Идентификация автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив

Authorship identification with support vector machine in case of two possible alternatives

Романов А. С. (ras@ms.tusur.ru), **Мещеряков Р. В.** (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск

В статье проблема идентификации автора текста рассматривается как задача классификации. Обоснована важность решения задачи бинарной классификации для идентификации автора. Приведено описание и результаты экспериментов по идентификации автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив.

1. Постановка задачи

Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующем образом. Имеется множество текстов $T = \{t_1,\dots,t_k\}$ и множество авторов $A = \{a_1,\dots,a_n\}$. 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) $T'' = \{t_{|T'|+1},\dots,t_k\} \subseteq T$.

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 интерпретируется как степень принадлежности объекта классу: 1 соответствует положительному решению, 0 — отрицательному.

Задачу многоклассовой классификации можно свести к решению нескольких бинарных задач. Для этого существуют следующие основные стратегии выбора решения [3]:

- «Один против всех» (one-against-all). Для решения задачи строится n классификаторов таким образом, что каждый класс a_j сопоставляется с остальными $(n-1)$ классами, т.е. в каждом из j случаев выбор осуществляется из двух вариантов: «класс a_j» и «не класс a_j». Итоговое решение по всем классам принимается по схеме «победитель забирает всё» (winner takes all) — победителем считается класс, имеющий максимальное значение функции F.
- «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.
- Ориентированный ациклический граф (DNA). На этапе обучения работает также как стратегия «каждый против каждого». На этапе тестирования и непосредственной классификации используется корневой бинарный ориентированный ациклический граф (ориентированное дерево) с $n(n-1)/2$ внутренними узлами — обученными бинарными классификаторами, и n листьями. Классифицируемый объект проходит путь от корня до одного из листьев, при этом в зависимости от результатов классификации в каждом узле один из классов отвергается, и дальнейшие действия продолжаются по ветке, соответствующей второму классу. После выполнения $(n-1)$ подобных операций, алгоритм достигает листа, который принимается как итоговое решение классификатора.

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

2. Классификатор на основе машины опорных векторов

В исследованиях используется классификатор на основе метода «машины опорных векторов» (Support Vector Machine, SVM), математический аппарат которого был предложен В.Н. Вапником в работах [4, 5] и одна из его популярных реализаций — библиотека libsvm [6]. Исследования отечественных и зарубежных авторов [1, 7] показывают, что SVM, на сегодняшний день, является одним из лучших методов классификации.

Пусть имеется помеченное тренировочное множество примеров $D = \{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in X \subset R^d$, метки могут принимать значения $y_i \in Y = \{-1,+1\}$. SVM строит линейный классификатор в пространстве признаков с высокой размерностью таким образом, чтобы зазор между граничными точками двух классов, называемых опорными векторами, был максимальным. Для отображения исходных данных в пространство, в котором разделяющая их поверхность будет линейной, используются ядровые преобразования — некоторая функция

$$(\Phi(x), \Phi(x')) = k(x, x').$$

Классифицирующая функция, реализуемая SVM, записывается следующим образом:

$$f(x) = \left\{ \sum_{i=1}^{\ell} \alpha_i y_i k(x_i, x) \right\} + b$$

Чтобы найти оптимальный коэффициент α достаточно максимизировать функционал

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

в положительном квадранте $0 \le \alpha_i \le C$, $i = \overline{1, \ell}$. Условие максимизации:

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0.$$

Параметр регуляризации *C* отвечает за соотношение между величиной зазора и количеством ошибок обучающего множества.

Необходимыми условиями решения задачи нелинейного программирования являются условия Каруша-Куна-Такера:

$$\alpha_i = 0 \Rightarrow y_i f(x_i) \ge 1,$$
$$0 < \alpha_i < C \Rightarrow y_i f(x_i) = 1,$$
$$\alpha_i = C \Rightarrow y_i f(x_i) \le 1.$$

Эти условия удовлетворяют множеству допустимых множителей Лагранжа $\alpha^0 = \{\alpha_1^0, \alpha_2^0, ..., \alpha_\ell^0\}$, максимизирующих целевую функцию $W(\alpha)$. Параметр смещения *b* выбирается, чтобы обеспечить выполнение второго условия Каруша-Куна-Такера для всех входных образцов, соответствующих множителям Лагранжа, лежащим не на границах. В общем случае только часть множителей Лагранжа α будет иметь ненулевые значения, они и составляют опорные вектора.

Пусть *I* — множество индексов образцов, относящихся к множителям Лагранжа, лежащим внутри границ:

$$I = \{i : 0 < \alpha_i^0 < C\},$$

а *J* множество индексов со значениями множителей Лагранжа, лежащих на верхней границе *C*:

$$J = \{i : \alpha_i^0 = C\},$$

тогда можно переписать следующим образом:

$$f(x) = \left\{ \sum_{i \in \{I,J\}} \alpha_i^0 y_i k(x_i, x) \right\} + b.$$

В отличие от искусственных нейронных сетей, применявшихся авторами ранее [8], 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 интегрирована в общую программную оболочку, описанную в работе [9].

3. Описание экспериментов

Для оценки точности классификатора в случае двух предполагаемых авторов были проведены эксперименты на корпусе, составленном из 47 текстов 11 русских писателей (см. табл. 1), взятых с Интернет-ресурса [10]. Количество обучающих примеров 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. В табл. 2 представлены признаки текста (частоты встречаемости тех или иных групп символов и слов), использованные в экспериментах.

Таблица 1. Корпус текстов для исследований

Автор	Название произведения
Айтматов Ч. Т.	«Белое облако Чингисхана»
	«Пегий пес, бегущий краем моря»
	«Плаха»
	«Прощай, Гульсары!»
Акунин Б.	«Азазель»
	«Пелагея и Бульдог»
	«Внеклассное чтение»
	«Статский советник»
Астафьев В. П.	«Печальный детектив»
	«Так хочется жить»
	«Царь-рыба»
	«Ода русскому огороду»
	«Жизнь прожить»
Беляев А. Р.	«Голова профессора Доуэля»
	«Остров погибших кораблей»
	«Светопреставление»
	«Последний человек из Атлантиды»
	«Человек Амфибия»
Булгаков М. А.	«Мастер и Маргарита»
	«Собачье сердце»
	«Театральный роман»
	«Белая гвардия»
Достоевский Ф. М.	«Братья Карамазовы»
	«Преступление и наказание»
	«Идиот»
	«Бесы»
Горький М.	«Дело Артамоновых»
	«Мать»
	«Фома Гордеев»
	«Коновалов»
Тургенев И. С.	«Дворянское гнездо»
	«Отцы и дети»
	«Накануне»
	«Рудин»
Набоков В. В.	«Лолита»
	«Защита Лужина»
	«Дар»
	«Король, дама, валет»
Распутин В. Г.	«Деньги для Марии»
	«Пожар»
	«Прощание с Матерой»
	«Живи и помни»
Булычев К.	«Лиловый шар»
	«Любимец»
	«Марсианское зелье»
	«Подземелье ведьм»
	«Смерть этажом ниже»

Таблица 2. Исследованные признаки текста

Обозначение признака	Расшифровка
УНИГРАММЫ	Буквы русского алфавита
УСЛОВНЫЕ	Условные вероятности появления одной буквы после другой
БИГРАММЫ	Пары букв русского алфавита
БИГРАММЫ_ГЛ	Биграммы, состоящие только из гласных
БИГРАММЫ_СГЛ	Биграммы, состоящие только из согласных
БИГРАММЫ_ВЧ	Биграммы с высокой частотой встречаемости
БИГРАММЫ_СЧ	Биграммы со средней частотой встречаемости
БИГРАММЫ_НЧ	Биграммы с низкой частотой встречаемости
БИГРАММЫ_100	100 наиболее частых биграмм
ТРИГРАММЫ	Тройки букв русского алфавита
ТРИГРАММЫ_100	100 наиболее частых триграмм
ТРИГРАММЫ_500	500 наиболее частых триграмм
ТРИГРАММЫ_1000	1000 наиболее частых триграмм
ТРИГРАММЫ_2000	2000 наиболее частых триграмм
ТРИГРАММЫ_3000	3000 наиболее частых триграмм
ТРИГРАММЫ_ВЧ	Триграммы с высокой частотой встречаемости
ТРИГРАММЫ_СЧ	Триграммы со средней частотой встречаемости
ТРИГРАММЫ_НЧ	Триграммы с низкой частотой встречаемости
ШАРОВ	Частоты всех слов из словаря Шарова [11]
ФОМЕНКО	Частоты «опорных слов» Фоменко [12]
ШАРОВ_100	100 наиболее частых слов из словаря Шарова
ШАРОВ_500	500 наиболее частых слов из словаря Широва
ШАРОВ_1000	1000 наиболее частых слов из словаря Шарова
ШАРОВ_2000	2000 наиболее частых слов из словаря Шарова

При использовании метода на основе частотного словаря С. А. Шарова все слова были приведены к нормальной форме с помощью алгоритма стемминга Snowball для русского языка [13].

Для получения характеристик с разделением по частоте встречаемости был проведен их частотный анализ для всех имеющихся текстов. Би-

граммы, триграммы и слова были упорядочены по частоте встречаемости в убывающем порядке. Часть биграмм и триграмм символов была отсеяна как нехарактерная для русского языка и как шумы, связанные с автоматической обработкой текстов. Границами для характеристик с высокой частотой (ВЧ) выбраны квантили уровней 0,66 и 1, для характеристик с низкой частотой (НЧ) — квантили уровней 0 и 0,33, и квантили уровней 0,33 и 0,66 — для характеристик со средней частотой (СЧ).

Параметры обучения моделей SVM 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 по двум авторам 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» заменялась 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 пар классов из всего множества авторов (в данном случае количество пар классов равно $C_{11}^2 = 55$).
4. Для каждого автора из текущей пары формируется по 3 обучающих выборки необходимого объема и одна тестовая. Выборки извлекаются из разных текстов автора.
5. Подсчет 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 следующей пары классов.

Для каждого объема выборки было проведено по 280 экспериментов (учитывались все сочетания авторов и текстов). В качестве результирующей оценки точности по данному признаку и объему выборки подсчитывалась средняя частота правильных классификаций.

4. Результаты экспериментов, обсуждение, выводы

По сформулированной выше методике были проведены эксперименты, результаты которых представлены на рис. 1–3.

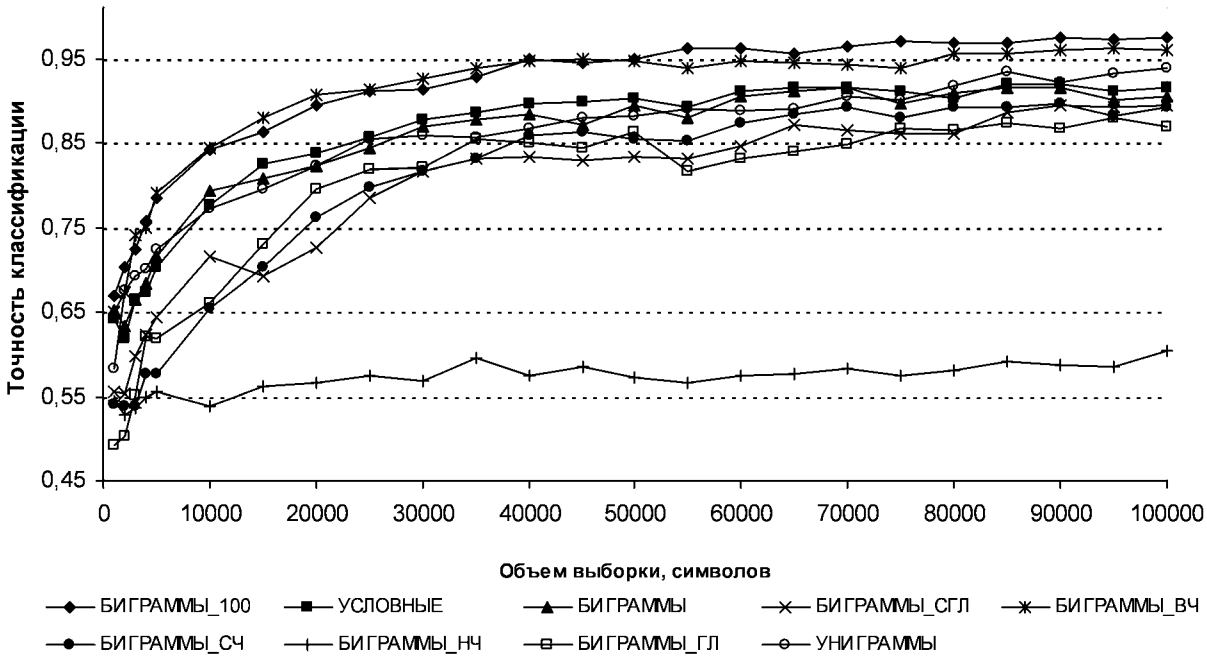

Рис. 1. Результаты исследований по унограмммам и биграммам символов

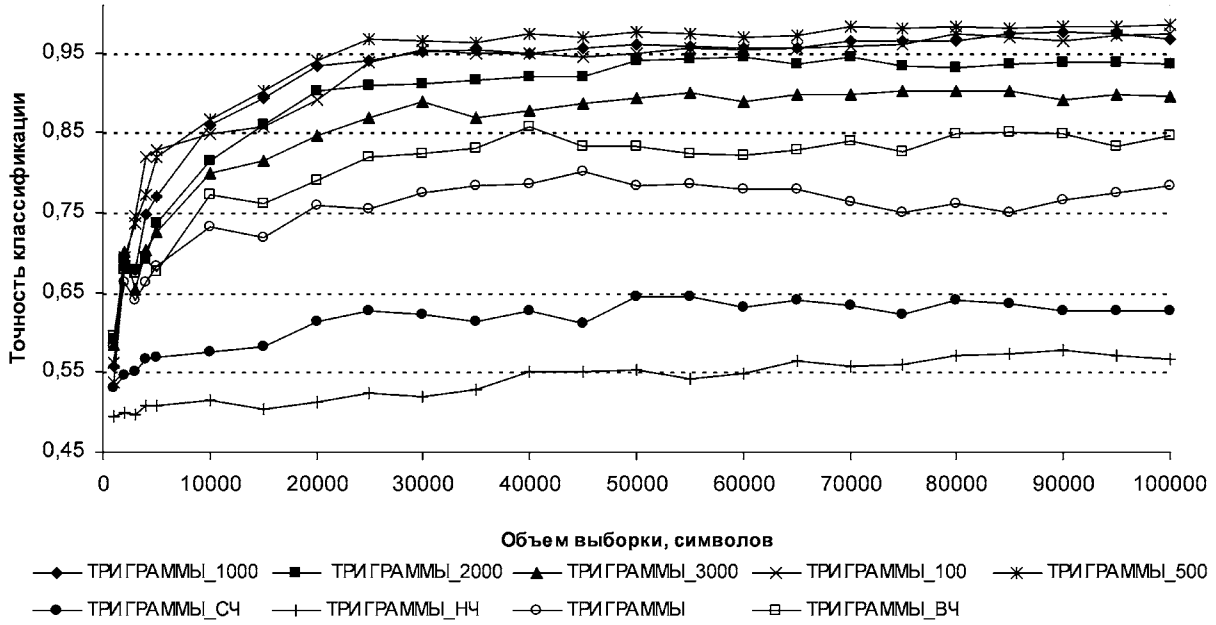

Рис. 2. Результаты исследований по триграммам символов

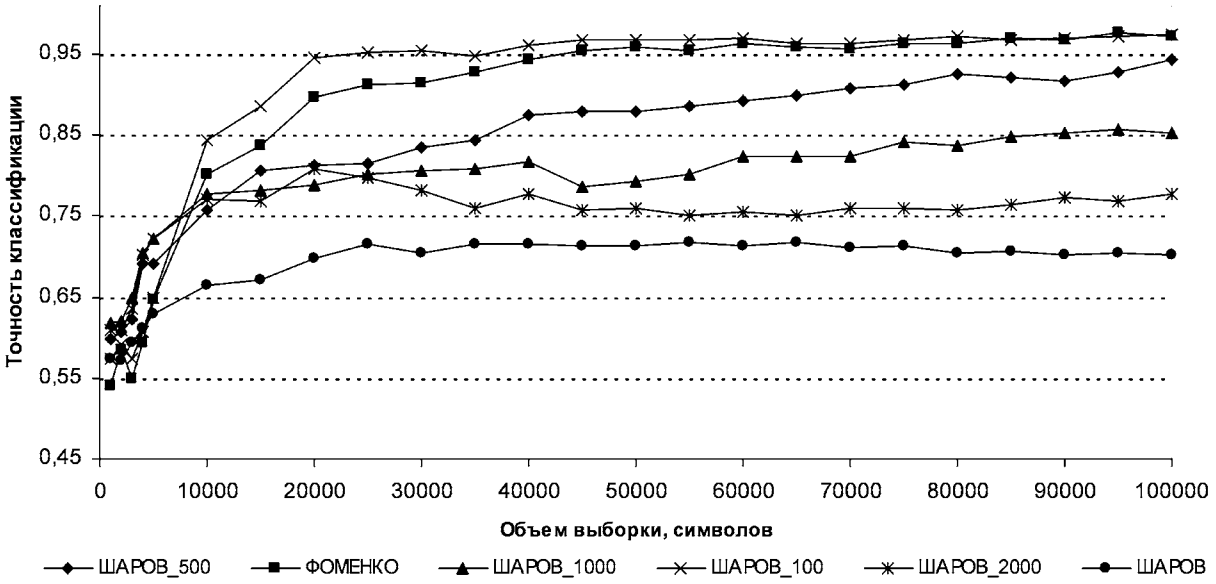

Рис. 3. Результаты исследований по частотному словарю русского языка

Из первой группы признаков (см. рис.1) наиболее точно классификация проходит при использовании признаков БИГРАММЫ_100 и БИГРАММЫ_ВЧ. Значение точности стабилизируется при объеме выборки равном 40000 символов и далее колеблется около 0,96. Дальнейшее увеличение количества признаков и использование биграмм со средними и низкими частотами встречаемости ведет к снижению точности, к аналогичным результатам приводит использование биграмм, составленных отдельно из гласных и согласных букв (средняя точность 0,77). Приблизительно одинаковые результаты дает использование признаков БИГРАММЫ и УНИГРАММЫ (средняя точность 0,83), немного выше точность классификации по признаку УСЛОВНЫЕ — 0,84.

Среди признаков, основанных на триграммах символов (см. рис.2), наиболее точной оказывается классификация по признаку ТРИГРАММЫ_500, стабилизация наступает при объеме выборки равной 25000 и далее точность колеблется около 0,97. При увеличении размерности признаков, аналогично экспериментам с биграммами символов, наблюдается снижение качества классификации.

Использование 100 наиболее частых слов русского языка предпочтительнее предложенного Фоменко набора служебных слов за счет более высо-

кого качества классификации на выборках любого объема, средняя точность классификации соответственно 0,87 и 0,86 для признаков ШАРОВ_100 и ФОМЕНКО (см. рис. 3). Стабилизация при использовании признака ШАРОВ_100 наступает при объеме выборки равном 20000 символов и далее точность колеблется около 0,96. Повышение количества признаков до 500 и более снижает качество классификации.

В целом по результатам экспериментов можно сделать вывод, что идентификация автора с помощью аппарата SVM в случае двух альтернатив возможна при объеме выборки 20000 символов и больше. При этом нецелесообразно использовать более 500 признаков.

В известных авторам работах по автоматическому определению авторства текста на русском языке приводятся результаты исследований при количестве классов, равном 10 и более, задача идентификации автора в случае двух возможных альтернатив не рассматривается. Сравнение же с аналогичными исследованиями для других языков проводить некорректно в силу особенностей строения каждого языка.

Высокая точность бинарной классификации позволит в дальнейшем применять наиболее эффективные наборы признаков для идентификации автора в случае трех и более предполагаемых авторов.

Однако необходимый для точной идентификации объем текста пока слишком велик для решения большинства практических задач. В дальнейших работах авторами планируется исследовать техники сглаживания [14] для уменьшения требуемого объема выборок, а также продолжить тему поиска статистически устойчивых характеристик на малых текстовых фрагментах и провести эксперименты с наиболее эффективными характеристиками на более представительном корпусе текстов.

Работа поддержана грантом 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: Учебное пособие. Томск:ТМЛ-Пресс, 2007. 144 с.
3. *Hsu C.-W., Lin C.-J.* A comparison of methods for multi-class support vector machines // IEEE Transactions on Neural Networks, 2003. № 13(2). P. 415–425.
4. *Vapnik V. N.* Statistical Learning Theory // Wiley, New York, 1998. 732 pages.
5. *Vapnik V. N.* The nature of statistical learning theory // Springer-Verlag, New York, 2000. 332 pages.
6. *Hsu C.-W., Chang C.-C., Lin C.-J.* A practical guide to support vector classification. — Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/ guide.pdf, свободный.
7. *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.
8. *Романов А. С.* Подходы к идентификации авторства текста на основе n-грамм и нейронных сетей // Молодежь и современные информационные технологии. Сборник трудов VI Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых. Томск, 26–28 февраля 2008 г. Томск: Изд-во ТПУ, 2008. С.145–146.
9. *Романов А. С.* Структура программного комплекса для исследования подходов к идентификации авторства текстов // Доклады Томского государственного университета систем управления и радиоэлектроники. Томск: Изд-во ТУСУР, 2008. Ч. 1. №2(18). С. 106–109.
10. *Библиотека* Максима Мошкова. — Режим доступа: http://www.lib.ru, свободный.
11. *Шаров С. А.* Частотный словарь русского языка. — Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный.
12. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов // Фоменко А. Т. Новая хронология Греции: Античность в средневековье. М.: Изд-во МГУ, 1996. Т. 2. С. 768–820.
13. *Porter M. F.* Russian stemming algorithm. — Режим доступа: http://snowball.tartarus.org/algorithms/ russian/stemmer.html, свободный.
14. *Katz S. M.* Estimation of probabilities from sparse data for the language model component of a speech recognizer // IEEE Transactions on Acoustics, Speech and Signal Processing, 1987. № 35(3). P. 400–401.