УДК 519.25: 004.8

А.С. Романов, А.А. Шелупанов, С.С. Бондарчук

Обобщенная методика идентификации автора неизвестного текста

Рассмотрена проблема идентификации автора текста при ограниченном наборе альтернатив. Предложена обобщенная методика для идентификации автора неизвестного текста и получения модели для разделения авторских стилей в виде обученного классификатора. Приводятся результаты экспериментов по применению методики на корпусе литературных текстов и корпусе коротких электронных сообщений.
Ключевые слова: идентификация автора текста, классификатор, нейронные сети, машина опорных векторов, сглаживание вероятностей.

Задача определения авторства неизвестного текста является важной проблемой информационной безопасности. Это связано, прежде всего, с широким распространением программ для обмена сообщениями в сети Интернет (интернет-мессенджеров), возросшей ролью электронной почты в деловой переписке, высокой популярностью интернет-форумов и блогов. Пользователи имеют возможность отправлять сообщения без регистрации и указания какой-либо информации о себе, а регистрация сама по себе зачастую носит чисто символический характер. То же самое касается интернет-мессенджеров и электронной почты – регистрационные данные не позволяют однозначно идентифицировать личность собеседника, адрес отправителя можно легко изменить. Анонимность сообщений в сети Интернет всё чаще привлекает злоумышленников для совершения преступлений в киберпространстве.

За более чем 120-летнюю историю развития данного вопроса отечественными и зарубежными исследователями было предложено множество методов определения автора текста, начиная от простого подсчета количества определенных слов в сравниваемых текстах и заканчивая разработками в области искусственного интеллекта. Главной проблемой традиционных работ по данной тематике является использование при проведении экспериментов текстов объемом более 30000–40000 символов и большого количества обучающих примеров (от 5 до 100 и более). Нерешенной задачей является идентификация авторства коротких текстов.

Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующем образом. Имеется множество текстов $T=\{t_1,\ldots,t_k\}$ и множество авторов $A=\{a_1,\ldots,a_l\}$. Для некоторого подмножества текстов $T'=\{t_1,\ldots,t_m\}\subseteq T$ авторы известны, т.е. существует множество пар «текст–автор» $D=\{(t_i,a_j)\}_{i=1}^{m}$. Необходимо установить, кто из множества A является истинным автором остальных текстов (анонимных или спорных) $T''=\{t_{m+1},\ldots,t_k\}\subseteq T$.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами [1]. В этом случае множество A составляет множество предопределенных классов и их меток, D – обучающие примеры, а множество T'' – классифицируемые объекты. Целью является построение классификатора, решающего данную задачу, т.е. нахождение некоторой целевой функции $F: T\times A\to[-1,1]$, относящей произвольный текст множества T к его истинному автору. Значения функции интерпретируется как степень принадлежности объекта классу: 1 соответствует полностью положительному решению, -1 – отрицательному.

Для решения задачи можно использовать любой из разработанных на данный момент алгоритмов классификации.

Обобщенная методика идентификации автора неизвестного текста показана на рис. 1.

Методика включает последовательность следующих действий:

1. Выбор модели представления текстов в виде наборов признаков.
2. Выбор группы признаков для проверки и формирования из неё авторского инварианта.
3. Выбор классификаторов и их параметров.
4. Формирование модели авторского стиля, позволяющей разделять двух и более авторов на основе полученного авторского инварианта и обученного классификатора.
5. Непосредственно определение авторства неизвестного текста.
6. Принятие итогового решения об авторе текста ансамблем классификаторов в случае, если удалось найти несколько информативных групп признаков текста.

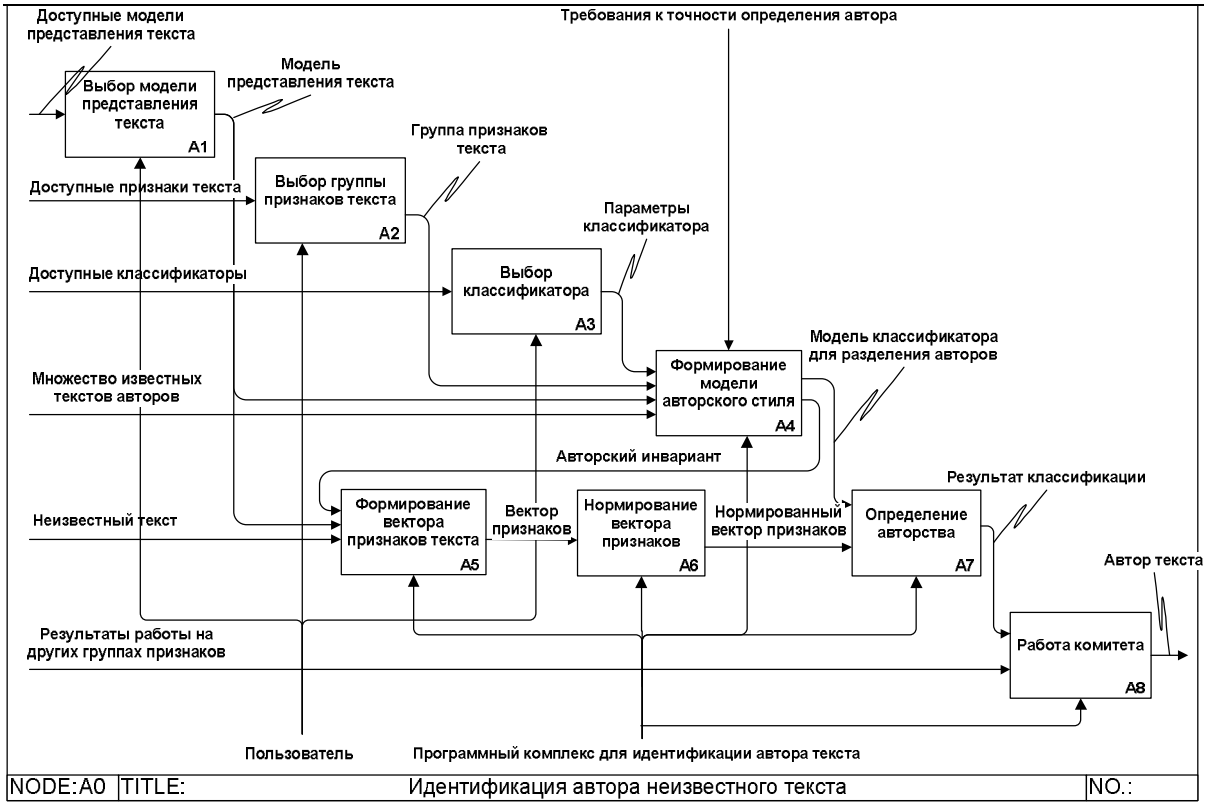

Рис. 1. Методика идентификации автора неизвестного текста

Важным этапом является процесс формирования модели отличий авторских стилей (рис. 2).

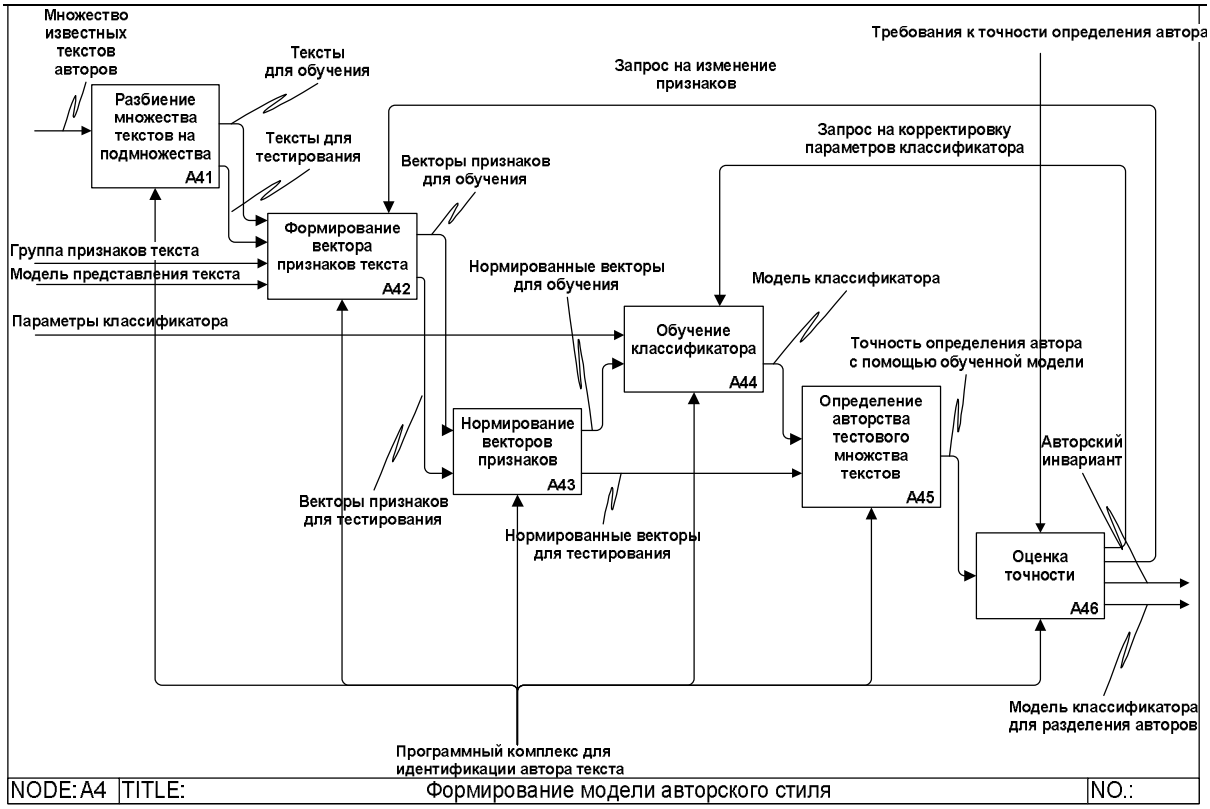

Рис. 2. Формирование модели авторского стиля

Для определения отличий стилей авторов предлагается следующая последовательность действий:

1. Разбиение имеющегося множества текстов на две группы. Первая используется для обучения модели классификатора. Вторая – для проверки точности идентификации автора с помощью обученной модели.

2. Формирование вектора признаков текста из характеристик полученного авторского инварианта в соответствии с выбранной моделью представления текста в виде набора признаков.

3. Приведение значений признаков в единый диапазон с помощью операций нормирования и шкалирования.

4. Корректировка параметров классификатора, позволяющих обеспечить высокую разделяющую способность исследуемых авторов, путем обучения классификатора на нормированных векторах признаков группы обучающих текстов и проверки точности обученного классификатора на векторах признаков тестовой группы текстов. Первоначальное обучение классификатора происходит с параметрами по умолчанию.

5. Изменение перечня групп характеристик и/или признаков, составляющих группу, в случае если изменением параметров классификатора достичь приемлемых результатов не удается.

Итогом является обученный классификатор, веса связей которого настроены таким образом, чтобы классификатор был способен разделить стили авторов, на текстах которых он обучался, при подаче на его входы подобранного набора признаков.

Разработанная методика, помимо информативности признаков текста, анализируемых в статистических методах идентификации авторства, учитывает влияние общей способности классификатора к разделению данных и его точность. Главной особенностью предложенного подхода является принятие итогового решения об авторе текста несколькими классификаторами (ансамблем классификаторов) по принципу мажоритарного голосования в случае, если удалось найти несколько информативных групп признаков текста.

Обозначим ключевые параметры методики.

В качестве модели представления текста в виде набора признаков использовались N-граммы уровня символов, сглаженные аддитивным методом Лапласа [2]:

$$P_{ADD}(a_i,\ldots,a_{i+n-1}) = \frac{1 + C(a_i,\ldots,a_{i+n-1})}{\delta W + \sum_i C(a_i,\ldots,a_{i+n-1})},$$

где $C(\cdot)$ – количество употреблений N-граммы; W – количество всех N-грамм в используемом словаре или алфавите.

Всего было исследовано порядка 45 различных признаков текста уровней символов, слов и предложений. Рассмотрены случаи 2, 5, 10 предполагаемых авторов. Количество обучающих примеров в экспериментах выбиралось исходя из потребностей при решении реальных задач идентификации автора, когда количество материала ограничено. Использовались выборки объемом 1000–100000 символов (~200–20000 слов). Количество обучающих примеров каждого автора бралось равное 3, для тестирования использовалось по 1 выборке автора. Корпус для исследований состоит из 215 текстов 50 русских писателей. Тексты взяты из электронной библиотеки М. Мошкова [3]. Размер каждого текста составляет более 100000 символов.

В качестве результирующей точности по данному признаку и объему выборки подсчитывалась средняя частота правильных классификаций. В результате сделан вывод, что авторский стиль лучше всего описывается характеристиками, представленными в таблице.

Средняя точность идентификации автора по всем объемам (1000–100000 символов)

Признак	Средняя точность идентификации по всем объемам								
	2 автора			5 авторов			10 авторов		
	MLP	CCN	SVM	MLP	CCN	SVM	MLP	CCN	SVM
УНИГРАММЫ	0,905	0,857	0,835	0,684	0,518	0,618	0,544	0,625	0,49
ТРИГРАММЫ_300	0,943	0,943	0,927	0,786	0,708	0,881	0,657	0,657	0,776
ШАРОВ_500	0,920	0,903	0,931	0,753	0,786	0,838	0,586	0,670	0,823
ПУНКТУАЦИЯ	0,921	0,917	0,861	0,757	0,774	0,745	0,609	0,746	0,739

Примечание: УНИГРАММЫ – частоты букв русского алфавита; ТРИГРАММЫ_300 – частоты 300 наиболее частых триграмм; ШАРОВ_500 – частоты 500 наиболее частых слов из словаря Шарова [4]; ПУНКТУАЦИЯ – частоты знаков пунктуации.

В качестве классификаторов использовались искусственные нейронные сети двух архитектур: классический многослойный перцептрон (MLP) и сети каскадных корреляций (CCN), и метод на основе машины опорных векторов (SVM).

Параметры перцептрона были выбраны исходя из результатов исследований, приведенных в работе [5], подтверждающихся также собственными исследованиями автора:

– алгоритм обучения – алгоритм обратного распространения ошибки;

– функция активации скрытых слоев – сигмоид;

– функция активации выходного слоя – сигмоид;

– скорость обучения 0,7, момент 0,0;

– количество скрытых слоев 1;

– количество нейронов в скрытом слое 10;

– максимальное количество эпох обучения 50000;

– допустимый уровень ошибки 0,00001.

Параметры обучения нейронных сетей каскадных корреляций выбирались аналогичными параметрам многослойного перцептрона:

– алгоритм обучения – быстрого распространения ошибки;

– функция активации скрытых слоев – сигмоид;

– функция активации выходного слоя – сигмоид;

– максимальное количество нейронов, которое можно добавить 100;

– допустимый уровень ошибки 0,00001.

Параметры обучения метода опорных векторов выбирались исходя из рекомендаций, приведенных в [6]. После проведения экспериментов параметры в общем случае были выбраны следующие:

– алгоритм обучения – метод последовательной оптимизации;

– ядро – линейное;

– параметр регуляризации C = 1;

– допустимый уровень ошибки – 0,00001.

Эксперименты показали, что автора можно определить с точностью в среднем 0,95–0,98 при объеме текстовой выборки 20000–25000 символов при использовании признаков ТРИГРАММЫ_300 и ШАРОВ_500. При этом начиная с 10000 символов, машина опорных векторов показывает лучшие из трех исследуемых классификаторов результаты.

Объединение результатов классификации с помощью трех описанных методов и признаков из таблицы по принципу мажоритарного голосования позволяет увеличить точность на 6–12% на объемах текста до 10000 символов. Итоговые графики для случаев 2, 5, 10 и 50 предполагаемых авторов представлены на рис. 3–4.

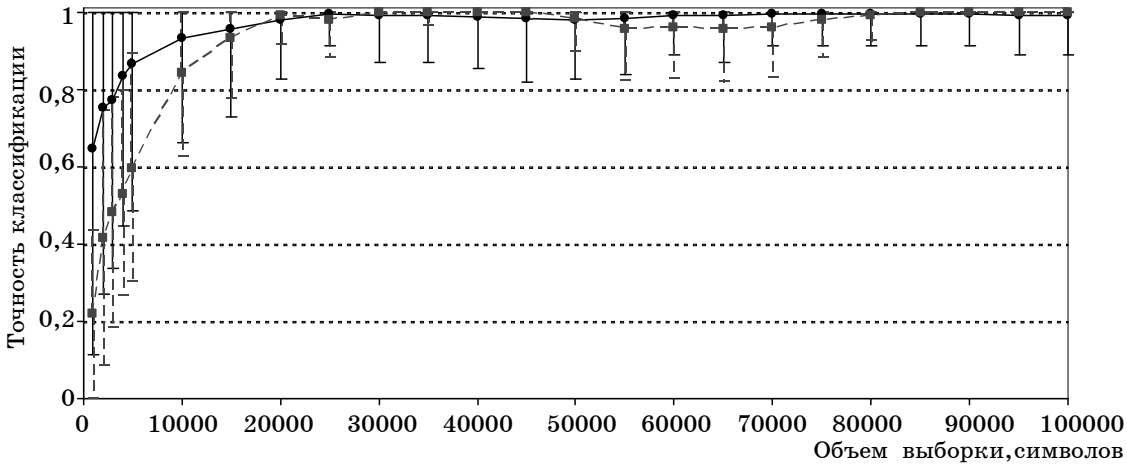

Рис. 3. Результаты исследования методики с применением мажоритарного голосования: (–•–) 2 и (–■–) 5 предполагаемых авторов

Средняя точность классификации для первых трех случаев равна 0,98 при объеме выборки, равном 20000 символов. Такие показатели для русского языка были достигнуты впервые.

Полученная методика была применена на практике для идентификации автора коротких электронных сообщений. Результаты показали, что авторство коротких электронных сообщений длиной 100 символов можно определить с точностью 0,70±0,17 в случае двух потенциальных авторов. При решении частных задач путем исключения из обучающего множества не характерных для автора текстов достигнута точность 0,86±0,06.

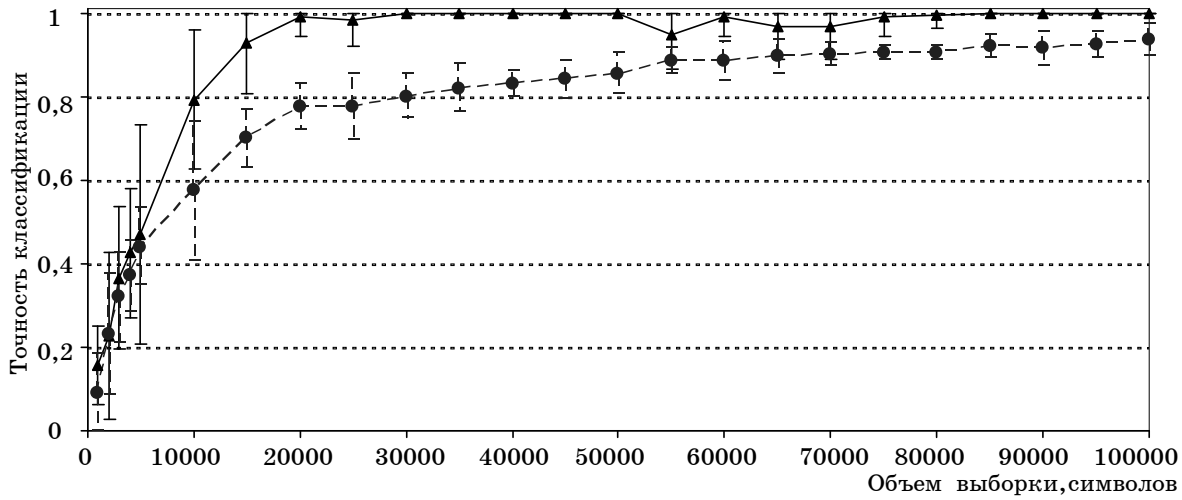

Рис. 4. Результаты исследования методики с применением мажоритарного голосования (–▲–) 10 и (–•–) 50 предполагаемых авторов

Литература

1. Романов А.С. Идентификация автора текста с помощью аппарата опорных векторов / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной междунар. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). – М. : РГГУ, 2009. – Вып. 8 (15). – С. 432–437.

2. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов / А.С. Романов // Докл. Том. гос. ун-та систем управления и радиоэлектроники. – 2009. – № 1 (19), Ч. 2. – С. 36–42.

3. Библиотека Максима Мошкова [Электронный ресурс]. – Режим доступа: http://www.lib.ru, свободный (дата обращения: 21.05.2010).

4. Шаров С.А. Частотный словарь [Электронный ресурс]. – Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный (дата обращения: 21.05.2010).

5. Шевелев О.Г. Методы автоматической классификации текстов на естественном языке: учеб. пособие. – Томск : ТМЛ-Пресс, 2007. – 144 с.

6. Hsu C.-W. A practical guide to support vector classification [Electronic resource] / C.-W. Hsu, C.-C. Chang, C.-J. Lin. – Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, свободный (дата обращения: 21.05.2010).

Романов Александр Сергеевич
Аспирант каф. комплексной информационной безопасности
электронно-вычислительных систем ТУСУРа
Тел.: 8-952-883-84-34
Эл. адрес: alexx.romanov@gmail.com

Шелупанов Александр Александрович
Зав. каф. комплексной информационной безопасности электронно-вычислительных систем ТУСУРа
Эл. адрес: saa@tusur.ru

Бондарчук Сергей Сергеевич
Доктор физ.-мат. наук, профессор,
Томский государственный педагогический университет
Тел.: (3822) 41-34-26
Эл. адрес: office@security.tomsk.ru

A.S. Romanov, A.A. Shelupanov, S.S. Bondarchuk
Generalized authorship identification technique

In the article authorship identification problem in the case of the limited set of alternatives is described. Generalized technique for authorship identification and for authors' style model generating in the form of studied classifier is given. Also results of authorship identification experiments for corpuses of Russian literary texts and short messages are represented.
Keywords: authorship identification, classifire, artifical neural networks, support vector machine, smoothing.