УДК 517.977.56 517.977.58

Айда-заде К.Р.[1], Талыбов С.Г.[2]
Институт Систем Управления НАНА, Баку, Азербайджан
[1]Kamil_ayda-zade@rambler.ru, [2]saxavat@yahoo.com

АНАЛИЗ МЕТОДОВ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВ НА АЗЕРБАЙДЖАНСКОМ ЯЗЫКЕ

В статье проведен анализ методов, алгоритмов, используемых для распознавания авторства текстов. Применяемые признаки распознавания основаны на n-граммах при n=1 и n=2. Приводятся результаты компьютерных экспериментов по распознаванию авторства текстов на азербайджанском языке.

Ключевые слова: *идентификация, идентификация автора, распознавание, n-грамм, метод опорных векторов.*

1. Введение

Как известно, одной из важных проблем обработки текстов является их классификация по авторам, т.е. определение, кто из заранее заданной группы авторов является предполагаемым автором конкретно данного текста.

Автоматизацией решения этой проблемы наиболее интенсивно начали заниматься в 70-х годах прошлого века. Первоначально методы решения этой проблемы базировались на использовании созданных специальных глоссариев для ключевых слов.

Одним из первых байесовский анализ для решения проблемы распознавания авторства использовал Mosteller [1]. Далее Burrows в работе [2] использовал частоты наиболее часто используемых слов авторами, Brinegar [3] использовал длины слов, Morton [4] − длины предложений, Brainerd [5] − среднее число слогов, Holmes [6] − количество используемых слов и объем документа, Twedie [7] − отношение числа используемых слов к общему количеству слов в тексте, *n*-грамм (2-грамм и 3-грамм) − Фюрнкранц [9] и Тан [10]. Çatal [11] создал систему идентификации NECL. В работах [18, 19] для распознавания авторов русской художественной литературы была использована частота встречаемости букв и буквосочетаний.

В работах [12, 13] впервые для распознавания авторства азербайджанских текстов исследована частота использования букв и длины слов, но, тем не менее, до сих пор компьютерной системы по распознаванию авторства текстов на азербайджанском языке нет. В данной работе изучается проблема идентификации авторства на основе анализа авторских статей небольшого объема. Основная трудность распознавания авторства текстов (статей) малого объема на азербайджанском языке заключается в том, что в словах используется большое количество малоинформативных суффиксов, окончаний, а автоматический разбор слов на составные части для азербайджанского языка до сегодняшнего дня остается нерешенной проблемой.

2. Постановка задач

Формально постановку задачи идентификации авторства текстов можно описать следующим образом.

В базе данных имеются тексты n авторов и от каждого из них m_i текстов $D_{i,j}$, $j = 1, ..., m_i$, $i = 1, ..., n$. Класс (группу) текстов i -го автора обозначим через Y_i. Рассматриваемая в статье задача состоит в том, что при появлении нового текста D требуется определить, какому из n авторов или, другими словами, к какому классу Y_i эта работа принадлежит.

Введем следующие обозначения, определения и формулы.

Каждому из текстов $\mathcal{D}_{i,j}$ и D сопоставим множество значений признаков $\{M_{i,j}^s, s \in K_i\}$ и $\{d_s,\ s \in K_i,\}, K = \bigcap_{s=1}^{n} K_s,\ \ i = 1, \dots, n, j = 1 \dots, m_i,$ на основе которых происходит классификация текстов по авторам. Здесь K_i – множество признаков для определения авторства *i*-го автора, $i = 1, \dots, n$.

Пусть $N_{i,j}$ − длина (объем) *j*-го текста *i*-го автора, m_i − количество статей *i*-го автора, находящихся в базе данных, тогда ясно, что среднее значение *s*-го признака *i*-го автора в *j*-й статье определяется формулой:

$$\varepsilon_{i,j}^s = \frac{M_{i,j}^s}{N_{i,j}}, \quad s \in K_i, j = 1, \dots, m_i,\ \ i = 1, \dots, n, \tag{1}$$

среднее значение *s*-го признака во всех статьях *i*-го автора равно

$$\xi_i^s = \frac{\sum_{j=1}^{m_i} M_{i,j}^s}{\sum_{j=1}^{m_i} N_{i,j}}, \ \ s \in K_i, i = 1, \dots, n. \tag{2}$$

Пусть среднее значение *s*-го признака в новой статье *D* равно

$$x_D^s = \frac{m_D^s}{N_D}, \qquad s \in K_i. \tag{3}$$

Здесь m_D^s – значение *s*-го признака в новой статье *D*, а N_D – ее длина (объем).

Очевидно, что дисперсия *s*-го признака для *i*-го автора равна

$$(d_i^s)^2 = \frac{\sum_{j=1}^{m_i} (M_{i,j}^s - \xi_i^s)^2}{\sum_{j=1}^{m_i} N_{i,j}}, \ \ s \in K_i, \ \ i = 1, \dots, n. \tag{4}$$

Вариация *s*-го признака для *i*-го автора равна

$$v_i^s = \frac{d_i^s * 100}{\xi_i^s}, \quad s \in K_i, \ \ i = 1, \dots, n. \tag{5}$$

Рассмотрим величину

$$R_i = \sum_{s \in K} \alpha_s (x_D^s - \xi_i^s)^2, i = 1, \dots, n, \tag{6}$$

определяющую близость (норму) значений признаков нового текста *D* к значениям признаков, характеризующих *i*-го автора; α_s – вес (важность) *s*-го признака для определения авторства статей.

3. Классификация компьютерных систем идентификации авторов

Известны следующие типы компьютерных систем идентификации авторов текстов:

- *системы, не учитывающие язык текстов;*
- *системы, учитывающие язык текстов;*
- *комбинированные системы распознавания с иерархической структурой.*

Существующие системы идентификации, не учитывающие язык текстов, как правило, используют следующие признаки: длины предложений; частоту использования абзацев; длины слов; длины тем подзаголовков; средний размер абзацев; количество знаков препинания; частоту наиболее часто используемых автором слов; суффиксы, наиболее часто используемые автором, и др.

Многие существующие компьютерные системы идентификации, не учитывающие язык написания текстов, основаны на использовании методов, анализирующих частоту использования всевозможных буквосочетаний, состоящих из *n*-букв *n*-грамм.

Использование *n*-грамм для многих языков имеет свои трудности, особенно для контекстно свободных языков и языков, у которых корень слова может использоваться с

несколькими суффиксами, например, как в языках тюркской (турецкий, азербайджанский, узбекский и т.д.), славянской (русский, белорусский, польский, словацкий и т.д.) групп. Например, в словах на азербайджанском языке для словообразования могут быть использованы префикс, суффикс и так далее. Использование в компьютерных системах наиболее часто употребляемых автором слов для идентификации авторства требует отделения корня от суффиксов. Но, к сожалению, до сих пор, как было сказано выше, для азербайджанского языка эта проблема не решена, а для идентификации автора текста важно игнорировать артикли, разделители, специальные слова, цифры и наиболее часто используемые слова (stop words).

Системы идентификации, учитывающие язык текстов, как правило, используют следующие признаки: частота использования stop word в тексте; частота использования определенных частей речи (например, существительных, местоимений и т.д.); частота наиболее часто используемых автором суффиксов и префиксов и другие.

Комбинированные системы иерархической структуры идентификации создаются на основе проведения нескольких этапов (уровней) распознавания. Если текст не художественный, то эффективность использования выделения суффиксов для одного языка может оказаться неправильной для другого языка и дать совершенно ошибочный результат. Например, в случае английского или немецкого языка отбрасывание префикса не столь существенно влияет на распознавание авторства текста, но это имеет большое значение, в частности для азербайджанского и турецкого языков.

Следует отметить, что используемые конкретные акценты и диалекты языка также являются ключевыми факторами, влияющими на идентификацию автора, хотя в некоторых случаях (например, для художественных и научных произведений) использование акцентов и диалектов не имеет большого значения. Из опыта создания систем идентификации авторства текстов известно, что использование комбинированных алгоритмов иерархической структуры может существенно повысить эффективность системы распознавания авторства текстов.

4. Используемые методы и алгоритмы распознавания авторства текстов

В статье описываются разработанные алгоритмы, не использующие специфику языка, и приводятся результаты их работы по определению авторства статей, взятых с газетных и информационных сайтов на азербайджанском языке.

Следует отметить, что серьезной проблемой получения качественных результатов распознавания авторства этими алгоритмами являлось предположение о небольшом объеме имевшихся статей каждого из авторов. При этом большая часть информации, содержащаяся в доступных статьях, как правило, относилась к конкретным рассматриваемым темам, т.е. использовались специфические слова, термины, определяемые конкретной тематикой, и, следовательно, это не является достаточно информативным для распознавания авторов.

Алгоритмы функционирования методов идентификации авторства в общем случае включают выполнение следующей последовательности процессов:

- проводится первичная обработка имеющихся текстов (статей, произведений) различных авторов, для каждого автора определяются числовые значения выбранных признаков;
- проводится анализ значений признаков и определяется множество информативных признаков для каждого автора (множества признаков могут быть неодинаковыми для разных авторов);
- определяются значения признаков представленной новой статьи пока неизвестного автора;
- по определенному критерию на основе известных алгоритмов определяется предполагаемый автор представленной статьи.

Приведем признаки авторства, основанные на статистическом анализе буквосочетаний. Отметим, что многие известные алгоритмы и системы распознавания авторства текстов применяют признаки, основанные на анализе использования различных буквосочетаний из *n*-букв, называемых в литературе n-граммами. Таким образом, в данном случае грамм означает, что в качестве единицы берется одна буква, при этом слова, предложения или абзацы всего текста в зависимости от значения *n* разбиваются на буквосочетания, содержащие *n* последовательных букв, *n=1,2,….* Например, фраза «Баку − столица» монограммами (*n* равно 1 букве) и диаграммами (*n* равно *2* буквам) соответственно разбивается следующим образом:

«Б», «а», «k», «у», «_»,«с», «т», «о», «л», «и», «ц», «а» ,

«Ба», «ak», «ky», «у_», «_с», «ст», «то», «ол», «ли», «иц», «ца».

Здесь «_» – знак пробела.

Отметим, что число букв, а следовательно, 1-граммов в азербайджанском языке 32, а практическое число всевозможных 2-граммов равно 835. Были реализованы следующие три алгоритма, использующие признаки, основанные на *n*-граммах.

Приведем общее описание алгоритма 1, использующего монограммы.

Шаг 1. Для каждого текста (статьи) i-го автора, входящего в класс Y_i, в качестве признаков по формуле (1) определяются частоты использования всех букв алфавита (1-граммы).

Шаг 2. Объединяя все произведения каждого автора в соответствии с формулой (2), рассчитываются средние значения всех признаков.

Шаг 3. Для новой исследуемой статьи D по формуле (3) рассчитывается вектор значений x_D^s признаков, $s \in K$.

Шаг 4. В формуле (6), взяв $\alpha_s = 1$ (все веса равны 1), определяем такое ν, что $R_\nu = \min\limits_{1 \le i \le n} R_i$, следовательно, автором статьи D является ν -й автор.

Использованный нами алгоритм 2 на базе диаграмм такой же, как алгоритм 1 для монограмм, но анализируется не частота использования отдельных букв, а частота использования всевозможных комбинаций из двух букв алфавита (2-грамм − диаграмм), используемых автором в своих статьях.

Приведем описание модифицированного алгоритма 3, использующего монограммы. Основная идея алгоритма, основанного на модифицированных монограммах, состоит в том, что в этом случае используются устойчивые признаки, не включающие нехарактерные для автора наборы признаков.

В предлагаемом алгоритме для каждого отдельного автора с помощью формулы (5) вычисляется вариация каждого признака. Вес *s*-го признака α_s в формуле (6) выбирается в зависимости от значения вариации *s*-го признака по всем авторам следующим образом. Обозначим:

$$v^s = \begin{cases} \min\limits_i v_i^s, & \min\limits_i v_i^s > \varepsilon, \\ \varepsilon, & \min\limits_i v_i^s \le \varepsilon, \end{cases} \qquad s \in K.$$

Положительная величина ε выбирается, исходя из величины значений нехарактерных для авторов признаков. Тогда

$$\propto_s = \frac{(v^s)^{-1}}{\sum_{j=1}^{k}(v^j)^{-1}}, \quad s \in K.$$

Ясно, что $\propto_s$ удовлетворяют условиям:

$$0 \le \propto_s \le 1 \quad , \quad \sum_{s \in K} \propto_s = 1.$$

Первые два шага предлагаемого алгоритма совпадают с двумя шагами первого алгоритма.

Шаг 3. Используя (4) и (5) для каждого признака класса Y_i, на основании вариации проверяется устойчивость признаков и устанавливаются значения весов α_s.

Шаг 4. Для новой исследуемой статьи D по формуле (3) рассчитываются значения признаков x_D^s, $s \in K$.

Шаг 5. Определяем ν, при котором $R_\nu = \min_i R_i$, следовательно, автором статьи $\mathcal{D}$ является ν -й автор.

5. Результаты компьютерных экспериментов

Для проверки и сравнения эффективности вышеизложенных алгоритмов для их обучения в базу данных были включены взятые из Интернета 32 (первый вариант выборки) и 50 (второй вариант выборки) газетные информационные статьи на азербайджанском языке случайно выбранных четырех авторов, условно названных *A1, A2, A3, A4*. Идентификация авторов проводилась соответственно по 5 и 8 дополнительным статьям одного из четырех авторов, авторство которых было скрыто.

В качестве признаков в случае применения монограммы (n=1) использовались 32 буквы, в случае диаграмм (n=2) использовались 835 реально возможных для азербайджанского языка сочетаний букв.

5.1. Результаты экспериментов для первого варианта объема обучения

Для обучения были рассмотрены 9 статей автора A1, 7 статей автора A2, 6 статей автора A3 и 10 статей автора A4. Общее количество букв в каждой статье составляло от 3438 до 6859 символов.

Для распознавания были взяты 1 статья z^1 автора A1 (число букв в статье 3926), 1 статья z^2 автора A2 (число букв 3470), 2 статьи $z^3 = (z_4^3,\ z_2^3)$ автора A3 (число букв 4067 и 4243) и 1 статья z^4 автора A4 (число букв 7463).

В первых четырех столбцах таблицы 1 приведены результаты работы алгоритма 1, использующего признаки, основанные на монограммах. В j-й строке i-го столбца приведено значение $R_i(z^j) * 10^3$. Ясно, что правильное распознавание авторства j-й статьи соответствует случаю, когда значение j-го элемента является наименьшим среди элементов j-й строки (соответствующие значения подчеркнуты).

Таблица 1

	Алгоритм 1 (n=1)				**Алгоритм 3 (n=1)**				**Алгоритм 2 (n=2)**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z^1	***108***	126	124	140	**423**	474	441	494	**164**	341	296	299
z^2	109	**086**	120	075	433	**385**	439	403	408	**238**	295	311
z_1^3	**078**	093	090	118	388	435	**377**	473	339	283	**237**	282
z_2^3	**163**	194	172	233	441	480	**437**	543	412	369	**314**	394
z^4	113	093	134	**083**	432	399	462	**332**	364	304	297	**272**

Как видно из таблицы 1, было правильно установлено лишь авторство статей первого и четвертого авторов, а качество распознавания составляло 60%.

Во второй группе из четырех столбцов таблицы 1 приведены результаты применения предлагаемого модифицированного алгоритма, основанного на монограммах и использующего веса для признаков. Как видно из таблицы 1, эффективность распознавания существенно улучшилась и равна 100%.

В третьей группе из четырех столбцов таблицы 1 приведены результаты работы алгоритма 2, использующего в качестве признаков диаграммы. Полученные результаты распознавания более устойчивы, т.е. для $R_\nu = \min\limits_{i \le i \le n} R_i$, и $j \neq \nu$ R_j существенно превышает R_ν , при этом эффективность распознавания равна 100%.

Отметим, что признаки, основанные на диаграммах для русского языка, впервые использовал Д.Хмелев. Он рассмотрел последовательность букв как реализацию марковской цепи. Исследовав соединения букв (диаграммы) в литературных произведениях 82 авторов на русском языке, экспериментально выявил наиболее характерные признаки для конкретного автора [19]. Результаты исследования показали, что и для азербайджанского языка этот подход более эффективен. В этом случае авторство распознаваемых авторов статей было определено со 100-процентной точностью.

В таблице 2 приводятся результаты проверки устойчивости признаков-монограмм, основанных на анализе значений вариации этих признаков, рассчитанных для всех текстов выбранных авторов**.** Ясно, что чем больше значение вариации для какого-либо признака, тем менее надежно использование этого признака в распознавании авторства.

В этом случае выяснилось, что число существенных признаков, характеризующих авторов A1, A2, A3 и A4, соответственно равно 18, 20, 20 и 20 и эти признаки для разных авторов различны.

Использование модифицированных признаков позволило для всех авторов рассмотренных статей определить авторство со 100-процентной точностью.

В этом случае были отброшены признаки с высокой вариацией (в нашем случае выше 20%), которые недостаточно характеризуют автора.

5.2. Результаты экспериментов с использованием второй обучающей выборки

Рассмотрены газетные статьи автора A1 в количестве 13, автора A2 соответственно 11, автора A3 − 12 и автора A4 − 14. Общее количество букв в колонках записи составляет от 3438 до 6859.

Для распознавания у каждого представленного автора были взяты по две статьи.

В первых четырех столбцах таблицы 3 приведены результаты работы алгоритма 1, использующего признаки, основанные на монограммах. В *j*-й строке *i*-го столбца приведено значение $R_i(z^j) * 10^3$. Ясно, что правильное распознавание авторства *j*-й статьи соответствует случаю, что значение *j*-го диагонального элемента является наименьшим среди элементов *j*-й строки (при правильном распознавании значение $R_i(z^j)$ подчеркнуто одной чертой снизу, при неправильном − одной чертой сверху).

Как видно из таблицы 3, с применением алгоритма 1 было правильно установлено лишь авторство статей четвертого автора, а в целом качество распознавания составляло 50%.

Во второй группе из четырех столбцов таблицы 3 приведены результаты применения предлагаемого модифицированного алгоритма, основанного на монограммах и использующего для признаков весовые коэффициенты.

Как видно из таблицы 3, эффективность распознавания существенно улучшилась и равна 62,5%.

В третьей группе из четырех столбцов таблицы 3 приведены результаты работы алгоритма 2, использующего в качестве признаков диаграммы.

Таблица 2

Буква	A1		A2		A3		A4	
	Средняя оценка	Вари-ация	Средняя оценка	Вари-ация	Средняя оценка	Вари-ация	Средняя оценка	Вари-ация
A	0.1060	5.54	0.1057	6.46	0.1068	4.09	0.1048	5.34
B	0.0275	***22.06***	0.0293	12.13	0.0327	5.12	0.0216	6.90
C	0.0091	***22.27***	0.0088	16.16	0.0098	***28.12***	0.0104	19.86
D	0.0545	17.49	0.0475	12.45	0.0544	6.50	0.0466	9.77
E	0.0244	16.06	0.0241	8.76	0.0242	16.24	0.0254	16.84
F	0.0058	***26.80***	0.0041	10.62	0.0055	***27.30***	0.0068	***26.75***
G	0.0079	***39.87***	0.0074	***20.48***	0.0080	***27.49***	0.0061	***21.26***
H	0.0144	***30.08***	0.0122	19.38	0.0120	4.12	0.0099	***30.29***
I	0.0380	***22.34***	0.0366	12.34	0.0388	6.56	0.0323	8.51
J	0.0005	***106.4***	0.0005	***39.6***	0.0004	***77.70***	0.0003	***91.57***
K	0.0224	15.84	0.0214	11.98	0.0224	15.04	0.0245	15.57
L	0.0607	14.24	0.0599	8.83	0.0570	8.66	0.0613	5.4822
M	0.0427	15.05	0.0390	15.29	0.0416	7.01	0.0377	7.8337
N	0.0683	15.37	0.0781	7.66	0.0692	8.86	0.0742	10.82
O	0.0240	19.21	0.0203	11.79	0.0216	8.15	0.0174	11.38
P	0.0060	***34.72***	0.0046	***21.03***	0.0063	16.72	0.0062	***29.49***
Q	0.0196	16.88	0.0226	9.20	0.0183	7.16	0.0191	13.73
R	0.0662	15.70	0.0650	7.24	0.0735	1.55	0.0670	4.98
S	0.0309	***21.75***	0.0388	10.92	0.0292	6.86	0.0415	13.87
T	0.0325	19.58	0.0325	10.56	0.0293	3.29	0.0428	9.71
U	0.0271	***21.49***	0.0218	10.26	0.0272	14.87	0.0213	***21.30***
V	0.0123	18.92	0.0129	***22.83***	0.0116	19.42	0.0118	13.16
X	0.0103	***23.42***	0.0072	***30.90***	0.0109	***23.72***	0.0090	***20.05***
Y	0.0320	14.49	0.0328	16.24	0.0324	10.41	0.0335	7.83
Z	0.0178	***24.46***	0.0138	12.70	0.0204	***20.16***	0.0118	13.73
Ç	0.0085	19.08	0.0088	***22.24***	0.0089	12.07	0.0075	17.07
Ö	0.0079	***23.13***	0.0104	***32.95***	0.0090	23.53	0.0080	15.50
Ü	0.0172	18.50	0.0200	***20.90***	0.0215	16.55	0.0166	10.16
Ğ	0.0057	***28.57***	0.0070	17.02	0.0044	***34.08***	0.0057	***28.22***
İ	0.0922	7.03	0.0953	6.18	0.0892	10.33	0.1089	6.32
Ş	0.0147	14.16	0.0129	17.47	0.0156	18.15	0.0119	16.66
Ə	0.0930	6.13	0.0977	4.58	0.0876	3.81	0.0948	7.21

Таблица 3

	Алгоритм 1 (*n*=1)				**Алгоритм 3 (*n*=1)**				**Алгоритм 2 (*n*=2)**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1	**758**	930	825	1070	**3356**	3631	3625	3895	**2139**	2301	3755	2621
z_2^1	1020	925	**726**	892	3999	3966	**3167**	3607	2150	3684	**1962**	2379
z_1^2	1414	1084	1515	**1215**	4381	**3711**	4700	3770	3844	**2571**	5435	2773
z_2^2	1484	1360	1475	**1308**	4653	**4036**	4225	4155	3813	3666	**2660**	3276
z_1^3	**1495**	1825	1697	2118	**4086**	4634	4965	4987	3471	**2918**	4687	3289
z_2^3	2066	2204	**1909**	1936	4794	5240	**4510**	4981	3606	4214	**2761**	3237
z_1^4	1578	1085	1721	**981**	5005	4035	5123	**3059**	3525	2697	5702	**2223**
z_2^4	1573	1375	1442	**1032**	5641	4516	4616	**3524**	3946	4152	3770	**2674**

Как видно из таблицы 3, использование алгоритмов, основанных на монограммах, и алгоритма модифицированного метода монограммы для распознавания автора недостаточно эффективно. Алгоритм, основанный на монограммах, правильно определяет автора лишь в 50% случаев, а модифицированный алгоритм − в 62,5%.

Как видно из таблицы 3, использование диаграмм позволяет определить автора с точностью около 62,5%.

Анализ результатов экспериментов, приведенных в таблицах 1–3, показывает, что с увеличением объема как обучающей, так и проверяющей выборок результаты распознавания с применением статистических методов ухудшаются.

Сравнительно низкий процент распознавания авторов в двух вариантах объясняется в основном большой размерностью пространства признаков и близостью значений признаков друг к другу. Это связано с невозможностью отделить множества признаков, характеризующих каждого автора, из пространства признаков с обычными линейными гиперповерхностями. С другой стороны, это также связано с малым объемом записей в газетных статьях и низкой информативностью этих записей.

5.3. Результаты применения метода опорных векторов

Метод опорных векторов (SVM – Support Vector Machin) впервые был предложен В.Н.Вапником [21, 22]. Наиболее популярная версия этого метода реализована в пакете LIBSVM [23]. Отметим, что благодаря современным разработкам многих исследователей можно констатировать, что SVM в настоящее время является одним из эффективных методов классификации (упорядочивания). В отличие от нейронных сетей, применение SVM в многомерном пространстве признаков более эффективно, в частности, при применении в качестве признаков *n*-граммов (диаграмм и монограмм).

В качестве ядра для опорного вектора была выбрана функция gaussian radial basis function (RBF):

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}, \gamma \geq 0.$$

Численные эксперименты проводились с помощью пакета LIBSVM с применением признаков, основанных на монограммах и диаграммах.

В таблице 4 приведены результаты распознавания авторства с использованием в качестве признаков 1-грамм, в таблице 5 результаты использования признаков при *n=2.*

Таблица 4 (Monogram)

	C=1				C=10				C=100			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1				+	+				+			
z_2^1				+	+				+			
z_1^2				+		+				+		
z_2^2				+		+				+		
z_1^3				+			+				+	
z_2^3				+			+				+	
z_1^4				+				+				+
z_2^4				+				+				+

Для обучения и распознавания использовался вариант второй обучающей выборки. Значение параметра γ было взято γ =1. Величина параметра штрафа C выбиралась равной 1, 10 и 100. При малых значениях параметра C точность распознавания составила 60−70%.

Отметим, что описанные выше модифицированные алгоритмы, использующие монограммы и диаграммы**,** впервые предложены нами. Как показали численные эксперименты, модифицированные *n*-граммы были всегда более эффективны по сравнению с классическими вариантами *n*-грамм.

Таблица 5 (Digram)

	***C*=1**				***C*=10**				***C*=100**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1	+				+				+			
z_2^1	+				+				+			
z_1^2				+		+				+		
z_2^2				+		+				+		
z_1^3	+						+				+	
z_2^3	+						+				+	
z_1^4				+				+				+
z_2^4				+				+				+

Заключение

В статье проведен анализ методов алгоритмов распознавания авторства азербайджанских текстов.

Основными признаками являлись *n*-грамм при n=1 и n=2. Алгоритмы распознавания строились с использованием статистического подхода и метода опорных векторов.

Объектом распознавания были авторы газетных информационных статей, характеризуемых малым объемом.

В статье предложены модифицированные алгоритмы статистического метода распознавания, использующие величины вариаций *n*-грамм, характерных для статей рассматриваемых авторов.

Приведены результаты экспериментов по распознаванию авторов статей с применением разработанных алгоритмов, программ. Проведено сравнение полученных результатов с результатами применения метода опорных векторов, показавшего более надежное распознавание по сравнению с байесовским подходом.

Литература

1. Mosteller F., Wallace D.L. Applied Bayesian and Classical Inference, The Case of the Federalist Papers. Reading, MA: Addison-Wesley, 1984, 303 p.
2. Burrows J.F. Not unless you ask nicely: the interpretative nexus between analysis and information // Literary Linguist Computing, 1992, vol.7, no.2, pp.91–109.
3. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author // Computational Linguistics, 2001, vol. 26, no.4, pp.471–495.
4. Morton A.Q. The Authorship of Greek Prose // Journal of the Royal Statistical Society, Series A, 1965, vol. 128, no.2, pp.169–233.
5. Brainerd B. Weighting Evidence in Language and Literature // A Statistical Approach, University of Toronto Press, 1974, 288 p.
6. Holmes D.I. Authorship Attribution // Computers and The Humanities, 1994, vol.28, no.2, pp.87–106.
7. Tweedie F., Baayen H. How Variable may a Constant be Measures of Lexical Richness in Perspective // Computers and The Humanities, 1998, vol.32, no.5, pp.323–352.
8. Stamatatos E., Fakotakis N., Kokkinakis G. Computer-Based Authorship Attribution Without Lexical Measures // Computers and The Humanities, 2001, no.35, pp.193–214.
9. Fürnkranz J. A Study using n-gram Features for Text Categorization, Austrian Research Institute for Artifical Intelligence, 1998, 10 p.
10. Tan C.M., Wang Y.F., Lee C.D. The Use of Bigrams to Enhance // Journal Information Processing and Management, 2002, vol.30, no.4, pp.529–546.
11. Çatal Ç., Erbakırcı K., Erenler Y. Computer-based Authorship Attribution for Turkish Documents / Turkish Symposium on Artificial Intelligence and Neural Networks, 2003, pp.539–541.
12. Aida-zade K.R., Talibov S.G. Analysis of the effectiveness of the methods of recognition of authorship of texts in the Azerbaijani language // The 5th International Conference on Control and Optimization with Industrial Applications (COIA-2015), 27−29 August, 2015, Baku, Azerbaijan, pp.183.
13. Gasimov S., Ibrahimov I. Analysis of sentences and words used in azerbaijani texts // The Second International Conference Problems of Cybernetics and Informatics, September 10–12, 2008, Baku, pp. 117–119.
14. Doğan S., Diri B. Türkçe Dokümanlar için N-gram Tabanlı Yeni Bir Sınıflandırma // Yazar, Tür ve Cinsiyet. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2010, 3, s.11–20.
15. Biricik G., Diri B., Sönmez A. A New Method For Attribute Extraction with Application on Text Classification / 5th International Conference on Soft Computing, Computing with Words, ICSCCW, North Cyprus, Famagusta, 2009, pp.1–4.

16. George H. Estimating Continuous Distributions in Bayesian Classifiers / 11th Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp.338–345.
17. Yasdi M., Diri B. Soyut Özellik Çıkarımı İle Yazar Tanıma / IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, SIU 2012, Fethiye (18−20 Nisan), 2012, s.4.
18. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов, М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012, 326 с.
19. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А.Маркова // Вестник МГУ, сер.9: Филология, 2000, №2, с.115–126.
20. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады ТУСУРа, № 1 (19), часть 2, июнь 2009, с.36−42.
21. Vapnik V.N. Statistical Learning Theory, New York: Wiley, 1998, 732 p.
22. Vapnik V.N. The nature of statistical learning theory, New York: Springer-Verlag, 2000, 332 p.
23. C.-W. Hsu, C.-C. Chan, C.-J. Lin. A practical guide to support vector classification. // www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
24. Романов А.С., Мещеряков Р.В. Идентификация автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив. www.dialog-21.ru/digests/dialog2009/materials/pdf/67.pdf

UOT 517.977.56 517.977.58
Ayda-zadə Kamil R.[1], Talıbov Səxavət Q.[2]
AMEA İdarəetmə Sistemləri İnstitutu, Bakı, Azərbaycan
[1]Kamil_ayda-zade@rambler.ru, [2]saxavat@yahoo.com
Azərbaycan dilində yazılmış mətnlərin müəllifinin müəyyən olunması üsullarının analizi
Məqalədə mətn müəllifinin tanınması üsulları və alqoritmləri təhlil edilir. Əlamətlərin seçilməsi *n*=1 və *n*=2 qiymətlərinə müvafiq *n*-qramlara əsaslanır. Azərbaycan dilində yazılmış mətnlərin müəllifinin tanınması üçün kompüter eksperimentlərinin nəticələri verilir.
Açar sözlər: *identifikasiya, müəllifin identifikasiyası, tanıma, n-qram, dayaq vektor maşını.*

Kamil R. Ayda-zade[1], Sakhavat G. Talibov[2]
Institute of Control Systems of ANAS, Baku, Azerbaijan
[1]Kamil_ayda-zade@rambler.ru, [2]saxavat@yahoo.com
Analysis of methods for the identification authorship of the text in Azerbaijani language
In this paper, the methods, algorithms used for recognition texts authorship is analyzed. The feature selection based on n-grams with n = 1, and n = 2. The results of computer experiments to recognize the authorship of texts in Azerbaijani are presented.
Keywords*: identification, author identification, recognition, n-qram, support vector machine.*