определенных правил представления материала, а также их особой направленностью, заключающейся в доступном изложении новой информации с целью ее последующего воспроизведения и применения.

Исследования по анализу читабельности учебных материалов чаще всего проводят опытно-статистическим путем, заключающемся в получении по результатам опроса определенных качественных характеристик, их переводе в количественную форму с последующей обработкой, а также построении математической модели, описывающей дискриминацию объектов по тому или иному признаку.

Использование обучающих выборок в таких исследованиях объясняется тем, что достоверную автоматизированную оценку учебного материала можно получить только на основе учета экспертных оценок, а не только прямом анализе структуры текста. Известные формулы читабельности построены именно на таком подходе и позволяют получить достоверную оценку сложности материала.

Основная особенность данной работы заключается в том, что впервые получены устойчивые классификационные правила оценки учебного текстового материала по специальности, а также разработано программное средство, позволяющее проводить расширенный статистический анализ текстовой информации и на основе этих данных определять уровень сложности всего издания в целом.

I. Содержание документа

При построении модели были выполнены следующие этапы исследования: 1. Определение минимального объема текстового фрагмента, достаточного для описания статистической структуры издания в целом.

По результатам дисперсионного анализа 378 фрагментов объемом 5002000 символов установлено, что объем текстовой информации, при котором ее статистические показатели находятся на относительно однородном уровне, составляет 18002000 символов. F-проверка результатов на равноточность и совместимость при использовании текстовых фрагментов объемом 30 000 символов и более, анализ показателей квадратичного отклонения разностей s и максимальной погрешности этой разности подтвердили сформулированные выводы.

2. Экспериментальная часть.

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. По результатам эксперимента 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.

3. Статистический анализ текстовых учебных материалов.

Для формирования факторного пространства статистической структуры анализируемых текстовых учебных материалов определены 14 существенных параметров: N1 средняя длина слов в слогах; N2 средняя длина слов в буквах; N3 средняя длина слов по Деверу; N4 средняя длина слов в 3 слога и более; N5 средняя длина слов в 4 слога и более; N6 средняя длина слов в 5 слогов и более; N7 средняя длина слов в 6 слогов и более; N8 средняя длина слов в 7 слогов и

более; N9 процент односложных слов; N10 средняя длина предложения в словах; N11 средняя длина предложения в слогах; N12 процент чисел от общего количества слов; N13 – отношение показателя N4 к N7; N 14 – N5 к N7.

4. Распознавание объектов с использованием обучающей выборки и без нее.

В качестве методов распознавания были определены: кластерный анализ, факторный анализ, метод корреляционных плеяд, главных компонент, множественного регрессионного анализа, искусственных нейронных сетей, эталонов, трех и пяти ближайших соседей, меры l, деревьев решений и дискриминантный анализ. Анализ распознавания объектов методом дискриминантного анализа при n=5 показал наилучшие результаты. 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».

5. Анализ устойчивости полученных решающих правил.

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. В качестве меры устойчивости определен коэффициент вариации, рассчитанный для каждого из представленных способов преобразования. Наименьшие значения коэффициентов вариации наблюдаются при использовании преобразования с помощью десятичного логарифма.

6. Разработка программного средства на основе построенной модели оценки качества текстовых учебных материалов.

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. Таким образом, в результате опытно-статистического анализа внутренней структуры текста учебного издания были построены и сформулированы устойчивые решающие правила, позволяющие оценить трудность учебного материала с позиций обучающихся. Примененные методы анализа данных позволили вывести уровень обработки текста на новый уровень и описать результат процесса восприятия текста математическим языком.