На правах рукописи

Шевелев Олег Геннадьевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМОВ СРАВНЕНИЯ СТИЛЕЙ ТЕКСТОВЫХ ПРОИЗВЕДЕНИЙ

АВТОРЕФЕРАТ
диссертации на соискание учёной степени кандидата технических наук по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Томск – 2006

Работа выполнена в Томском государственном университете на кафедре прикладной информатики факультета информатики

Научный руководитель: доктор технических наук, профессор Поддубный В.В.

Официальные оппоненты: доктор технических наук, профессор Матросова А.Ю.

кандидат физико-математических наук, доцент Новосельцев В.Б.

Ведущая организация – Московский государственный университет.

Защита состоится 20 апреля 2006 г. в 10-30 на заседании диссертационного совета Д 212.267.08 в Томском государственном университете по адресу: г. Томск, пр. Ленина 36, корп. 2, ауд. 102.

С диссертацией можно ознакомиться в научной библиотеке Томского государственного университета.

Отзывы на автореферат (2 экз.), заверенные печатью, высылать по адресу: 634050, г. Томск, пр. Ленина, 36, ученому секретарю ТГУ.

Автореферат разослан 10 марта 2006 г.

Ученый секретарь
диссертационного совета,
доктор технических наук, доцент Скворцов А.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы

Проблема сравнения стилей текстовых произведений является актуальной в целом ряде областей человеческой деятельности. В исторических исследованиях сравнивать стили приходится для того, чтобы определить время написания того или иного исторического документа или установить личность его автора. В филологических дисциплинах – для изучения стилистических особенностей текстов или языка произведений различных жанров, авторов и т.д. В психологии и психиатрии сравнение стилей актуально при изучении мыслительной деятельности, тестировании и диагностике авторов текстов. Многие практические задачи сравнения стилей текстов возникают в криминалистике, например, для установления личности автора письменной угрозы или определения индивидуальных особенностей автора при проведении оперативно-розыскных мероприятий.

Количественные подходы к решению данных проблем в настоящее время особенно актуальны, так как они позволяют автоматизировать процедуру сравнения стилей текстов, дать формализованное объективное решение. Развитие этих подходов важно также и для информатики, поскольку с их помощью можно улучшить качество классификации и упорядочивания текстовых коллекций, что чрезвычайно актуально для поисковых систем и крупных хранилищ текстовых данных.

Сравнение стилей текстов проводится, как правило, на основе совокупности ряда признаков, отражающих свойства стилей текстов. Обычно рассматриваются частотные признаки (частоты появления определенных слов, буквосочетаний и др.), которые могут быть легко формализованы для проведения с их помощью количественного (частотного) анализа текстов.

На базе сравнения стилей текстовых произведений решаются три основные задачи: 1) проверка текстов на близость стилей или однородность по стилю, 2) кластеризация и 3) классификация текстов.

Проверкой текстов на близость стилей впервые занимались, в частности, Mendenhall T.C., Морозов Н.А., Фоменко Т.Г. и Фоменко В.П. Серьезный вклад в исследования по проверке однородности текстов внесли Morton A.Q., Ashford T., Farringdon J.M., Ковалевский А.П. и др.

В рамках задачи кластеризации текстов применяются различные методы кластеризации (метод k-средних, метод ближайшего соседа, нейронные сети SOM и др.), а также их модификации. Иерархические методы кластеризации использовали в своих работах Leouski A.V., Croft W.B., Karger D.R., Pedersen J.O., Tukey J.W., Tantrum J., Murua A.,

Stuetzle W. Неиерархические методы кластеризации текстов исследовали Zhong S., Gosh J., Steinbach M., Karypis G., Kumar V. и др.

Наибольшее число работ в области сравнения стилей текстов посвящено задаче классификации текстов. Среди методов классификации рассматриваются нейронные сети (Matthews R., Merriam T., Kjell B., Tweedie F.J., Singh S., Holmes D.I., Lowe D., Matthews R.), метод опорных векторов (de Vel O., Joachims T., Diederich J. J.), дискриминантный анализ (Baayen H., Tweedie F., Patton J.M., Can F.A, Peng R.D., Hengartner N.W.), метод сжатия данных (Frank E., Chui C., Witten I.H., Teahan W.J., Хмелев Д., Benedetto D.), метод Хмелева Д., методы, основанные на извлечении правил (Apte C., Damerau F., Weiss S., Oakes M., Holden N., Freitas A.A.), и др.

В настоящее время существует ряд программных систем, позволяющих производить разнообразные виды анализа текстов. Наиболее известными среди таких систем являются «Лингвоанализатор» Д. Хмелева, информационная система «СМАЛТ», система «ВААЛ» 9.0, PolyAnalyst 4.6 (с модулем для работы с текстом TextAnalyst), система DICTUM.

Несмотря на множество работ по сравнению стилей текстов, имеется ряд не исследованных или мало исследованных областей. Нет работ по применению мер близости стилей текстов, основанных на точных статистических критериях сравнения частот появления признаков. Недостаточно исследованы зависимости качества классификации различными методами от объемов фрагментов и от числа классов. Нет исследований по сравнению качества классификации по сложным (в т.ч. грамматическим) признакам на размеченном и выверенном текстовом материале. Несмотря на то, что в ряде работ говорится об использовании свойств марковости текста, никто не проводил проверку последовательностей символов текста на марковость. В существующих программных продуктах нет механизма, обеспечивающего возможность задания признаков стилей текстов пользователем (а не выбора признаков из числа предлагаемых разработчиком). Наконец, имеющиеся программы анализа текстов не ориентированы на комплексное исследование и сравнение стилей текстов (для разных задач анализа стилей текстов с использованием различных методов их решения, различных частотных признаков, различного текстового материала и т.д.).

Решению задач, в той или иной степени заполняющих указанные пробелы, посвящена настоящая работа.

Цель работы

Целью работы является разработка алгоритмов и инструментария для сравнения стилей текстовых произведений. В рамках указанной цели поставлены и решены следующие задачи:

1) исследование качества работы ряда существующих методов математической статистики и искусственного интеллекта для сравнения стилей текстовых произведений по частотным признакам, задаваемым пользователем;

2) модификация известных и разработка новых мер сравнения частот для задач кластеризации и классификации текстов;

3) создание языка задания частотных признаков стилей текстовых произведений и его интерпретатора;

4) разработка и реализация программного комплекса для сквозного количественного анализа текстов от их первичной обработки до получения решений.

Методика исследований

Для решения задач, обеспечивающих достижение поставленной цели, использовались методы математической статистики, искусственного интеллекта, а также методы объектно-ориентированного программирования.

Научная новизна работы,

1. Предложены новые подходы для сравнения стилей текстов с использованием гипергеометрического критерия (двустороннего точного критерия Фишера) и критерия хи-квадрат по отдельным частотным признакам текстов, совокупности признаков, а также по их распределению.

2. Предложен новый подход к кластеризации текстов с использованием ранее не применявшихся в области обработки текстов таких мер сходства, как «частота рассогласования» (сложный признак) и интегральная мера рассогласования (совокупность признаков), получаемых на основе проверки гипотез о сходстве стилей текстов по частотным признакам.

3. Предложены модификации известного метода Хмелева классификации текстов по авторскому стилю с использованием для оценки расхождения частот мер Кульбака и хи-квадрат, а также модульных мер. Показано, что мера Хмелева является частным случаем меры Кульбака.

4. Доказана несостоятельность гипотезы о том, что последовательность символов текста обладает свойствами простой цепи Маркова.

5. Разработан оригинальный язык задания частотных признаков, позволяющий декларировать признаки и представлять их в виде шаблонов, пригодных для автоматического преобразования текстов к набору частот.

Практическая ценность работы

Разработанный программный комплекс «СтилеАнализатор» для анализа стилей текстов, обеспечивающий полный цикл проведения количественного анализа текстов, включающий предварительную обработку текстов, извлечение частотных признаков, их обработку и представление результатов в наглядном для человека виде, может быть широко использован специалистами в различных областях знаний (лингвистами, филологами, криминалистами, историками).

Положения, выносимые на защиту

1. Новые подходы для сравнения стилей текстов с использованием гипергеометрического критерия и критерия хи-квадрат по отдельным частотным признакам текстов, совокупности признаков, а также по их распределению.
2. Новый подход к кластеризации текстов на основе проверки гипотез о равенстве частотных признаков стилей текстов с использованием таких мер сходства, как «частота рассогласования» и интегральная мера рассогласования.
3. Модификации известного метода Хмелева с использованием для оценки расхождения частот мер Кульбака и хи-квадрат, а также модульных мер.
4. Доказательство несостоятельности гипотезы о том, что последовательность символов текста обладает свойствами простой цепи Маркова.
5. Язык задания частотных признаков стилей текстов.
6. Программный комплекс «СтилеАнализатор» для анализа стилей текстов.

Внедрение полученных результатов

Реализованный программный комплекс внедрен в лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ.

Апробация работы

Результаты работы докладывались и обсуждались на следующих конференциях:

1. IV Межвузовская конференция студентов аспирантов и молодых ученых «Наука и образование», Томск, 2000.

2. V Общероссийская межвузовская конференция студентов, аспирантов и молодых ученных «Наука и образование», Томск, апрель 2001 г.

3. Нейроинформатика и ее приложения: XII Всероссийской семинар, Красноярск, октябрь 2004 г.

4. Информационные технологии и математическое моделирование: III Всероссийская научно-практическая конференция, Анжеро-Судженск, декабрь 2004 г.

5. XLIII Международная научная студенческая конференция «Студент и научно-технический прогресс»: Информационные технологии, Новосибирск, апрель 2005 г.

6. XI Международная научно-практическая конференция студентов и молодых ученых «Современные техника и технологии СТТ'2005», Томск, марта – апрель 2005 г.

7. IX Международная конференция студентов, аспирантов и молодых ученых «Наука и образование», Томск, апрель 2005 г.

8. Всероссийская научная конференция Квантитативная лингвистика: исследования и модели (КЛИМ - 2005), Новосибирск, июнь 2005 г.

9. Информационные технологии и математическое моделирование: IV Всероссийская научно-практическая конференция, Анжеро-Судженск, ноябрь 2005 г.

Структура диссертации

Диссертация состоит из введения, основного текста, заключения, библиографического списка (135 наименований), и 5 приложений. Основной текст состоит из 3 глав и содержит 37 таблиц и 36 рисунков. Общий объем работы 176 страниц, включая 12 страниц приложений.

Публикации по теме работы

Основное содержание работы отражено в 16 публикациях, в т.ч. в 11 статьях.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе диссертации проводится аналитический обзор имеющихся методов и программ количественного анализа текстов. Первые три пункта обзора касаются основных классов задач количественного анализа текстов – проверки текстов на близость стилей или однородность по стилю, кластеризации текстов и классификации текстов. Рассматриваются методы решения этих задач и исследования, в которых используются данные методы. Для исследований по возможности даются условия экспериментов: число классов, наборы признаков, используемый текстовый материал, приводятся результаты. В чет-

вертом пункте говорится о программных продуктах. В пятом – ставятся задачи исследования и разработок диссертации на основе выявленных не решенных проблем и их классификации.

Вторая глава посвящена алгоритмам и проведенным в работе исследованиям в области сравнения, кластеризации и классификации стилей текстов.

В п. 2.1 предлагаются подходы к сравнению стилей текстов на основе сравнения частот и распределений частот появления признаков стилей по статистическим критериям, предлагаются меры кластеризации, использующие результаты применения этих подходов.

Для сравнения стилей двух текстов по одному частотному признаку стиля предлагается использовать гипергеометрический критерий. В ходе сравнения стилей по данному критерию проверяется нулевая гипотеза о том, что тексты имеют одинаковый стиль по данному признаку, против альтернативы – тексты различаются по стилю. Достигнутый уровень значимости критерия рассчитывается по следующей формуле:

$$p_0 = \sum_{x=\max(0,s-n_2)}^{\min(n_1,s)} \left\{ h(x \mid s, n_1, n_2) \le h(m_1 \mid s, n_1, n_2) \right\},$$

где $s = m_1 + m_2$, m_1 и m_2 – числа появления признака в первом и втором тексте, n_1 и n_2 – объемы текстов, $h(x \mid s, n_1, n_2) = C_{n_1}^{x} C_{n_2}^{s-x} / C_{n_1+n_2}^{s}$ – гипергеометрическое распределение, $x = \overline{\max(0, s-n_2), \min(n_1, s)}$. Статистикой критерия является наблюдаемое значение *x*, то есть m_1. Решение в пользу альтернативы принимается при значении достигнутого уровня значимости критерия меньше или равном альфа. При значении большем альфа оснований отвергнуть нулевую гипотезу нет.

Аналогичным образом предлагается делать сравнение стилей текстов по отдельным частотным признакам на основе критерия χ^2. Статистика критерия рассчитывается по формуле:

$$\chi^2 = \frac{(m_1 - E_1)^2}{E_1} + \frac{(n_1 - m_1 - E_2)^2}{E_2} + \frac{(m_2 - E_3)^2}{E_3} + \frac{(n_2 - m_2 - E_4)^2}{E_4},$$

где $E_1 = (m_1 + m_2) n_1 / n$, $E_2 = (n - m_1 - m_2) n_1 / n$, $E_3 = (m_1 + m_2) n_2 / n$, $E_4 = (n - m_1 - m_2) n_2 / n$ – ожидаемые значения чисел событий при верной нулевой гипотезе. Достигнутый уровень значимости критерия вычисляется по формуле $p_0 = 1 - F(\chi^2)$, где $F(\chi^2)$ – интегральная функция распределения χ^2 с одной степенью свободы при наблюдаемом значении статистики χ^2.

С помощью критерия χ^2 предлагается также сравнение распределений частот признаков. Частоты признаков при этом должны соответствовать полной группе попарно несовместимых событий. Статистика критерия вычисляется по формуле:

$$\chi^2 = \sum_{i=1}^{N} \left(\left(\frac{n_1 n_2}{m_{1i} + m_{2i}} \right) \cdot \left(\frac{m_{1i}}{n_1} - \frac{m_{2i}}{n_2} \right)^2 \right).$$

При верной нулевой гипотезе данная статистика имеет χ^2-распределение с числом степеней свободы, равным $N - z - 1$, где z – число пар частот, в которых оба значения m_{1i}, m_{2i} равны нулю.

Для проведения кластеризации набора из K текстов по отдельному признаку или распределению признаков предлагается следующий подход. Первоначально производится попарное сравнение всех текстов набора по одному из указанных выше критериев. В результате таких сравнений будет получено K^2 достигнутых уровней значимости $\left\{ p_{0ij}, i, j = \overline{1, K} \right\}$. Эти значения размещаются в матрице, из которой при фиксированном допустимом уровне значимости критерия делается индикаторная матрица $y_{ij} = \left\{ \begin{matrix} 0, p_{0ij} > \alpha, \\ 1, p_{0ij} \le \alpha, \end{matrix} i, j = \overline{1, K} \right\}$, состоящая из нулей и единиц. Нули в ней соответствуют принятию нулевой гипотезы, а единицы – альтернативы. Каждая строка этой матрицы представляет собой K-мерный булев вектор, компоненты которого характеризуют статистическую значимость различия вероятностей появления признака или распределений признаков для всех пар текстов. Далее, на основе меры расстояния строится матрица расстояний. В качестве меры расстояния между двумя текстами предлагается взять частоту несоответствия элементов строк этих текстов, вычисляемую по формуле:

$$r_{ij} = \frac{1}{K} \sum_{k=1}^{K} 1\left(y_{ik} \neq y_{jk} \right)$$

где $1(\cdot)$ – индикатор события, указанного в скобках. Эта мера обладает всеми свойствами расстояния, и на ее основе может быть корректно проведена кластеризация текстов (по расстоянию «частота несоответствия»). На основе матрицы расстояний производится иерархическая кластеризация текстов с помощью метода дальнего соседа. В результате получается дендрограмма кластеризации.

Для кластеризации текстов по L различным признакам (не обязательно соответствующим полной группе попарно несовместимых событий) предлагается похожий подход, за тем исключением, что в ходе сравнений строится столько различных матриц, по скольким признакам производится сравнение, а в качестве меры используется мера интегрального рассогласования, которая вычисляется следующим образом: $r_{ij} = \sum_{l=1}^{L} y_{ij}^{l}$. С использованием данной меры различие стилей пары текстов определяется суммой всех различий (единиц) по всем индикаторным матрицам.

В завершение пункта по подходам к сравнению и кластеризации текстов по стилям в работе приводятся примеры кластеризации текстов по авторству. Показывается, что с помощью данных подходов можно получить группы текстов разной степени близости по стилю (рис. 1).

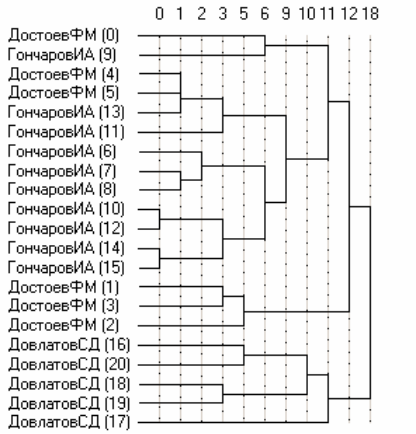

Рис. 1. Пример дендрограммы кластеризации

В п. 2.2 рассматривается использование деревьев решений для классификации текстов. В пп. 2.2.1-2.2.2 приводятся полученные формы алгоритмов построения деревьев решений и отсечения ветвей, удобные для реализации и применения в задачах классификации текстов по частотным признакам. Для построения дерева используется алгоритм C4.5. Для отсечения ветвей – алгоритм, основанный на статистическом тесте независимости, с использованием критерия χ^2.

В п. 2.2.3 приводятся результаты экспериментов по классификации текстов по авторству с помощью деревьев решений. В качестве текстового материала используются наборы художественных текстов 30, 20 и 10 авторов (русские классики и современники, общий объем примерно 50 Мб) и газетные статьи 10 журналистов (10 Мб). Для исследования зависимости качества классификации от объемов фрагментов рассматриваются различные наборы данных, в каждом из которых тексты разбиты на фрагменты определенной длины. Данные получены для трех наборов признаков: частот появления пар букв, частот появления 100 самых часто встречаемых словоформ из частотного словаря Шарова,

частот появления предложений с определенным числом слов. Длины фрагментов берутся в числах тех элементов, в которых задан признак (буквах, словах, предложениях). Всего рассматривается 146 наборов данных.

Качество классификации в ходе каждого эксперимента оценивается по частоте правильно классифицированных фрагментов на тестовой выборке. Каждый набор данных участвует в классификации 10 раз в соответствии с методом *k*-подмножеств. Суть метода состоит в разделении исходных данных на *k* равных частей и запуске (обучении и тестировании) алгоритма *k* раз, причем в ходе каждого запуска (*k*-1) частей участвует в обучении, одна – в тестировании, а тестовая часть постоянно меняется. В работе *k* выбрано равным 10.

Результатами тестирования в работе выступают средняя частота $\overline{p_{ra}}$ правильных классификаций (среднее качество), полученное как среднеарифметическое каждого из *k* запусков алгоритма, и границы 95% интерквантильного интервала, задающие разброс частот. Для нормального распределения (а гипотеза о нормальности для исследуемых данных не отвергается) такие границы определяются интервалом $\overline{p_{ra}} \pm 2\sigma$.

Всего рассматривается 12 графиков зависимостей качества классификации от объемов текстовых фрагментов (3 набора признаков, 4 набора текстовых данных). Основной вывод состоит в том, что качество классификации растет в среднем с увеличением объемов фрагментов, причем на кривой роста четко выделяются две области: область быстрого роста при малых объемах фрагмента (меньше критического) и область практической стабилизации при больших объемах фрагментов (больше критического). Критическое значение объема имеет порядок 30000-40000 символов или 5000-6000 слов, или 400-600 предложений. В целом, по рассмотренным признакам деревья решений довольно плохо классифицируют тексты: даже в самом лучшем случае (10 классов) частота правильных классификаций сильно колеблется и в среднем не превышает 80%.

В п. 2.2.4 приведены графики зависимостей качества классификации от порога отсечения. Рассмотрено 11 различных порогов для разных комбинаций признаков и наборов данных и установлено, что на рассмотренных данных выбранный алгоритм отсечения не дает существенного увеличения частоты правильных классификаций.

В п. 2.2.5 приводятся исследования качества классификации текстов по жанровым типам с помощью деревьев решений. В качестве текстового материала взят грамматически размеченный корпус рус-

ских газет конца XX века. Корпус предоставлен лабораторией общей и компьютерной лексикологии и лексикографии филологического факультета МГУ. Статьи корпуса сгруппированы в работе по жанровым типам, полученные массивы разбиты на фрагменты по 40000 символов и 6000 слов. Всего рассматривается 4 жанровых типа. Классификация производится по 14 наборам признаков: 5 – уровня букв, 5 наборов грамматических признаков и 4 набора словарных признаков. Вывод: деревья решений дают почти одинаковые низкие показатели частот правильных классификаций (примерно 50%±20%) по жанровым типам ядерного корпуса на любом из рассмотренных наборов признаков.

В п. 2.2.6 по тем же наборам признаков и на том же газетном корпусе проведены эксперименты по классификации текстов по 10 источникам газет. С учетом большего числа классов качество классификации с помощью деревьев решений по источникам газет заметно выше качества классификации по жанровым типам.

В п. 2.2.7 рассматривается подход к оценке индивидуальной информативности признаков на основе значения количества информации, используемого при построении деревьев решений. Получены 10 наиболее информативных признаков (в плане различения авторов) из 3-х рассмотренных наборов для разного числа авторов.

В п. 2.3 рассматривается перспективный метод Хмелева для классификации текстов. В ряде работ при упоминании данного метода нередко говорится о последовательности символов текста как о простой цепи Маркова. Но на деле это утверждение никем не проверено. Для его проверки в п. 2.3.1 предлагается два алгоритма проверки гипотезы о марковости последовательности символов текста – с применением критерия χ^2 и критерия доверительных интервалов – для сравнении матриц частот в левой и правой частях уравнения Колмогорова-Чепмена. Показывается, что гипотеза о марковости последовательности символов текста отвергается на уровне значимости, меньшем 5%, при объеме фрагментов 10-15 Кб для критерия χ^2 и 100-200 Кб для метода доверительных интервалов на большом количестве художественных текстов.

В п. 2.3.2 подробно рассматривается метод Хмелева. Ключевой идеей метода Хмелева является подсчет и обработка парных сочетаний элементов текста (например, букв). Обучение алгоритма производится на текстах заданного множества классов. Для каждого класса подсчитывается матрица-эталон появления всех пар рассматриваемых элементов в его текстах. При классификации произвольного текста подсчитывается аналогичная матрица и сравнивается со всеми матрицами-эталонами. Рассматриваемый текст относится к классу с наиболее по-

хожей матрицей-эталоном. Сравнение матриц производится по мере Хмелева. В п. 2.3.2 показывается, что мера Хмелева является частным случаем меры Кульбака, а также предлагаются новые меры сравнения матриц частот появления признаков: безусловная мера Кульбака (далее, просто «мера Кульбака»), мера хи-квадрат и модульные варианты мер Хмелева и Кульбака. Особенностью данных мер является то, что они не используют специфическую информацию о частотах переходов, в отличие от меры Хмелева, поэтому могут работать с произвольными признаками.

В п. 2.3.3 приведены исследования по классификации текстов по авторству на том же материале, признаках и объемах, что и для деревьев решений, но для метода Хмелева и его модификаций. Найдены зависимости качества классификации от объемов фрагментов и числа классов. Показано, что использование метода Хмелева и мер Хмелева, модульной меры Кульбака и меры хи-квадрат дает примерно одинаковые результаты и позволяет с точностью до 100% классифицировать тексты по авторству. Меры же Кульбака (не модульная) и модульная Хмелева работают плохо. Качество классификации так же, как и для деревьев решений, растет в среднем с увеличением объемов фрагментов, но, начиная с критического значения, стабилизируется.

В п. 2.3.4 для метода Хмелева и его модификаций проведены исследования по классификации текстов по жанровым типам и источникам, аналогичные тем, что были проведены для деревьев решений. Выявлено, что меры Хмелева, хи-квадрат и модульная мера Кульбака работают примерно одинаково, мера Кульбака работает плохо. Метод Хмелева позволяет с довольно высоким качеством (75-100%) классифицировать газетные статьи по 4 жанровым типам. Наилучшие результаты классификации на всех наборах признаков достигнуты с использованием меры хи-квадрат. Метод Хмелева и некоторые его модификации позволяют с очень высоким качеством (99-100%) производить классификацию газетных статей по 10 источникам.

В п. 2.4 проведены эксперименты по классификации тех же текстов, что и для деревьев решений и метода Хмелева, по авторству с помощью нейронных сетей с различными параметрами сети. Выявлено, что на рассмотренном текстовом материале использование любого момента инерции или финального значения скорости обучения, меньшего единицы, почти всегда ухудшают качество классификации.

В п. 2.5 приводится сравнение рассмотренных методов классификации. Показывается, что метод Хмелева и его модификации выигрывают как в скорости обучения, так и в качестве классификации. Нейронные сети дают сопоставимое качество, но сильно проигрывают в

скорости. Деревья решений обеспечивают наихудшее качество классификации, но при этом дают наглядный вид решения и по ходу производят отбор самых информативных признаков.

В третьей главе диссертации приводится описание разработанного языка задания частотных признаков, его интерпретатора и разработанной программы анализа стилей текстов «СтилеАнализатор».

В п. 3.1 дается описание языка. В основе разработанного языка лежит взаимодействие с элементами текста, естественными для человека, такими как предложение, буква, слово и их последовательностями. Результаты выполнения запросов языка могут быть представлены в трех видах: таблице частотных данных, списке последовательностей данных, списке подходящих элементов.

Запрос на разработанном языке состоит и набора признаков. Признаки задаются набором однородных элементов. Элементы состоят из содержимого и свойств. В языке имеется три вида элементов: буквосочетания, слова, предложения. Содержимым элементов является последовательность элементов уровнем ниже. Свойства могут задавать общий вид элемента и указывать на позицию элемента в цепочке. Для элемента слова они могут задавать, например, длину слова в буквах, позицию слова в предложении.

Язык позволяет задавать как простые признаки (например, запрос для подсчета числа слов «он» в текстах), так и сложные, состоящие из нескольких элементов (например, запрос для подсчета предложений, включающих определенные слова). С помощью свойства «позиция по отношению к предыдущему элементу» можно задавать признаки появления связанных последовательностей элементов. Можно задавать сразу множество запросов с помощью механизма задания диапазонов (например, одним запросом можно задать сразу 1024 признака появления всех возможных сочетаний пар букв).

Интерпретатор языка реализован на Microsoft Visual C# .Net и является частью «СтилеАнализатора». В п. 3.1.3 дается описание устройства интерпретатора. Основными функциями интерпретатора являются: разбор строк запросов признаков и преобразование этих строк во внутренние структуры программы, просмотр текстов на предмет соответствия их элементов заданным признакам, формирование результатов в нужном для пользователя виде. В пункте описаны внутренние структуры программы, последовательность разбора отдельной строки запроса, процедура просмотра текста и формирования результатов.

В пп. 3.2-3.7 дается описание разработанного программного комплекса «СтилеАнализатор». Процесс исследований в «СтилеАнализаторе» разделен на отдельные этапы. Каждый этап является относи-

тельно независимым, предусматривает различные варианты исполнения, свой набор и формат представления данных, пригодные для использования на других этапах и в других программах. На первом, подготовительном этапе пользователь может сделать предварительную обработку текстов. Второй этап предусматривает извлечение частотных признаков текста или набора текстов. На третьем этапе исследователь, выбрав метод обработки, может привести к нужному виду или проанализировать полученные данные.

В программе можно работать как с обычными текстами, так и с размеченными, так называемыми вертикальными. Вертикальный текст, основная структура которого взята из программы DICTUM, помимо самого текста содержит служебную информацию и дополнительную информацию о словах (нормальную форму слова, грамматические характеристики). В «СтилеАнализаторе» можно просматривать вертикальный текст, редактировать его свойства, размечать диалоги, разбивать файл вертикального текста на отдельные тексты с различным видом группировок, разбивать текст на фрагменты по главам, частям, задавать сложные грамматические признаки на основе информации в вертикальном тексте.

Для извлечения частотных признаков в программе имеется диалоговое окно, где пользователь может выбрать текст или список текстов, добавить, удалить или изменить признак, очистить список, сохранить список признаков в файл и т.д. При добавлении признака открывается форма, которая для удобства задания признаков содержит кнопки-макросы. После запуска подсчета открывается форма, на которой отображается время начала подсчета, текущий обрабатываемый текст, фрагмент, время окончания подсчета.

При извлечении признаков одновременно с созданием таблицы результатов создается описание данных, представленное в виде небольшой базы данных XML. Эта база содержит информацию о текстах и фрагментах, такую как путь к файлу с текстом, имя автора, название, жанровый тип, дату публикации и источник публикации, номер фрагмента в тексте, размеры фрагмента и т.д. Эта информация используется в ходе анализа данных.

В программе имеется функции предобработки таблиц частотных данных и списка последовательностей данных. Можно преобразовывать списки частотных данных в таблицы, фильтровать таблицы по признакам (т.е. столбцам) с учетом информативности, фильтровать по блокам, разделять случайным образом данные на тестовую и обучающую выборки.

Исходными данными для анализа в программе являются таблицы частотных данных. Для проведения анализа имеется форма, в которой собраны все основные методы анализа. В данной форме можно выбрать анализируемые таблицы, задачу анализа, метод для решения данной задачи и его параметры, виды представления результата и т.д.

Задачи анализа разделены на две большие подгруппы: решаемые с помощью одношаговых и многошаговых алгоритмов. На базе одношаговых алгоритмов реализованы задачи сравнения частот появления признаков по различным критериям, иерархическая кластеризация текстов, классификация текстов с помощью деревьев решений, классификация на основе метода Хмелева и его модификаций.

На базе многошаговых алгоритмов реализована классификация текстов на основе нейронных сетей прямого распространения. В программе имеется возможность задания числа слоев и нейронов сети, этапов обучения сети, числа итераций, скорости обучения и т.п.

В заключении подводятся итоги проделанной работы.

1. Предложены новые подходы для сравнения стилей текстов по частотным признакам с использованием гипергеометрического критерия (двустороннего точного критерия Фишера) и критерия хи-квадрат.

2. Предложен новый подход к кластеризации текстов с использованием мер сходства «частота рассогласования» и интегральная мера рассогласования, получаемых на основе проверки гипотез о сходстве стилей текстов по частотным признакам.

3. Проведены исследования зависимости от объемов текстовых фрагментов качества классификации текстов по авторству с помощью деревьев решений. Показано, что качество классификации сначала растет в среднем с увеличением объемов фрагментов, а затем стабилизируется.

4. Проведены исследования по классификации с помощью деревьев решений текстов газетных статей по жанровым типам, источникам и различным наборам признаков. Установлено, что качество классификации с помощью деревьев решений является в целом не высоким, слабо зависит от набора признаков и по источникам несколько выше, чем по жанровым типам.

5. Предложены модификации метода Хмелева с использованием для оценки расхождения частот мер Кульбака и хи-квадрат, а также модульных мер. Показано, что мера Хмелева является частным случаем меры Кульбака.

6. Показано, что последовательность символов текста не обладает свойствами простой цепи Маркова.

7. Проведены исследования качества классификации текстов по авторству с помощью метода Хмелева и его модификаций в зависимости от объемов фрагментов. Показано, что, как и в случае использования деревьев решений, качество классификации сначала растет в среднем с увеличением объемов фрагментов, а затем стабилизируется.

8. Проведены исследования по классификации с помощью метода Хмелева и его модификаций текстов газетных статей по жанровым типам и источникам. Показано, что их применение позволяет с высоким качеством производить классификацию и по жанровым типам (75-100%), и по источникам (99-100%).

9. Произведено сравнение рассмотренных методов классификации текстов. Показано, что нейронные сети и метод Хмелева дают примерно одинаковые высокие показатели качества, а деревья решений – самые низкие показатели качества.

10. Разработан язык задания частотных признаков и реализован интерпретатор с этого языка.

11. Создан программный комплекс «СтилеАнализатор» для анализа стилей текстов.

В приложениях приводится список рассмотренных в работе наборов признаков стилей текстов, список авторов, текстов и объемов текстов из набора 156 художественных произведений, список журналистов, число и общий объем их статей из набора 5697 газетных текстов, использованных в главе 2 диссертации, количественные характеристики реализации программного комплекса «СтилеАнализатор», а также акты о внедрении и использовании разработанного комплекса.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Поддубный В.В., Шевелев О.Г. Кластеризация объектов по мерам сходства частот событий // Обработка данных и управление в сложных системах: Сборник статей / Под ред. А.Ф. Терпугова. – Томск: Изд-во Том. ун-та, 2005. – Вып. 7. – С. 175-185.

2. Поддубный В. В., Шевелев О. Г. О мерах расстояния при кластеризации текстов по частотным признакам // Обозрение прикладной и промышленной математики, 2005, Т. 12. – Вып. 2. – С. 478.

3. Поддубный В.В., Шевелев О.Г. Кластеризация объектов по частотам событий // IV Всероссийская ФАМ конференция (25-27 февраля 2005 г.): Тезисы докладов / Под ред. к.ф.-м.н. Д.В. Семеновой. – Красноярск: Красноярский гос. ун-т, 2005. – С. 67-68.

4. Поддубный В. В., Шевелев О.Г. Образует ли последовательность символов текста простую цепь Маркова? // Информационные технологии и математическое моделирование (18-19 ноября 2005 г.):

Материалы IV Всероссийской научно-практической конференции, Ч. 2. – Томск: Изд-во Том. ун-та, 2005. – С.14-16.

5. Поддубный В.В., Шевелев О.Г. Сравнение и кластерный анализ текстов по частотным признакам на основе гипергеометрического критерия // Квантитативная лингвистика: исследования и модели (КЛИМ – 2005, 6-10 июня 2005 г.): Материалы Всероссийской научной конференции. – Новосибирск: Изд-во НГПУ, 2005. – С. 205-217.

6. Поддубный В.В., Шевелев О.Г. Сравнение стилей текстовых произведений по частотному признаку на основе гипергеометрического критерия // Теоретическая и прикладная информатика: Сборник статей / Под ред. А.Ф. Терпугова. – Томск: Изд-во Том ун-та, 2004, Вып. 1. – С. 101-110.

7. Поддубный В.В., Шевелев О.Г. Сравнительный анализ стилей текстов по частотным признакам на основе гипергеометрического критерия // Информационные технологии и математическое моделирование (11-12 декабря 2004 г.): Материалы III Всероссийской научно-практической конференции, Ч. 2. – Томск: Изд-во Том. ун-та, 2004. – С. 48-51.

8. Тютерев В.В., Шевелев О.Г., Анализ текстов с помощью семантических карт нейронных сетей топографического отображения // Нейроинформатика и ее приложения: Материалы IX Всероссийского семинара / Под ред. А.Н. Горбаня; Отв. за выпуск Г.М. Цибульский. – Красноярск: ИПЦ КГТУ, 2001. – С. 199-200.

9. Тютерев В.В., Шевелев О.Г. Использование нейронных сетей GTM для редукции многомерных пространств // V Общероссийская межвузовская конференция студентов, аспирантов и молодых ученных «Наука и образование» (23-26 апреля 2001 г.): Материалы конференции в 5 т., Т. 1. – Томск: Изд-во Томского государственного педагогического университета, 2003. – С. 188-192.

10. Шевелев О.Г. Анализ частоты встречаемости различных длин предложений в литературном тексте как возможной характеристики авторского стиля с помощью самоорганизующихся карт Кохонена // Нейроинформатика и ее приложения (1-3 октября 2004 г.): Материалы XII Всероссийского семинара / Под ред. А.Н. Горбаня, Е.М. Миркеса. – Красноярск: ИВМ СО РАН, 2004. – С. 177-178.

11. Шевелев О.Г., Бурков Д.В. Предобработка текстов для целей лингвистического анализа. // IX Всероссийская конференция студентов, аспирантов и молодых ученых «Наука и образование» (25-29 апреля 2005 г.): Материалы конференции в 6 т., Т.1, Ч.2: Естественные и точные науки, инновационные технологии. – Томск: Изд-во ТГПУ, 2005. – С. 53-58.

12. Шевелев О.Г. Общая схема программного комплекса для проведения стилеметрических исследований // Материалы XLIII Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск: Новосибирский гос. ун-т, 2005. – С. 244-245.

13. Шевелев О.Г. Представление набора текстов в реляционной базе данных для целей лингвистического анализа // Вестник Томского государственного университета, 2004, № 284. – С. 225-229.

14. Шевелев О.Г. Преобразование текстов к набору частотных признаков для проведения лингвистических исследований // XI Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Современные техника и технологии» (29 марта – 2 апреля 2005 г.): Труды конференции в 2-х т. – Томск: Изд-во Томского политехн. ун-та, 2005. – Т.2. – С. 264-267.

15. Шевелев О.Г., Тютерев В.В. Многослойные персептроны в задаче разрешения спорного авторства текста // Сборник трудов научно-технической конференции «Нейроинформатика-2003» (29-31 января 2003 г.), Ч.2. – М.: МИФИ, 2003. – С. 206-212.

16. Шевелев О.Г. Энтропийный отбор входов нейронной сети в задаче классификации текстов по авторству // Нейроинформатика и ее приложения (7-9 октября 2005 г.): Материалы XIII Всероссийского семинара / Под ред. А.Н.Горбаня, Е.М. Миркеса. – Красноярск: ИВМ СО РАН, 2005. – С. 131-132.