УДК 004.442: 519.25

А.С. Романов

Структура программного комплекса для исследования подходов к идентификации авторства текстов

Описывается структура программного комплекса для исследования подходов к идентификации авторства текстов и характеристик текста. Предложенная структура программы позволит решать все базовые задачи идентификации авторства текстов.

Современное информационное общество использует вычислительные машины различного рода практически во всех сферах жизнедеятельности и, прежде всего, в научных исследованиях. В своем современном воплощении компьютеры и сопутствующие им информационные системы представляют собой идеальное техническое решение задач обработки больших объемов статистических данных и решения сложных вычислительных задач, необходимых, в частности, в образовательном процессе, лингвистических и криминалистических исследованиях для определения авторства текста.

Для исследований разрабатывается программный комплекс, функциональные блоки которого представлены на рис. 1.

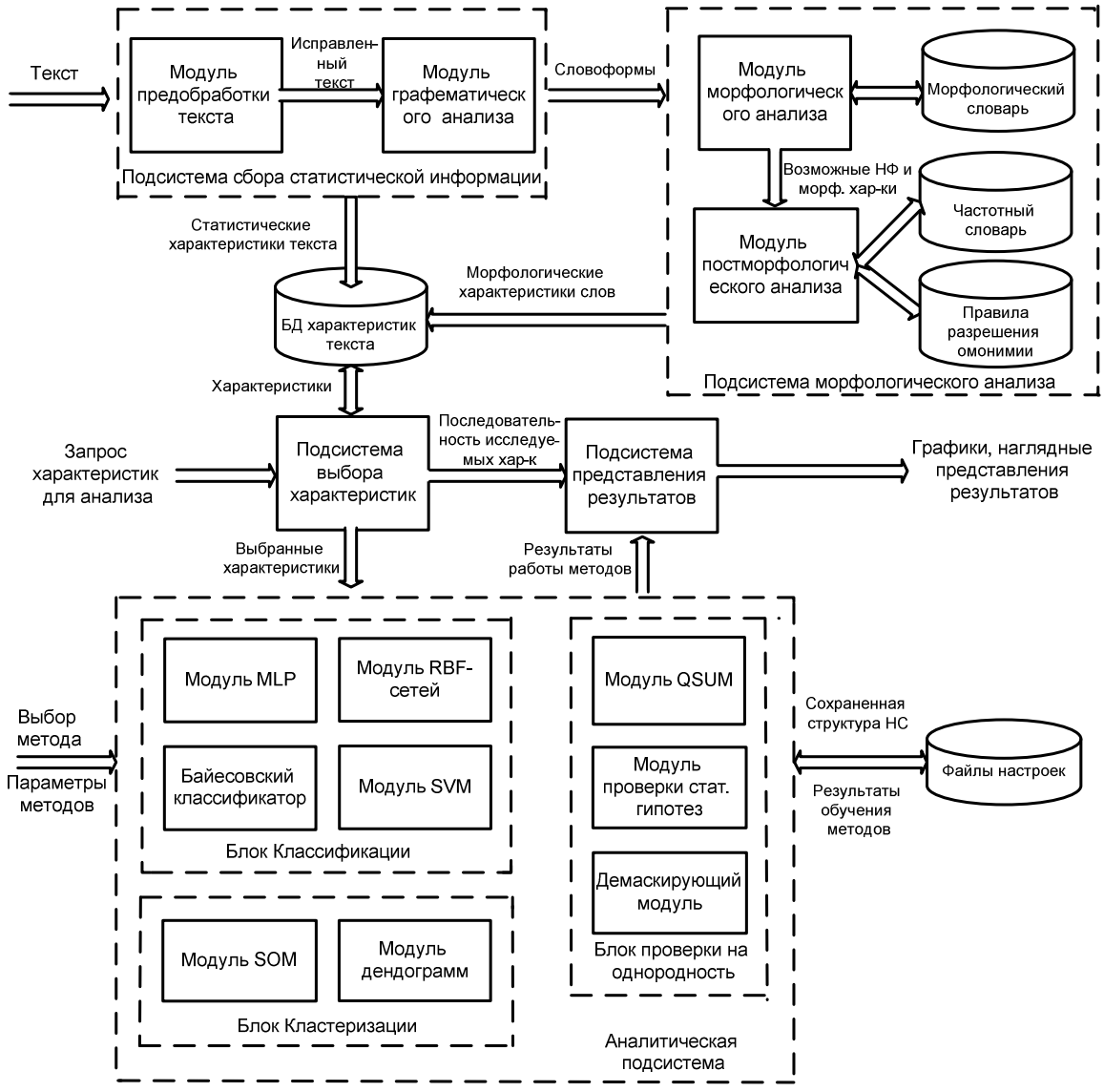

Рис. 1. Структура программного комплекса

Функции каждого из блоков строго определены, что позволяет реализовывать систему поэтапно и исследовать применяемые методы по мере их программной реализации. Опишем процесс анализа текста и модули программы более подробно.

Подсистема сбора статистической информации состоит из двух модулей: предобработки текста и графематического анализа.

Функциями *модуля предобработки текста* являются: поддержка возможности работы с различными кодировками текста; очистка текста от «лишних» символов (перевода строки, двойных тире и пробелов и т.д.); исправление явных ошибок, связанных с вводом текста, которые не влияют и не учитываются при проведении непосредственно идентификации и др.

Функции *модуля графематического анализа*: сегментация текста на предложения, слова, словосочетания, символы; сбор статистических сведений о выделенных единицах.

Выделенные на этапе графематического анализа словоформы поступают в *подсистему морфологического анализа*.

Подсистема морфологического анализа предназначена для определения основных морфологических характеристик слов, а также нормальной формы слова. Она состоит из *морфологического словаря* (в нем хранятся неизменяющиеся основы слов, изменяющиеся части слов – аффиксы, слова-исключения, а также связанная с ними морфологическая информация), и непосредственно *модуля морфологического анализа*, который занимается поиском слова и его характеристик в словаре. В программе используются морфологические модули Диалинг [1].

На случай возникновения ситуации морфологической омонимии в подсистему добавлен *модуль постморфологического анализа*. В настоящий момент снятие омонимии происходит с помощью частотного словаря Шарова [2]. Выбирается тот вариант, частота употребления которого в русском языке больше или меньше других (в зависимости от настроек программы).

Вся собранная двумя подсистемами информация помещается в *базу данных характеристик текста*. Используется СУБД MySQL 5.1.

Текст в базе данных хранится как список предложений. Предложения хранятся как список составляющих их словоформ, последовательностей знаков препинания, цифр и т.д., которые, в свою очередь, представлены в БД в виде составляющих их символов. Такой подход позволяет извлекать из БД выборки любой сложности с помощью запросов на языке SQL, учитывая при этом иерархическую структуру текста.

Пользователь может сформировать запрос и получить выборку из БД с помощью *подсистемы выбора характеристик*. Основное же её назначение – отбор характеристик для использования одним из методов идентификации, реализованных в *Аналитической подсистеме*.

Первый блок аналитической подсистемы – *модуль проверки на однородность и близость авторского стиля* – предназначен для выделения фрагментов, не соответствующих общему авторскому стилю текста, а также для сравнения двух текстов на предмет сходств и различий.

Модуль QSUM (накопительных сумм). Принцип работы метода [3] заключается в сопоставлении двух графиков, масштабированных относительно друг друга. Первый из них составляют точки, показывающие накопительную сумму отклонений предложений от средней длины предложения текста на текущем этапе вычисления. Второй – накопительную сумму отклонений некоторой характеристики текста, являющейся функцией предложения, от её среднего значения в тексте. Для однородного стиля графики должны практически совпадать, тогда как неоднородный текст покажет их различие. Исследования автора показали, что такой характеристикой для русского языка может быть количество функциональных слов в предложении. Возможны некоторые вариации в зависимости от конкретного текста.

Модуль проверки статистических гипотез. Реализация вычисления основных формул математической статистики и теории вероятности (используются также другими методами), а также проверка статистических гипотез о равенстве средних на основе критерия Стьюдента, проверка текста на однородность с помощью критерия Колмогорова–Смирнова, вычисления мер хи-квадрат, Кульбака и др.

Демаскирующий модуль. Задача проверки текста на однородность представляется как задача классификации по одному классу. Суть демаскирующего подхода [4] заключается в итеративном обучении классификатора, способного разделять два класса. На каждом шаге обучающие выборки «ослабляются» путем исключения наиболее эффективных в распознавании характеристик (причем как оказывающих положительный, так и отрицательный эффект) по отношению к модели, полученной на предыдущем шаге. Затем проводится анализ снижения качества классификации в процессе подбора модели: если после последнего «ослабляющего» этапа классы по-прежнему могут быть разделены с небольшими ошибками, то предполагаем, что они взяты от разных авторов.

Вторая составляющая аналитической подсистемы – *блок классификации*. В общем случае необходимо обучить выбранный классификатор на некотором наборе характеристик текстов, принадлежащих предполагаемым авторам. Затем на входы обученного классификатора подается тот же набор характеристик текста, авторство которого необходимо определить. Классификатор формирует вектор-ответ, содержащий информацию об авторстве.

Байесовский классификатор – реализация классификатора, в основе которого лежит формула Байеса [5]: по известным априорным вероятностям и функции правдоподобия необходимо найти максимум апостериорной вероятности.

В *модуле MLP* реализуется работа с одной из самых популярных архитектур нейронных сетей – «многослойным перцептроном» [6]. В качестве алгоритма обучения используется алгоритм обратного распространения ошибки. У этого метода есть ряд существенных недостатков, главными из которых являются сравнительно низкая скорость обучения, высокая склонность к переобучению и необходимость эксперименталного подбора параметров для каждой конкретной задачи (количества слоев, количества нейронов в них и т.д.).

Модуль RBF – сети радиальных базисных функций [7]. Сети имеют один скрытый слой, который состоит из радиальных элементов, каждый из которых воспроизводит гауссову поверхность отклика. Преимущества перед MLP: обучение сети происходит быстрее, процессы принятия решений в RBF-сети легче поддаются объяснению, существует возможность использования некоторой предварительной информации в качестве отправной точки для обучения.

В дальнейшем планируется отказаться от использования MLP и RBF и полностью перейти на использование метода опорных векторов [8, 9] – *модуль SVM*. SVM используют ядровые преобразования для увеличения рамерности пространства признаков таким образом, чтобы разделяющая их поверхность была линейной и позволяла проводить классификацию по двум классам. В методе опорных векторов нет необходимости задавать количество элементов в промежуточном слое, как это делается в MLP, а обучение происходит быстрее. Полнота и точность классификации метода SVM выше, чем других классификаторов. Применение сигмоидального или гауссовского ядра позволяет, соответственно, смоделировать сети MLP и RBF.

Третий блок аналитической подсистемы – *блок кластеризации* – состоит из *модуля дендограмм* [10], реализующего иерархический метод кластеризации, и *модуля SOM*, реализующего неиерархический метод кластеризации на основе самоорганизующихся карт Кохонена [11].

Результатом работы первого модуля является дерево, на котором графически представлен процесс объединения текстов в кластеры на основе матриц расстояний. При этом для объединения используется метод «дальнего соседа».

В процессе обучения сетей Кохонена происходит постепенная подстройка весов нейронов под обучающие данные. При этом на каждом шаге корректируются центры кластеров. Результаты кластеризации текстов отображаются на топологической карте.

В аналитической подсистеме зарезервировано место для добавления новых методов идентификации.

Методы, использование которых подразумевает обучение, хранят параметры обучения в *файлах настроек*, там же сохраняются общие настройки программы.

Последней подсистемой является *подсистема представления результатов*. Её основной функцией являются обработка результатов работы аналитического блока и вывод их в понятной и наглядной для исследователя форме (например, иерархического дерева как результата работы метода дендограмм, расшифровки компонентов вектора-ответа для методов классификации и т.д.), а также представление сформированных пользователем выборок из базы данных в виде таблиц и графиков.

Представленная структура ПО позволит решать такие базовые задачи идентификации автора [12], как:

– множественная неопределенность;

– конкуренция образцов;

– сравнение по образцу.

В настоящее время ведется активная работа по усовершенствованию блока проверки текста на однородность и выявлению возможного плагиата без базы возможных источников заимствований. После окончательной доработки и проведения испытаний планируется выделить её в отдельный программный продукт.

Работа поддержана грантом ФСРМПНТ.

Литература

1. Сокирко А.В. Морфологические модули на сайте www.aot.ru. – Режим доступа: http://www.aot.ru/docs/sokirko/Dialog2004.htm, свободный. – Загл. с экрана.

2. Шаров С.А. Частотный словарь. – Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный. – Загл. с экрана.

3. Morton A.Q. The Authorship of Greek Prose // Journal of the Royal Statistical Society (A). – 1965. – 128. – P. 169–233.

4. Benno Stein, Sven Meyer zu Eissen Intrinsic Plagiarism Analysis with Meta Learning / In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors // SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07). – 2007. – P. 45–50.

5. Kevin B. Korb Bayesian Artificial Intelligence. – CRC Press, 2003. – 392 p.

6. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – М.: Горячая линия - Телеком, 2001. – 382 с.

7. Mark J. L. Orr Introduction to Radial Basis Function Networks // Technical Report. – Center for Cognitive Science, University of Edinburgh, 1996. – Режим доступа: http://www.cns.ed.ac.uk/people/mark.html, свободный. – Загл. с экрана.

8. Vapnik V.N. The nature of statistical learning theory. – New York: Springer-Verlag, 2000. – 332 p.

9. Burges C. A Tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. – 1998. № 2. – P. 955–974.

10. Поддубный В.В., Шевелев О.Г., Бормашов Д.А. Сравнение качества подходов к кластеризации текстов на основе гипергеометрического критерия // Вестник Том. гос. ун-та. – 2006. – № 293. – С. 120–125.

11. Kohonen T. Self-Organizing Maps (Third Extended Edition). – New York: Springer-Verlag, 2001. – 501 p.

12. Баранов А.Н. Авторизация текста: пример экспертизы // Цена слова: Из практики лингвистических экспертиз текстов СМИ в судебных процессах по защите чести, достоинства и деловой репутации: Сб. науч. тр. / Под ред. М.В. Горбаневского. – 3-е изд., испр. и доп. – М.: Галерия, 2002. – С. 238–242.

Романов Александр Сергеевич
ГОУ ВПО Томский государственный университет систем управления и радиоэлектроники, аспирант.
Эл. почта: ras@ms.tusur.ru.

A.S. Romanov

The architecture of software for authorship attribution techniques research

In this paper the architecture of software for research of authorship attribution techniques and text's features are described in detail. Given architecture of software make it possible to solve main problems of texts' authorship identification.
