УДК 519.25: 004.8

А.С. Романов

Методика идентификации автора текста на основе аппарата опорных векторов

Рассматривается проблема идентификации автора текста при ограниченном наборе альтернатив как задача классификации. Для её решения предлагается использовать классификатор на основе аппарата опорных векторов и N-граммные признаки текста. Рассматривается возможность повышения точности классификации за счет использования методов сглаживания вероятностей. Приводятся результаты экспериментов, подтверждающие теоретические предположения.

Ключевые слова: идентификация автора текста, признаки текста, классификатор, машина опорных векторов, сглаживание вероятностей.

Важность задачи идентификации автора печатного текста обуславливается повсеместным переходом от рукописного письма к печатному способу набора. При возникновении ситуации спорного авторства, при криминалистическом исследовании печатного текста развитые методы идентификации по почерку могут оказаться бесполезными. Кроме того, почерковедческая экспертиза позволяет определить лишь исполнителя, но не автора текста. Проведение авторoведческой экспертизы в настоящее время осуществляется с привлечением экспертов, имеющих соответствующее образование в области лингвистики и филологии. Эффективных же количественных методов и инструментов автоматического определения авторства небольших текстов на русском языке на данный момент не разработано, поэтому исследования в данном направлении остаются актуальными.

Подобные методы, помимо криминалистики, могут найти своё применение также и в других областях.

В лингвистических исследованиях данные методики могут использоваться для изучения феномена авторства. Интерес здесь представляет отличие в стиле того или иного писателя; черты, которые делают его речь легко узнаваемой; индивидуальность или общность каких-либо характеристик.

Существует ряд неатрибутированных литературных текстов, а также произведения, авторство которых до сих пор находится под сомнением. Так многие скептики приписывают авторство как минимум нескольких глав романа «Тихий Дон» М.А. Шолохова менее знаменитому казацкому писателю Ф.Д. Крюкову. И, несмотря на то, что почерковедческая экспертиза подтвердила, что черновики романа написаны рукой Шолохова, вопрос о том, пользовался ли он при этом какими-либо источниками, остается открытым. Аналогично зарубежные исследователи подвергают сомнению авторство ряда произведений, приписанных У. Шекспиру.

Очевидно, что существование точных количественных методов идентификации автора и проведение авторoведческой экспертизы на их основе могли бы разрешить большинство спорных вопросов в области литературоведения и истории.

Другой сферой применения является сфера образования. Школьники и студенты всё реже сами выполняют рефераты, курсовые работы и доклады, предпочитая не тратить на это время и просто скачать готовые работы из сети Интернет. Использование методик определения авторства в этом случае позволит более объективно оценивать учащихся.

Эффективные методы идентификации на основе устойчивых характеристик можно применять и для решения ряда смежных задач: идентификации пола и гендера, профессии, национальности, уровня образования автора и т.д.

Описание экспериментальной методики. Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующим образом. Имеется множество текстов $T = \{t_1, \ldots, t_k\}$ и множество авторов $A = \{a_1, \ldots, a_k\}$. Для некоторого подмножества текстов $T' \subseteq T$ авторы известны $D = \{(t_i, a_i)\}_{i=1}^{\ell}$. Необходимо установить, кто из множества A является истинным автором остальных текстов (анонимных или спорных) $T'' = \{t_{|T'|+1}, \ldots, t_k\} \subseteq T$.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами [1, 2]. 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 $F: T \times A \to [0,1]$, относящей произвольный текст множества T к его истинному автору. Значения функции интерпретируются как степень принадлежности объекта классу: 1 соответствует положительному решению, 0 – отрицательному.

Для решения поставленной задачи в исследованиях используется классификатор на основе аппарата опорных векторов (Support Vector Machine, SVM), математический аппарат которого был предложен В.Н. Вапником в работах [3, 4], и одна из его популярных реализаций – библиотека libsvm [5]. Исследования отечественных и зарубежных авторов [1, 6] показывают, что SVM на сегодня является одним из лучших методов классификации. В отличие от искусственных нейронных сетей, применявшихся автором ранее [7], SVM лучше подходит для работы с большим признаковым пространством, что важно при использовании N-граммных признаков текста. Нет необходимости в выборе количества скрытых элементов, скорость работы SVM существенно выше, чем нейронных сетей.

Программная реализация метода SVM интегрирована в общую программную оболочку, структура которой описана в работе [8].

Для того чтобы классификация проходила успешно, необходимо в первую очередь добиться высокой точности при решении задач бинарной классификации. Важными этапами при этом являются выбор параметров алгоритма классификации, количества обучающих примеров, а также выбор характеристик текста для анализа и необходимого объема выборок.

Текст можно рассматривать как иерархическую структуру [9] и анализировать на любом уровне как последовательность отдельных составляющих его элементов (символов, словоформ, грамматических классов и т.д.) или групп элементов длиной N–N-грамм.

Следует отметить, что анализ структуры текста усложняется при использовании признаков более высоких уровней иерархии и с каждым новым уровнем труднее поддается автоматизации. Так, в процессе морфологического анализа используется информация, полученная на этапе лексического анализа, на этапе синтаксического анализа – информация, полученная на этапе морфологического анализа, и т.д. На каждом из этапов могут возникнуть неточности, например из-за зашумленности анализируемого текста, особенностей русского языка (морфологическая, синтаксическая омонимия) и т.д., которые далее приведут к серьезным ошибкам на более высоких уровнях анализа. Поэтому в данном исследовании было решено ограничиться характеристиками уровней символов и слов – использовались биграммы и триграммы символов, наиболее частые слова русского языка [10].

При использовании частот появления в тексте N-грамм как признаков для идентификации исследователь может столкнуться с проблемой «разреженных данных» (sparse data), когда в обучающем или исследуемом корпусе, в силу того, что используется не генеральная выборка данных, отсутствует часть признаков. Особенно проблема актуальна для текстов небольших размеров и N-грамм высоких порядков. Решением является использование специальных техник сглаживания вероятностей (метод Лапласа, Гуда-Тьюринга, интерполяции, Катца, Кнезера-Нейя и др. [11]), позволяющих оценить вероятности ненаступивших событий.

Самым простым видом сглаживания является аддитивный метод, при котором к каждому элементу матрицы, содержащему количество подсчетов определенной N-граммы, перед переводом в вероятности добавляется некоторая величина δ. Для расчета вероятностей N-грамм используется выражение

$$P_{ADD}(w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta V + \sum_{w_i} c(w_{i-n+1}^{i})},$$

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

Наиболее простым случаем аддитивного сглаживания является метод, когда $\delta=1$ – метод сглаживания Лапласа.

Метод Гуда-Тьюринга лежит в основе более сложных методов сглаживания, поэтому целесообразно рассмотреть результаты его работы для оценки перспективности реализации и использования более трудоемких техник. Основная идея метода заключается в получении значений частот для N-грамм, частота встречаемости которых во фрагменте равна нулю, из N-грамм, встретившихся один раз (hapax legomenon).

Алгоритм базируется на подсчете N_c – количества N-грамм, встретившихся во фрагменте ровно c раз. Частота c заменяется оценкой Гуда-Тьюринга c^* как функцией от величины N_{c+1}:

$$c^* = (c+1)\frac{N_{c+1}}{N_c}. \quad (1)$$

Вместо непосредственно применения выражения (1) для вычисления c^* и дальнейшего расчета частот N-грамм для случая N_0 используется формула

$$P_{GT}^* = \frac{N_1}{N},$$

где N_1 – число hapax legomenon в тексте; N – общее количество рассматриваемых единиц текста. Для N_i, где $i > 0$, частоты рассчитываются по формуле

$$P_{GT}^* = \frac{c^*}{N}.$$

Так как величина c^* зависит от N_{c+1}, то значение выражения (1) становится неопределенным в случае N_{c+1}=0. Для решения данной проблемы, после подсчета значений N_c, но перед использованием их для расчета c^*, значения N_c сглаживаются для замены нулевых элементов в последовательности – заменяются значениями, полученными путем применения операции линейной регрессии, позволяющей найти линейную зависимость между N_c и c в логарифмическом масштабе:

$$\log(N_c) = a + b\log(c).$$

Корпус для исследований состоит из 215 текстов следующих русских писателей: Ч. Айтматов, Б. Акунин, В.П. Астафьев, А.Р. Беляев, М.А. Булгаков, Ф.М. Достоевский, М. Горький, И.С. Тургенев, В.В. Набоков, В.Г. Распутин, К. Булычев, А.И. Куприн, В.Н. Войнович, А. и Б. Стругацкие, Л.Н. Андреев, Г.И. Успенский, Л.А. Чарская, Н.С. Лесков, Ф.М. Решетников, В.С. Соловьев, В.Ю. Агафонов, М.П. Арцыбашев, Н.Г. Гарин-Михайловский, Н.Э. Гейнце, Д.В. Григорович, Д. Гранин, Л. Кассиль, В.Г. Короленко, Д. Мамин-Сибиряк, А.Ф. Писемский, Ю. Бондарев, И.А. Бунин, Г.А. Федосеев, С. Довлатов, И. Грекова, А. Гайдар, А.С. Грин, Е.И. Замятин, К.М. Станюкович, Б. Можаев, Ю. Нагибин, П. Нилин, О. Павлов, В. Пьецух, С.Н. Сергеев-Ценский, М. Симашко, Ф. Сологуб, А. Тарн, А.Н. Толстой, И. Шмелев. Тексты взяты из электронной библиотеки М. Мошкова [12].

Количество обучающих примеров в экспериментах выбиралось, исходя из потребностей при решении реальных задач идентификации авторства, когда количество материала ограничено. Использовались выборки объемом 1000–100000 символов (~200–20000 слов), количество обучающих примеров для каждого автора бралось равным 3, для тестирования использовалось по 1 выборке автора.

Параметры обучения моделей SVM по умолчанию были выбраны следующие:

– ядро на основе радиальных базисных функций (RBF): $k(t,t') = e^{-\gamma\|x-x'\|^2}$;

– значение параметра гамма $\gamma = 0{,}5$;

– значение параметра регуляризации $C = 1$.

Последовательность шагов проведения экспериментов для оценки точности классификации приведена ниже:

1. Выбор параметров обучения моделей SVM, параметров текста для исследований.

2. Применение к каждому тексту операции «склеивания»: все слова приводятся к нижнему регистру, буква «ё» заменяется буквой «е», из текста удаляются все символы форматирования и пунктуации, включая пробел (это позволяет учитывать при анализе также и соединительные биграммы на границе двух слов).

3. Формирование подмножеств сочетаний классов необходимой мощности (без повторений) из всего множества авторов.

4. Для каждого автора из текущей пары формируется по 3 обучающие выборки необходимого объема и одна тестовая. Выборки извлекаются из разных текстов автора.

5. Подсчет интересующих параметров в выборках.

6. Нормирование параметров выборок в диапазон [–1…1].

7. Обучение модели SVM на данных пары выборок.

8. Подача на вход обученной модели SVM данных тестовых выборок, работа классификатора, считывание результатов.

9. Замена для каждого автора тестовой выборки на одну из обучающего множества.

10. Повтор с шага 8 до тех пор, пока каждая из четырех выборок автора не будет использована в качестве тестовой.

11. Увеличение объема выборки на заданный шаг, если предел не достигнут. Повтор с шага 5.

12. Повтор с шага 4 для следующего сочетания классов.

В качестве результирующей оценки точности по данному признаку и объему выборки будем подсчитывать среднюю частоту правильных классификаций.

Результаты модельных экспериментов. С целью отбора наиболее результативных характеристик были проведены предварительные эксперименты на ограниченном корпусе, состоящем из 10 авторов. Для каждого объема выборки было проведено 280 экспериментов (учитывались все сочетания авторов и текстов).

Для ответа на вопрос, влияет ли частота используемых признаков на результаты идентификации, все триграммы символов были отсортированы по частоте встречаемости и разбиты на группы по 500 признаков. Далее с каждой группой были проведены эксперименты по классификации. Результаты представлены на рис. 1.

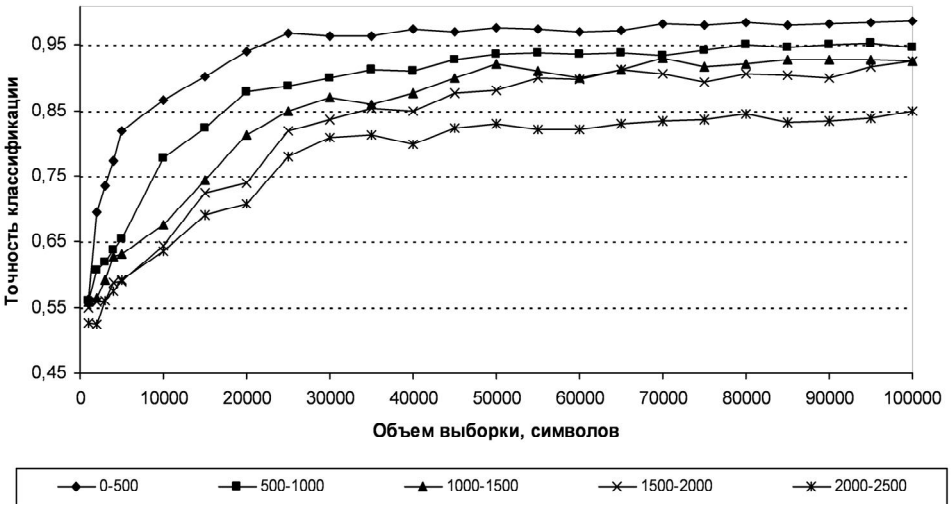

Рис. 1. Влияние частотности признаков на результат классификации на примере 500 самых частых триграмм русского языка

Из графиков видно, что точность классификации снижается по мере использования признаков с меньшей частотой встречаемости. Также в результате экспериментов было установлено, что с повышением количества признаков точность классификации снижается, поэтому использование более чем 500 признаков нецелесообразно.

В таблице представлены признаки текста, вызвавшие наибольший интерес в процессе экспериментов. Всего же было исследованно порядка 30 различных характеристик.

Признаки текста

Обозначение признака	Расшифровка
БИГРАММЫ_100	100 наиболее частых биграмм
ТРИГРАММЫ	Тройки букв русского алфавита
ТРИГРАММЫ_500	500 наиболее частых триграмм
ШАРОВ_100	100 наиболее частых слов из словаря Шарова
ТРИГРАММЫ_ГТ	Триграммы, сглаженные методом Гуда-Тьюринга
ТРИГРАММЫ_АДД1	Триграммы, сглаженные методом Лапласа
ТРИГРАММЫ_500_АДД1	500 наиболее частых триграмм, сглаженных методом Лапласа

Наилучшие результаты без применения методов сглаживания достигаются при использовании характеристик ТРИГРАММЫ_500, БИГРАММЫ_100 и ШАРОВ_100 (рис. 2).

Наиболее точной оказывается классификация по признаку ТРИГРАММЫ_500, средняя частота правильных классификаций для всех объемов выборок составляет 0,91. Стабилизация наступает при объеме выборки, равной 25000, и далее точность колеблется около 0,97. Чуть хуже работает классификация по признакам БИГРАММЫ_100 и ШАРОВ_100 – средняя точность по данным признакам для всех объемов выборок составляет соответственно 0,89 и 0,88.

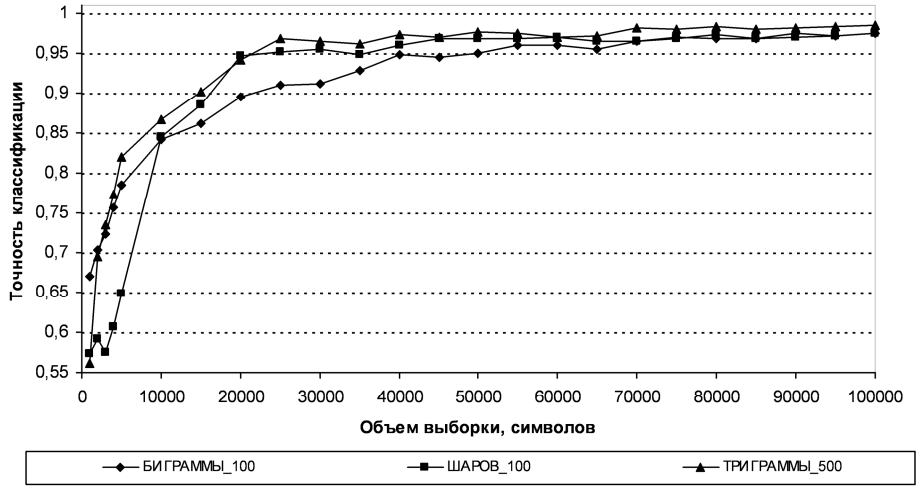

Рис. 2. Характеристики, показавшие наилучшие результаты

Показательны также результаты экспериментов по применению методов сглаживания вероятностей (рис. 3). Так, метод Гуда-Тьюринга не дает положительных результатов. Объяснить это можно тем, что для оценки вероятности ненаступивших событий используется априорная информация о языке, являющаяся общей для всех авторов. Методы Катца и Кнезера-Нейя используют априорную информацию в большей степени, поэтому эксперименты с ними были временно приостановлены. Использование метода Лапласа дает существенный прирост точности. Так, уже при размере выборки 5000 символов точность классификации составляет 0,83, при 10000 символов – 0,92, максимум точности достигается при размере выборки 25000–30000 символов – 0,95. Однако далее точность постепенно снижается до уровня несглаженных триграмм.

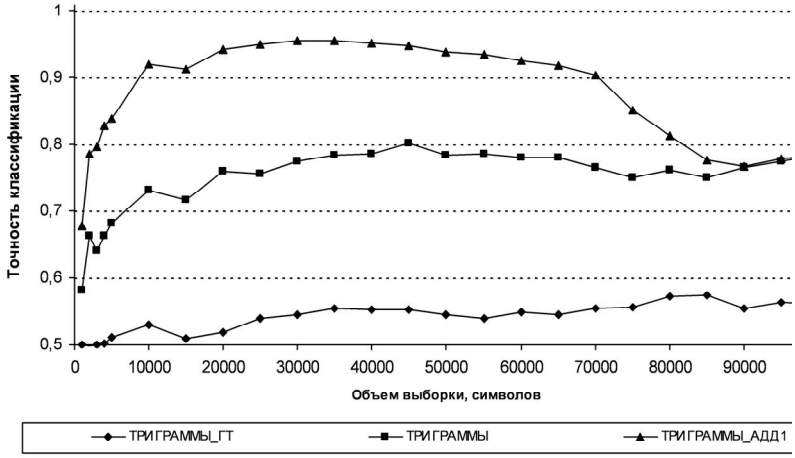

Рис. 3. Результаты исследования методов сглаживания

Результаты экспериментов на всем корпусе текстов. Из рис. 1–2 видно, что наиболее точно классификация проходит при использовании 500 наиболее частых триграмм. Учитывая положительные результаты применения метода сглаживания Лапласа к триграммам символов, за основу при проведении экспериментов на всей текстовой коллекции были взяты частоты встречаемости 500 наиболее частых триграмм русского языка, полученные после сглаживания всех триграмм методом Лапласа.

Классический метод опорных векторов предназначен для решения задачи бинарной классификации. Однако проблему с большим количеством классов можно свести к решению нескольких бинарных задач [13]. Для этого будем использовать стратегию «каждый против каждого» (one-against-one). Классификаторы строятся для каждой пары классов для того, чтобы можно было однозначно разделить любые два класса из множества A. Количество классификаторов в этом случае равно $n(n-1)/2$. После подачи на входы каждого из обученных классификаторов тестового образца получаем ответы, содержащие информацию о его принадлежности одному из двух классов, участвовавших в обучении. К полученному множеству ответов применяется схема мажоритарного голосования, и класс, выбранный большинством классификаторов, принимается как итоговое решение.

Результаты экспериментов на всем имеющемся корпусе представлены на рис. 4–6. Всего для случая 2 предполагаемых авторов было проведено 1225 экспериментов, для 5 авторов – 1764 экспериментов, для 10 авторов – 680 экспериментов с разными комбинациями авторов. Доверительные интервалы построены для доверительной вероятности 0,90.

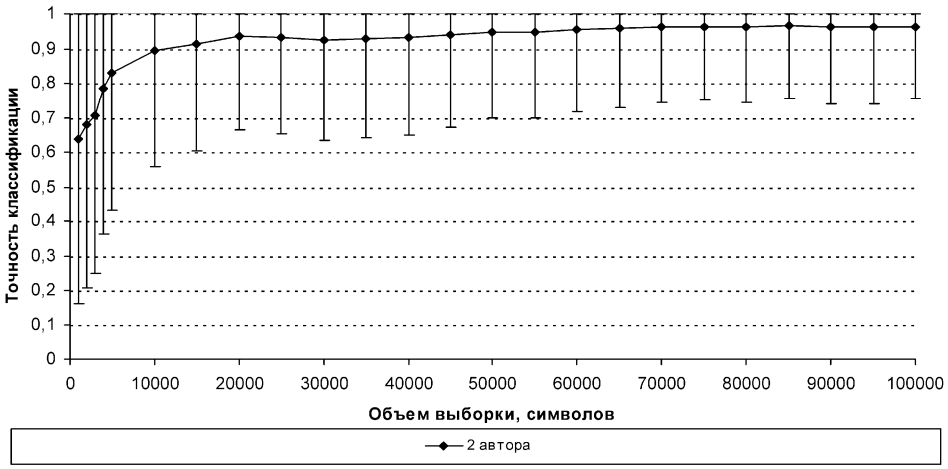

Рис. 4. Результаты исследований по 2 авторам на всем корпусе текстов с использованием признака ТРИГРАММЫ_500_АДД1

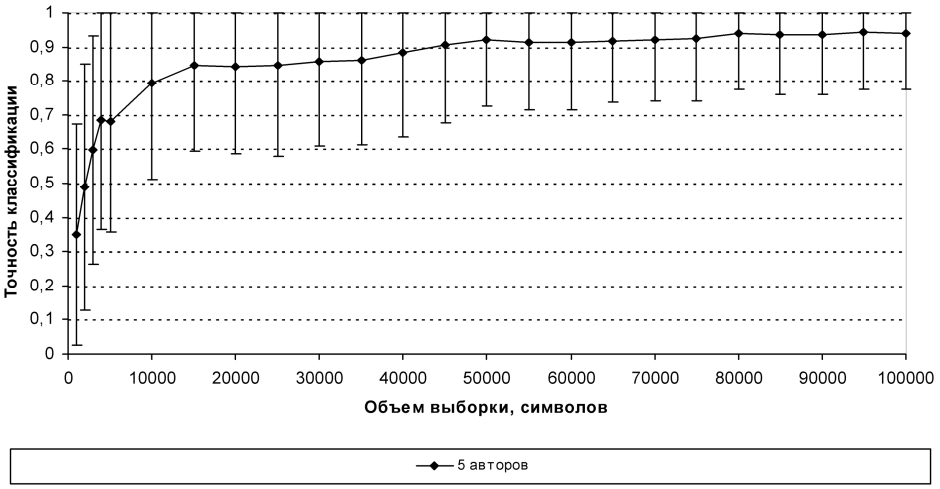

Рис. 5. Результаты исследований по 5 авторам на всем корпусе текстов с использованием признака ТРИГРАММЫ_500_АДД1

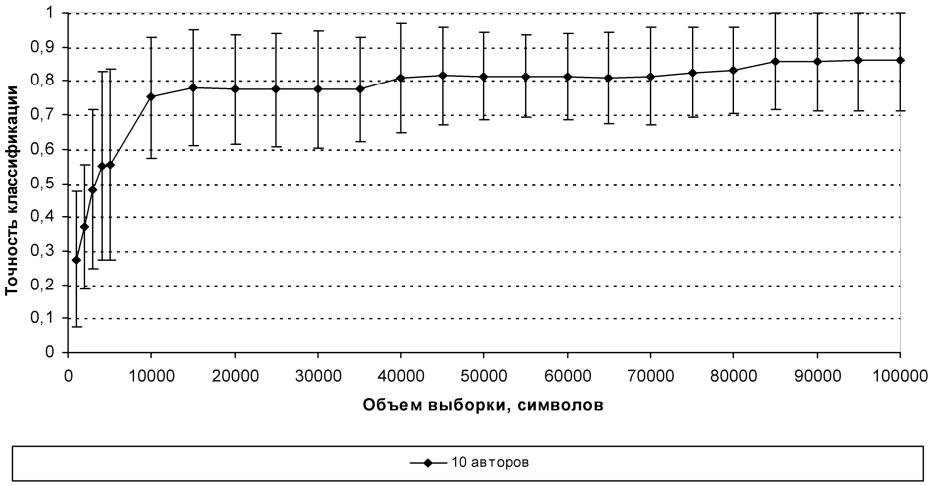

Рис. 6. Результаты исследований по 10 авторам на всем корпусе текстов с использованием признака ТРИГРАММЫ_500_АДД1

Как видно из рис. 4–6, с увеличением количества классов средняя частота правильных классификаций для всех объемов выборок падает с 0,90 для двух классов до 0,82 для 5 классов и до 0,73 для 10 классов. Стабилизация наступает, начиная с 15000 символов.

В целом по результатам экспериментов можно сделать вывод, что метод идентификации автора на основе аппарата SVM и *N*-граммных признаков текста обеспечивает высо-

кое качество классификации. Наилучшие результаты при этом достигаются при использовании в качестве признаков частот встречаемости 500 наиболее частых триграмм русского языка, полученные после сглаживания всех триграмм аддитивным методом Лапласа.

Однако необходимый для точной идентификации объем текста (15000–20000 символов) пока слишком велик для решения большинства практических задач. Открытым также остается вопрос о влиянии параметров модели SVM на итоговую точность идентификации.

В дальнейших работах автором планируется продолжить тему поиска статистически устойчивых характеристик на малых текстовых фрагментах. Особое внимание планируется уделить определению авторства коротких электронных сообщений, а также затронуть тему применения предложенной методики для идентификации пола автора.

Работа выполнена при финансовой поддержке ФСРМПНТ.

Литература

1. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys. – 2002. – Vol. 34, № 1. – P. 1–47.

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

3. Vapnik V.N. Statistical Learning Theory. – New York: Wiley; 1998. – 732 p.

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

5. Hsu C.-W. A practical guide to support vector classification / C.-W. Hsu, C.-C. Chan, C.-J. Lin. [Электронный ресурс]. – Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, свободный.

6. Васильев В.Г. Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). – Вып. 7(14). – М.: РГГУ, 2008. – С. 83–91.

7. Романов А.С. Подходы к идентификации авторства текста на основе *n*-грамм и нейронных сетей // Молодежь и современные информационные технологии: Сб. тр. VI Всерос. науч.-практ. конф. студентов, аспирантов и молодых ученых, Томск, 26–28 февраля 2008 г. – Томск: Изд-во ТПУ, 2008. – С. 145–146.

8. Романов А.С. Структура программного комплекса для исследования подходов к идентификации авторства текстов // Докл. Том. гос. ун-та систем управления и радиоэлектроники. – 2008. – Ч. 1, № 2(18). – С. 106–109.

9. Романов А.С. Модель базы данных для хранения текстов и их характеристик // Докл. Том. гос. ун-та систем управления и радиоэлектроники. – 2008. – № 1(17). – С. 70–73.

10. Шаров С.А. Частотный словарь русского языка [Электронный ресурс]. – Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный.

11. Chen S.F. An empirical study of smoothing techniques for language modeling / S.F. Chen, J. Goodman // Computer Speech & Language. – 1999. – Vol. 13, № 4. – P. 359–393.

12. Библиотека Максима Мошкова [Электронный ресурс]. – Режим доступа: http://www.lib.ru, свободный.

13. Hsu C.-W. A comparison of methods for multi-class support vector machines / C.-W. Hsu, C.-J. Lin // IEEE Transactions on Neural Networks. – 2003. – № 13(2). – P. 415–425.

Романов Александр Сергеевич
ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», аспирант каф. комплексной информационной безопасности электронно-вычислительных систем
Эл. адрес: ras@ms.tusur.ru

A.S. Romanov

Authorship identification technique on basis of support vector machine

In the article authorship identification problem in the case of the limited set of alternatives is posed as classification task. It is suggested to use support vector machine and text's *N*-gramm features for solving the problem. Classification accuracy improving based on smoothing techniques is considered. Results of experiments that confirm theoretical assumptions are given.

Key words: authorship identification, classifire, text features, support vector machine, smoothing.