УДК 004.93'14

АВТОМАТИЗИРОВАННЫЙ КОНТРОЛЬ КАЧЕСТВА ИЗДАНИЙ ПРИ ПОДГОТОВКЕ К ПЕЧАТИ

М.А. Зильберглейт[1], Ю.Ф. Шпаковский[2], М.М. Невдах[3]
[1]Институт общей и неорганической химии НАН Беларуси, Минск;
[2]Белорусский государственный технологический университет, Минск;
[3]ИООО «ЭПАМ Системз», Минск, Беларусь

Приведены результаты исследования, посвященного автоматизированному контролю качества учебных материалов на допечатной стадии полиграфического производства. Проведены эксперименты с использованием различных методик для получения объективных критериев качества текстов, выделены и вычислены значения 49 параметров издательских оригиналов. Снижение признакового пространства осуществлено методами многомерного статистического анализа. Для разработки решающего правила использован дискриминантный анализ. Для автоматизированного контроля качества материалов создана программа «Readability analysis».

Введение

В последнее время постоянно повышается уровень управления различными этапами полиграфического производства, однако допечатные процессы, ввиду своей сложности и разнообразия, не позволяют осуществлять эффективное решение задач планирования, управления, моделирования и автоматизации полиграфического процесса. Между тем особую актуальность приобретает проблема создания адекватных экономических, технических, математических и других формализованных моделей на допечатной стадии полиграфического производства.

Как известно, задачей допечатных процессов является обеспечение надлежащего качества подготовки издательской продукции к полиграфическому воспроизведению. С этой целью в Беларуси принят ряд нормативных документов, в частности, различные стандарты (СТБ 1339-2002 «Учебники и учебные пособия для системы высшего и среднего специального образования. Общие технические условия», СТБ 1021-2004 «Издания учебные для общего среднего образования. Общие технические условия», ГОСТ 7.89-2005 «Оригиналы текстовые авторские и издательские. Общие требования», СТБ 7.204-2006 «Издания книжные. Общие технические условия», СТБ 7.206-2006 «Издания книжные и журнальные для детей. Общие технические условия» и др.), но основное внимание в них уделено общим техническим требованиям. Поэтому на сегодняшний день актуальна разработка надежного контроля качества материалов на допечатной стадии полиграфического производства.

Результаты исследования

Следует отметить, что решения некоторых задач (атрибуция текста, оценка близости и однородности стилей, их классификация, читабельность) отражены в работах отечественных и зарубежных исследователей: О.Н. Гринбаума, Г.Я. Мартыненко, В.В. Поддубного, О.Г. Шевелева, М.С. Мацковского, Р.Г. Пиотровского, Д.В. Хмелева, Р. Флеша, Дж. Чолл и др. [1–10]. Кроме того, для решения конкретных прикладных задач разработано программное обеспечение: «ЛингвоАнализатор», «СМАЛТ», «ЛинДа», «PolyAnalyst», «DICTUM», «ВААЛ» и т. д. Однако специальных исследований, нацс-

ленных на разработку автоматизированного контроля качества издательской продукции, не проводилось. Не сформулированы и основные подходы к разработке методики ее использования.

В связи с этим была определена *цель исследования* – разработка модели контроля качества изданий на допечатной стадии полиграфического производства. Для ее реализации были поставлены и решены следующие задачи:

– провести классификацию, анализ и отбор методов проверки качества печатного материала на допечатной стадии полиграфического производства, установить объективные показатели качества исследуемых объектов выборки;

– определить и измерить количественные характеристики исходного материала и экспериментальным путем выявить их связь с качеством; снизить размерность признакового пространства методами многомерного статистического анализа и выявить наиболее информативные показатели;

– на основе экспертных данных и информативных признаков с помощью дискриминантного анализа разработать модель процесса в виде решающего правила для контроля качества материалов;

– разработать специализированное программное обеспечение для оценки качества авторского оригинала и управления качеством издательских проектов.

Конечной целью доклада выступают получение решающего правила в виде набора дискриминантных (классифицирующих) функций для автоматизированного контроля качества изданий и разработка специализированного программного обеспечения для внедрения в существующий технологический процесс издательских предприятий.

На *первом этапе* исследования были проведены эксперименты с использованием различных методик. В экспертизе участвовало 75 реципиентов, что позволило с вероятностью 99 % получить относительную ошибку в долях среднеквадратичного отклонения, равную 0,3. Экспериментальными материалами послужили издательские оригиналы для вузов по философии (первая выборка) и экономической теории (вторая выборка). Каждая выборка содержала 24 отрывка издательского оригинала. При этом наиболее надежными выступили методы: методика дополнения и метод балльных оценок.

Впервые для оценки качества учебного материала для вузов использовался метод парных сравнений. Для выявления связи между мнениями экспертов в последнем методе рассчитывался коэффициент конкордации. Оценка его значимости осуществлялась на основе χ^2-критерия Пирсона.

Обработка и анализ результатов экспериментов позволили выявить информацию относительно качества материалов. На основании полученных данных найдены пять объективных показателей качества: процент правильно заполненных пропусков (Y_1), относительное время работы с текстом (Y_2) – с использованием методики дополнения, средняя оценка качества текста (Y_3), относительное время работы с текстом (Y_4) – с использованием балльных оценок, ранг текста (Y_5).

Для каждого показателя была найдена середина диапазона всех полученных значений. В соответствии с ней производилось разбиение объектов на два класса, которые условно названы «материал требует доработки – материал не требует доработки». В итоге было получено разбиение текстов на группы по выделенным пяти показателям качества.

Второй этап исследования посвящен изучению информационных характеристик исследуемых объектов и выявлению объективных диагностических показателей, которые в наибольшей степени влияют на качество учебных материалов.

С этой целью выделены и вычислены значения 49 параметров учебных текстов (длина текста в абзацах, длина текста в словах, длина текста в буквах, средняя длина

абзаца в фразах, средняя длина абзаца в словах, средняя длина абзаца в буквах, средняя длина абзаца в печатных знаках, средняя длина предложения в фразах и др.). Очевидно, что использование большого количества показателей неэффективно по ряду причин: сильная взаимосвязанность признаков; неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); возможность агрегирования по некоторым признакам. Для снижения признакового пространства применялись методы многомерного статистического анализа (кластерный и факторный анализ, метод корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование).

В кластерном анализе для нахождения расстояния между объектами были приняты следующие меры сходства: евклидово расстояние, квадрат расстояния Евклида, косинус угла, коэффициент корреляции Пирсона, неравенство Чебышева, расстояние Минковского. Для кластеризации использовались методы: простого среднего, группового среднего, ближнего соседа, дальнего соседа, невзвешенный центроидный, взвешенный центроидный и метод Варда.

В результате кластеризации выделенных характеристик печатного материала была получена информация о формировании кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик объекта к тому или иному кластеру.

Кластерный анализ позволил выделить шесть групп для первой выборки и девять – для второй.

Для снижения признакового пространства в работе использовались следующие варианты факторного анализа: метод главных факторов, центроидный метод и метод главных компонент.

При проведении факторного анализа было установлено, что первые три фактора объясняют около 74 % разброса дисперсии для первой выборки, около 64 % – для второй. Для решения вопроса, какие из факторов следует оставить для дальнейшей обработки, рассматривались критерий Кайзера и критерий «каменистой осыпи» Р. Кэтелла.

С целью получения более простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору, в работе нашли применение ортогональные методы вращения: варимакс, квартимакс и эквимакс.

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между факторами. Анализ показал, что факторы по всем методам вращения для двух выборок практически идентичны.

Результаты, полученные методами факторного анализа, позволили выделить шесть групп для первой выборки и восемь групп – для второй.

Наряду с методами факторного анализа для снижения признакового пространства использовался метод корреляционных плеяд. Упорядочение производилось на основании принципа максимального корреляционного пути. Для удобства построения графа были составлены упорядоченные корреляционные матрицы. Опираясь на упорядочения всех признаков, были построены графы, представляющие собой кратчайший незамкнутый путь. После выбора порогового значения коэффициента корреляции исходный граф распался на пять подграфов (групп близких параметров) для первой выборки и шесть подграфов – для второй.

В методе вроцлавской таксономии (методе дендритов) точки многомерного пространства проецировались на плоскость, чем достигалось нелинейное упорядочение изучаемых элементов. Из дендрита, построенного на единицах разбиваемого множества, удалялось n–1 самых длинных связей. Тем самым получалось разбиение денд-

рита на n частей, которое характеризовалось минимальной суммой образующих их отрезков, а полученные подмножества включали элементы с близкими значениями признаков.

С целью построения дендрита вычислены матрицы расстояний (на основе расстояния Евклида) между изучаемыми характеристиками. Данный метод позволил выделить семь групп для первой выборки и пять групп – для второй.

При многомерном шкалировании матрица различий между объектами представлялась в пространстве относительно небольшого числа измерений с наименьшим возможным искажением геометрической структуры исходных данных. Качество метода характеризовалось величиной коэффициентов стресса и R^2.

Наилучшей моделью для первой выборки (stress = 0,210, R^2 = 0,856) стала модель, полученная с использованием меры сходства, основанной на неравенстве Чебышева; для второй выборки (stress = 0,230; R^2 = 0,763) – модель, полученная с использованием квадрата расстояния Евклида. На их основе было получено семь групп признаков для первой выборки и шесть групп – для второй.

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.

Таким образом, впервые в области автоматизированного контроля качества учебных материалов методами многомерного статистического анализа установлены диагностические признаки, т. е. те показатели исследуемых объектов, которые в наибольшей степени влияют на качество. Ими оказались следующие признаки:

– длина слов и предложений (кластерный анализ);

– сложность предложений, число предикативных ядер, длина фразы (факторный анализ);

– разнообразие словаря, длина абзаца, слов и предложений, средняя длина фразы и предложения (метод корреляционных плеяд);

– длина слов и предложений (методы вроцлавской таксономии и многомерного шкалирования).

Для дальнейшего изучения характеристик учебных текстов важнейшей задачей выступает выделение наиболее информативного признака из каждой полученной группы. В данной работе для оценки информативности признаков в качестве информационной использовалась мера С. Кульбака. На основе данной меры 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.

На *третьем этапе* на основе диагностических признаков и экспертных данных был проведен дискриминантный анализ, который позволил разработать решающее правило для автоматической проверки качества материалов. Точность классификации объектов первой выборки составила 91,7 %, второй – 83,3 %.

С помощью дискриминантного анализа для объектов первой выборки для дальнейшей программной реализации получены следующие дискриминантные функции:

$$F_1 = -53{,}062 - 0{,}015X_3 + 0{,}831X_9 - 15{,}106X_{24};$$

$$F_2 = -42{,}720 - 0{,}011X_3 + 0{,}554X_9 - 8{,}663X_{24},$$

где X_3 – длина текста в буквах; X_9 – средняя длина предложения в словах; X_{24} – средняя длина слов в печатных знаках.

Для объектов второй выборки получены следующие дискриминантные функции:

$$F_1 = -123{,}728 - 0{,}165X_5 + 0{,}268X_{10} - 3{,}100X_{39};$$
$$F_2 = -104{,}608 - 0{,}100X_5 + 0{,}2229X_{10} - 2{,}830X_{39},$$

где X_5 – средняя длина абзаца в словах; X_{10} – средняя длина предложения в слогах; X_{39} – процент неповторяющихся слов.

Очевидно, что в зависимости от происхождения изучаемой выборки факторы, влияющие на качество, различны. Дискриминантный анализ подтвердил этот факт и позволил выявить некоторые особенности изучаемых объектов, которые в обязательном порядке следует учитывать на подготовительной стадии издательско-полиграфического процесса. Для получения конкретного результата относительно качества учебных материалов решающее правило следует реализовать на программном уровне.

В соответствии с требованиями к программным средствам был разработан программный продукт «Readability analysis», предназначенный для автоматизированного контроля качества материалов подготовительной стадии.

На *заключительном этапе* проведена верификация с использованием 16 текстовых объектов (издательских оригиналов). Был проведен эксперимент на основе метода балльных оценок, а после этого произведен контроль качества выбранных объектов с помощью разработанной программы. Точность результатов – 94 %.

Положительный эффект от использования данной модели в технологии допечатных процессов обусловлен, прежде всего, расширением экспертных методов контроля качества наиболее ранней стадии издательско-полиграфического производства.

Заключение

Основные научные результаты проведенного исследования можно изложить следующим образом:

1) предложен новый научно обоснованный подход для автоматизированного контроля качества материалов на допечатной стадии полиграфического производства на основе отбора наиболее информативных признаков и разработки модели контроля качества в виде решающего правила;

2) на основе методов многомерного статистического анализа (кластерного анализа, факторного анализа, метода корреляционных плеяд, метода вроцлавской таксономии, многомерного шкалирования) выделены относительно однородные группы взаимосвязанных признаков и выявлена связь между данными группами и особенностями изучаемых объектов;

3) впервые выявлены объективные диагностические показатели, которые оказывают наибольшее влияние на качество учебных материалов на подготовительной стадии издательско-полиграфического производства;

4) впервые разработана модель контроля качества учебных материалов на подготовительной стадии издательско-полиграфического производства в виде решающего правила, что позволяет осуществить автоматизированную проверку качества на основе дискриминантного анализа. Точность классификации объектов составляет 83–92 %. Разработанный в исследовании алгоритм в виде программного продукта интегрирован в редакционный процесс издательско-полиграфических предприятий для повышения эффективности процесса.

Список литературы

1. Гринбаум, О.Н. Проект «ЛИНДА» – автоматизированная система обработки лингвостатистических данных / О.Н. Гринбаум, Г.Я. Мартыненко, С.Я. Фитиалов // Прикладная лингвистика и автоматический анализ текста. – Тарту : ТГУб, 1988. – С. 31–33.

2. Мартыненко, Г.Я. Основы стилеметрии / Г.Я. Мартыненко. – Л. : Ленинградский ун-т, 1988. – 176 с.

3. Мацковский, М.С. Проблемы читабельности печатного материала / М.С. Мацковский // Смысловое восприятие речевого сообщения (в условиях массовой коммуникации). – М. : Наука, 1976. – С. 126–142.

4. Пиотровский, Р.Г. Текст, машина, человек / Р.Г. Пиотровский. – Л. : Наука, 1975. – 327 с.

5. Поддубный, В.В. Сравнительный анализ эффективности алгоритмов распознавания авторства текстов по частотам переходов / В.В. Поддубный, О.Г. Шевелев, А.А. Фатыхов // Вестник Томского гос. ун-та. – 2006. – № 290. – С. 232–234.

6. Хмелев, Д.В. Распознавание автора текста с использованием цепей А.А. Маркова / Д.В. Хмелев // Вестник Моск. ун-та. – Сер. 9. Филология. – 2000. – № 2. – С. 115–126.

7. Хмелев, Д.В. Сложностной подход к задаче определения авторства текста / Д.В. Хмелев // Русский язык: исторические судьбы и современность : труды и материалы Междунар. конгр. (Москва, 13–16 марта 2001 г.). – М. : МГУ, 2001. – С. 426–427.

8. Шевелев, О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. … канд. техн. наук / О.Г. Шевелев. – Томск : ТГУ, 2006. – 20 с.

9. Chall, J.S. Readability: an appraisal of research and application / J.S. Chall // Bureau of educational research monographs. – Columbus : Ohio State University Press, 1958. – № 34. – P. 58–68.

10. Flesch, R. Estimating the comprehension difficulty of magazine articles / R. Flesch // Journal of general psychology. – 1943. – № 28. – P. 63–80.