УДК 519.85

DOI 10.21685/2307-4205-2017-1-10

АЛГОРИТМ ИДЕНТИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

А. В. Полтавский, Н. К. Юрков, А. В. Гриншкун

Введение

В нашем современном «компьютеризированном» мире наблюдается переизбыток неструктурированных потоков текстовой информации, поэтому так важен и актуален вопрос о возможностях ее объективного анализа. Само же понятие «информация» имеет достаточно широкий спектр определений, поэтому изначально следует обратиться к концептуальным истокам и, совместив философско-культурологический и технический подходы, выяснить, что информацией можно считать лишь тот феномен [1–3], который содержит следующие компоненты: автор, адресат, канал связи (передачи приема информации) и некий код [2, 4–6], к примеру, вербальный или математический язык в устной или письменной форме (или сам текст). А при такой постановке становится очевидным, что для определения авторства любой текстовой информации необходимо выработать определенный подход и разработать инновационный механизм для ее анализа. При разработке такого возможного инструментария следует применить современные научно-методические подходы и возможности вычислительных систем.

В плане инновационной проблематики такого рода разработку можно отнести к расширению арсенала аналитических возможностей программных средств. В частности, данный подход можно применять как дополнение к активно используемой в России программе типа Антиплагиат[1], поскольку предлагаемый алгоритм дает возможность повысить уровень достоверности определения авторства текста по вероятностным критериям [7, 8].

Изначально следует задать вопрос: каковы инновационные данные и технические возможности для анализа текстовой информации? В результате поисков ответа на этот вопрос *впервые* было создано вычислительное устройство [3], которое может быть использовано непосредственно при расчетах, связанных с идентификацией текстовой информации и в случаях необходимости определения ее авторства. Такая научно-техническая задача уже решалась и на нее получен патент [3]. Новизна подхода заключается в том, что в него положены формальные логические и числовые методы как основы для принципа работы вычислительного устройства. Данное обстоятельство в своем итоге повышает возможность принимать более объективные решения при определении и в случае необходимости при защите авторских прав непосредственно создателей текстовой информации. Покажем сущность инновационного подхода.

Алгоритм анализа текстовой информации

Числовой и логико-содержательный анализ текстовой информации осуществляется следующим образом. Два отрывка текстовой информации сравниваются путем сопоставления информации о вероятностях появления какой-либо буквы (или слов) в двух различных отрывках текста. Покажем на буквах. Среднее значение разности $\Delta P_{\text{ср}}$ между вероятностью появления i-й буквы для отрывка «а» $P_{\text{а}i}$ и вероятностью появления j-й буквы из отрывка «б» $P_{\text{б}j}$ оценивается по следующей формуле:

$$\Delta P_{\text{ср}} = \frac{1}{n}\sum_{k=1}^{n}\left|P_{\text{а}i} - P_{\text{б}j}\right|_k,\ k = 1, \ldots, n, \tag{1}$$

$$i = 1, \ldots, m_{\text{а}i}, j = 1, \ldots, m_{\text{б}j},$$

где n – количество букв в алфавите (в данном случае – русском); $m_{\text{а}i}$ – количество i-й буквы в отрывке текста «а»; $m_{\text{б}j}$ – количество j-й буквы в отрывке текста «б».

[1] Российский интернет-проект, представленный как программно-аппаратный комплекс для проверки текстовых документов на наличие заимствований из открытых источников в сети Интернет и других источников.

Сравнивая величину $\Delta P_{\text{ср}}$ с допустимым значением $\Delta P_{\text{д}}$ (критерием принятия решений), можно сделать вывод о принадлежности двух отрывков «а» и «б» текстовой информации одному автору. Если $\Delta P_{\text{ср}} \le \Delta P_{\text{д}}$, то отрывки текстов «а» и «б» принадлежат одному автору. В противном случае ($\Delta P_{\text{ср}} > \Delta P_{\text{д}}$) авторы этих отрывков могут быть различными.

Вероятности $P_{\text{а}i}$ и $P_{\text{б}j}$ определяются по следующим формулам:

$$P_{\text{а}i} = \frac{m_{\text{а}i}}{N_{\text{а}}},\ i = 1, ...,\ m_{\text{а}}; \tag{2}$$

$$P_{\text{б}j} = \frac{m_{\text{б}j}}{N_{\text{б}}},\ j = 1, \ldots,\ m_{\text{б}}, \tag{3}$$

где $N_{\text{а}}$ – общее количество букв в отрывке текста «а»; $N_{\text{б}}$ – общее количество букв в отрывке текста «б».

Апробация действия алгоритма

Для подтверждения работы инновационного алгоритма и иллюстрации разработанного подхода целесообразно рассмотреть определенный пример. В качестве объектов исследования и для содержательного анализа текстов будем использовать стихотворения Иосифа Бродского «Одиссей Телемаку» (отрывок «а») и «На смерть Жукова» (отрывок «б»).

Одиссей Телемаку

Мой Телемак,
Троянская война
окончена. Кто победил – не помню.
Должно быть, греки: столько мертвецов
вне дома бросить могут только греки….
И все-таки ведущая домой
дорога оказалась слишком длинной,
как будто Посейдон, пока мы там
теряли время, растянул пространство.
Мне неизвестно, где я нахожусь,
что передо мной. Какой-то грязный остров,
кусты, постройки, хрюканье свиней,
заросший сад, какая-то царица,
трава да камни…Милый Телемак,
все острова похожи друг на друга,
когда так долго странствуешь, и мозг
уже сбивается, считая волны,
глаз, засоренный горизонтом, плачет,
и водяное мясо застит слух.
Не помню я, чем кончилась война,
и сколько лет тебе сейчас, не помню.

***

Расти большой, мой Телемак, расти.
Лишь боги знают, свидимся ли снова.
Ты и сейчас уже не тот младенец,
перед которым я сдержал быков.
Когда б не Паламед, мы жили вместе.
Но, может быть, и прав он: без меня
ты от страстей Эдиповых избавлен,
и сны твои, мой Телемак, безгрешны.

На смерть Жукова

Вижу колонны замерших внуков,
гроб на лафете, лошади круп.
Ветер сюда не доносит мне звуков
русских военных плачущих труб.
Вижу в регалии убранный труп:
в смерть уезжает пламенный Жуков.
Воин, пред коим многие пали
стены, хоть меч был вражьих тупей,
блеском маневра о Ганнибале
напоминавший средь волжских степей.
Кончивший дни свои глухо, в опале,
как Велизарий или Помпей.
Сколько он пролил крови солдатской
в землю чужую! Что ж, горевал?
Вспомнил ли их, умирающий в штатской
белой кровати? Полный провал.
Что он ответит, встретившись в адской
области с ними? «Я воевал».
К правому делу Жуков десницы
больше уже не приложит в бою.
Спи! У истории русской страницы
хватит для тех, кто в пехотном строю
смело входили в чужие столицы,
но возвращались в страхе в свою.

Анализ полученных численных расчетов этих текстов показывает, что общее количество букв в отрывке для текста «а» $N_{\text{а}} = 745$, а в отрывке «б» $N_{\text{б}} = 611$. Количество букв в русском алфавите мы знаем – $n = 33$. Количество i-й буквы $m_{\text{а}i}$ в отрывке «а» и j-й буквы $m_{\text{б}j}$ в отрывке «*б*» приведены в соответствующей сводной табл. 1.

Вероятности появления i-й буквы отрезка «а» $P_{\text{а}i}$ и i-й буквы отрезка «б» $P_{\text{б}j}$, определяемые непосредственно по рабочим формулам, соответственно, (2) и (3), представлены также в сводной табл. 1.

Таблица 1

Вероятности появления различных букв в двух стихотворениях

Показатели	Буквы русского алфавита										
	А	Б	В	Г	Д	Е	Е	Ж	З	И	Й
$m_{\text{а}i}$	60	14	25	16	25	60	1	8	13	40	18
$P_{\text{а}i}$	0,081	0,019	0,034	0,021	0,034	0,081	0,001	0,011	0,017	0,054	0,024
$m_{\text{б}j}$	36	10	45	6	12	41	1	12	6	54	15
$P_{\text{б}j}$	0,059	0,016	0,074	0,010	0,020	0,067	0,002	0,020	0,010	0,088	0,025
$\lvert P_{\text{а}i} - P_{\text{б}j} \rvert$	0,022	0,003	0,040	0,011	0,014	0,014	0,001	0,009	0,007	0,034	0,001
Показатели	Буквы русского алфавита										
	К	Л	М	Н	О	П	Р	С	Т	У	Ф
$m_{\text{а}i}$	33	30	34	46	83	15	33	45	54	12	0
$P_{\text{а}i}$	0,044	0,040	0,046	0,062	0,111	0,020	0,044	0,060	0,072	0,016	0
$m_{\text{б}j}$	23	35	18	34	61	20	30	33	34	24	1
$P_{\text{б}j}$	0,038	0,057	0,029	0,056	0,100	0,033	0,049	0,054	0,056	0,039	0,002
$\lvert P_{\text{а}i} - P_{\text{б}j} \rvert$	0,006	0,017	0,017	0,006	0,011	0,013	0,005	0,006	0,016	0,023	0,002

Окончание табл. 1

Показатели	Буквы русского алфавита										
	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
m_{ai}	5	4	8	6	1	0	16	13	1	5	17
P_{ai}	0,007	0,005	0,011	0,008	0,001	0	0,021	0,017	0,001	0,007	0,023
$m_{бj}$	14	3	7	6	3	0	10	8	0	7	2
$P_{бj}$	0,023	0,005	0,011	0,010	0,005	0	0,016	0,013	0,	0,011	0,003
$\lvert P_{ai} - P_{бj} \rvert$	0,016	0	0	0,002	0,004	0	0,005	0,004	0,001	0,004	0,020

В нижней строке сводной табл. 1 размещены численные значения величины модуля разности для измеряемой информации величин $\lvert P_{ai} - P_{бj} \rvert$.

Сумма значений этих величин равна 0,334, а среднее значение разности $\Delta P_{ср}$ между вероятностью появления *i*-й буквы из отрывка текста «а» P_{ai} и вероятностью появления *i*-й буквы для отрывка текста «б» $P_{бj}$ оценивается по следующей формуле (1):

$$\Delta P_{ср} = \frac{1}{33} \cdot 0{,}334 = 0{,}01.$$

Если принять (в качестве критерия для оценки) допустимое значение этой вероятности как $\Delta P_{д} = 0{,}02$, то можно предполагать о том, что отрывки текста «а» и текста «б» принадлежат одному автору данных произведений [7, 8].

Предложение вычислительного устройства анализа текста

Технический результат непосредственно достигается тем, что само вычислительное устройство для измерения и содержательного анализа текстовой информации содержит первую и вторую группы входных регистров, состоящих из *n* элементов. Также устройство содержит с первого по четвертый входные регистры, первую и вторую группы блоков деления, состоящие из *n* элементов, группу блоков вычитания по модулю, состоящую из *n* элементов, накопительный сумматор, блок деления, блок сравнения, блок индикации, генератор тактовых импульсов и распределитель импульсов (РИ). Его тактовый вход непосредственно соединен с выходом генератора тактовых импульсов, а первый выход РИ – с входами записи первой и второй групп входных регистров. Он также соединен с входами записи первого, второго, третьего и четвертого входных регистров. Его второй выход соединен с входами считывания первой и второй групп входных регистров, а также первого и второго входных регистров. Его третий и четвертый выходы соединены с входами считывания соответственно третьего и четвертого входных регистров. Информационные входы с первого по *n*-й элемент первой группы входных регистров являются входом задания исходной информации, на которые поступают значения m_{ai}, характеризующие количество *i*-й буквы в отрывке текста «а». Информационные входы с первого по *n*-й элемент второй группы входных регистров являются входом задания исходной информации, на которые поступают значения $m_{бj}$, характеризующие количество *j*-й буквы в отрывке «б». Информационные входы с первого по четвертый входной регистр являются входами для задания исходной информации. На них поступают, соответственно, значение $N_{а}$, характеризующее общее количество букв отрывка текста «*а*», значение $N_{б}$, характеризующее общее количество букв в отрывке текста «б», и значение *n*, характеризующее количество букв в алфавите. Значение $\Delta P_{д}$ характеризует величину допустимого значения средней разности между вероятностью появления *i*-й буквы отрывка «а» и вероятностью появления *j*-й буквы в отрывке «б». Выходы с первого по *n*-й элемент первой и второй групп входных регистров соединены с входами делимого каждого соответствующего элемента, соответственно, первой и второй групп блоков деления. Их входы делителя подключены непосредственно к выходам, соответственно, первого и второго входных регистров, а выходы непосредственно подключены к выходам уменьшаемого и к входам вычитаемого группы блоков-вычитания по модулю. Их выходы соединены с входами с первого по *n*-й накопитель-

ного сумматора, а его выход уже подключен к входу делимого блока деления. Вход делителя непосредственно соединен с выходом третьего входного регистра вычислителя, а выход – с информационным входом блока сравнения. Его пороговый вход подключен к выходу четвертого входного регистра, а выход – к входу блока индикации.

Работа вычислительного устройства анализа текстов

Устройство вычислительной системы (ВС) для содержательного анализа текстовой информации работает следующим образом (рис. 1). На информационные входы ВС с первого по *n*-й элементов первой группы 1 входных регистров засылаются соответственно величины $m_{а1}$, …, $m_{аi}$, …, $m_{аn}$, а на информационные входы с первого по *n*-й элементов второй группы 2 входных регистров подаются соответственно значения как $m_{б1}$, …, $m_{бi}$, …, $m_{бn}$. На информационные входы первого 3, второго 4, третьего 5 и четвертого 6 входных регистров направляются, соответственно, величины $N_а$, $N_б$, n и $\Delta P_д$. При этом *управляющий сигнал* на входы записи всех элементов этих групп входных регистров и входных регистров подается с первого выхода РИ15, темп работы вычислительного устройства непосредственно задается генератором 14 тактовых импульсов (ГТИ).

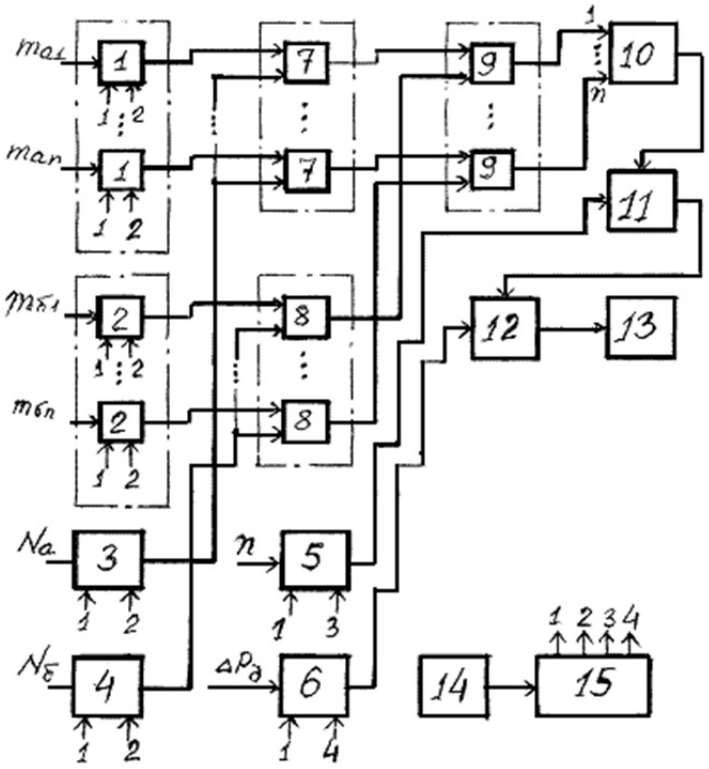

Рис. 1. Схематизация вычислительного устройства анализа текстов

По сигналу со второго выхода вычислительного устройства РИ15 на входы считывания первой 1 и второй 2 групп входных регистров величины $m_{аi}$ и $m_{бj}$ с их выходов засылаются на входы делимого соответственно первой 7 и второй 8 групп блоков деления, а на входы делителя этих групп направляются по сигналу со второго выхода РИ 15 с выходов соответственно первого 3 и второго 4 входных регистров значения $N_а$ и $N_б$. С выходов первой 7 и второй 8 групп блоков деления величины $P_{аi}$ и $P_{бj}$, определяемые по формулам (2) и (3), поступают, соответственно, на входы уменьшаемого и входы вычитаемого группы 9 блоков вычитания по модулю. С выходов этой группы величины $\left|P_{аi} - P_{бj}\right|_k$ засылаются на входы модели накопительного сумматора 10, с выхода которого значение $\frac{1}{n}\sum_{k\equiv 1}^{n} \left|P_{аi} - P_{бj}\right|_k$ подается на вход делимого блока 11 для деления.

На вход делителя этого блока по сигналу с третьего выхода РИ 15 направляется с выхода третьего входного регистра 5 величина *n*. С выхода блока 11 деления значение для $\Delta P_{ср}$, определяемое по формуле (1), поступает на информационный вход блока 12 сравнения, на пороговый вход которого по сигналу с четвертого выхода РИ 15 засылается с выхода четвертого входного регистра 6 величина $\Delta P_{д}$. Если выполняется условие $\Delta P_{ср} \leq \Delta P_{д}$ (отрывки текстов «а» и «б» принадлежат одному автору) на выходе блока 12 сравнения появится *управляющий сигнал*, который приведет к загоранию лампы-индикатора блока 13 индукции. В противном случае, когда $\Delta P_{ср} > \Delta P_{д}$ сигнала на выходе блока 12 сравнения не будет и индикатор блока 13 индукции не засветится. Это и будет свидетельствовать о том, что отрывки «а» и «б» принадлежат разным авторам.

Заключение

Таким образом, предлагаемая нами информационная технология и ее инновационные составляющие в целях ее практической реализации с использованием разработанной вычислительной системы заключаются в следующем:

1) предложены алгоритм и инновационный подход к анализу текстовой информации с описанием принципа работы вычислительного устройства для реализации программы созданного алгоритма;

2) разработан оригинальный механизм анализа текстовой информации, основанный на комбинировании формально-числовых и логико-содержательных методов исследования;

3) технический результат работы алгоритма и устройства достигнут не только за счет математического аппарата, но и за счет предлагаемых цифровых технических средств (блоков и электронных элементов);

4) расширены ресурсы современных информационных технологий и технических средств вычислительных систем, с помощью которых можно осуществлять непосредственно количественную оценку и определение авторства текстовой информации;

5) повышен уровень достоверности определения авторства текстовой информации в результате применения разработанного вычислительного устройства на основе предложенного алгоритма.

В заключение отметим, что применимость предлагаемого алгоритма и данной инновационной разработки ВС обосновываются прежде всего тем, что они могут быть использованы в разных областях (отраслях) управления знаниями, например, при расчетах, связанных с идентификацией текстовой информации в случаях необходимости определения ее автора (или соавторства) с целью принятия объективных решений при защите авторских прав создателей текста и других подобных объектов, связанных непосредственно с правом интеллектуальной собственности. Алгоритм может применяться в лингвистической области при построении графовых моделей, а также в современных системах с искусственным интеллектом (ИИ) для принятия решений, в поисковых справочных системах электронных библиотек, а также в современных моделях и объектах робототехники и т. п.

Кроме того, предложенный алгоритм может быть полезен для наглядной демонстрации возможности объективного математического анализа и идентификации различных текстов при подготовке лингвистов, филологов, математиков и учителей по информатике.

Библиографический список

1. Пелипенко, А. А. Постижение культуры. Ч. 1. Культура и смысл / А. А. Пелипенко. – М. : Роспен, 2012. – 608 с.
2. Шеннон, К. Работы по теории информации и кибернетике / К. Шеннон. – М. : Изд-во иностранной литературы, 1963. – 830 с.
3. Пат. № 2568272 РФ. Устройство для содержательного анализа текстовой информации / Полтавский А. В. [и др.]. Зарег. 16.10.2015.
4. Полтавский, А. В. Программные средства вычислительных систем. Ч. 1. ЭВМ первых поколений / А. В. Полтавский. – М. : МГПУ, 2015. – 92 с.
5. Полтавский, А. В. Компьютерный практикум : учеб. пособие / А. В. Полтавский, И. И. Кочегаров, Н. В. Горячев. – Пенза : Изд-во ПГУ, 2015. – 238 с.
6. Информационные технологии в предметной области / отв. ред. проф. В. А. Бубнов. – М. : МГПУ, 2004. – Вып. II. – 246 с. – (Мастер-класс МГПУ).

7. Кочегаров, И. И. Исследование влияния отверстий на собственные частоты пластинчатой конструкции / И. И. Кочегаров, С. И. Торгашин, А. В. Фомичев, А. В. Ляшенко // Труды Международного симпозиума Надежность и качество. – 2015. – Т. 2. – С. 297–298.
8. Исследование программных пакетов моделирования влияния электромагнитных воздействий на изделия радиоэлектронных средств / С. А. Бростилов, Т. Ю. Бростилова, Н. К. Юрков, Н. В. Горячев, В. А. Трусов, В. Я. Баннов, А. О. Бекбаулиев // Труды Международного симпозиума Надежность и качество. – 2015. – Т. 1. – С. 206–209.

Полтавский Александр Васильевич
доктор технических наук, профессор,
кафедра информатизации образования,
Институт математики, информатики
и естественных наук,
Московский государственный
педагогический университет;
ведущий научный сотрудник,
Институт проблем управления
им. В. А. Трапезникова Российской академии наук,
г. Москва
(117997, Россия, г. Москва ул. Профсоюзная, 65)
E-mail: avp57avp@yandex.ru

Poltavskiy Aleksandr Vasil'evich
doctor of technical sciences, professor,
sub-department of informatization of education,
Institute of Mathematics, Informatics
and Natural Sciences,
Moscow State Pedagogical University;
leading researcher,
Institute of management named after V. A. Trapeznikov
of Russian Academy of Sciences
(117997, 65 Profsoyuznaya street, Moscow, Russia)

Юрков Николай Кондратьевич
доктор технических наук, профессор,
заслуженный деятель науки РФ,
заведующий кафедрой конструирования
и производства радиоаппаратуры,
Пензенский государственный университет
(440026, Россия, г. Пенза, ул. Красная, 40)
E-mail: yurkov_NK@mail.ru

Yurkov Nikolay Kondrat'evich
doctor of technical sciences, professor,
honoured worker of science of the Russian Federation,
head of sub-department of radio equipment
design and production,
Penza State University
(440026, 40 Krasnaya street, Penza, Russia)

Гриншкун Александр Вадимович
ассистент,
Институт математики, информатики
и естественных наук,
Московский государственный
педагогический университет
(119991, Россия, г. Москва,
ул. Малая Пироговская, 1/1)
E-mail: grishkun@mail.ru

Grinshkun Aleksandr Vadimovich
assistant,
Institute of Mathematics, Informatics
and Natural Sciences,
Moscow State Pedagogical University
(119991, 1/1 Malaya Pirogovskaya street,
Moscow, Russia)

Аннотация. Предложена новая информационная технология и подход к анализу текстовой информации на основе применения формальных числовых методов и программируемых средств вычислительного устройства. Показано, что для определения авторства текстовой информации необходим инновационный механизм ее анализа. Предложен подход, расширяющий возможность программных средств повышения уровня достоверности определения авторства текста по вероятностным характеристикам. Описан алгоритм и принцип работы предлагаемого вычислительного устройства, которые могут быть использованы в расчетах, связанных с идентификацией текстовой информации, а также в случаях определения ее авторства. Представлен метод оценки качества принятия решений по содержательному анализу текстов. Дан подробный анализ работы предложенного устройства.

Abstract. A new information technology and approach to the analysis of textual information based on the use of formal methods of numerical and programmable resources of the computing device.
It is shown that in order to determine the authorship of textual information, an innovative mechanism for its analysis is needed. An approach is proposed that extends the possibility of software tools to increase the level of authenticity of the authorship of text on probabilistic characteristics. The algorithm and the operating principle of the proposed computing device are described, which can be used in calculations related to the identification of textual information, as well as in cases of determining its authorship. A method for evaluating the quality of decision-making on the meaningful analysis of texts is presented. A detailed analysis of the operation of the proposed device is given.

Ключевые слова: анализ текстовой информации, инновационный механизм, числовые методы, вычислительное устройство.

Key words: analysis of textual information, innovative mechanism, the numerical methods, the computing device.

УДК 519.85

Полтавский, А. В.

Алгоритм идентификации текстовой информации / А. В. Полтавский, Н. К. Юрков, А. В. Гринюкун // Надежность и качество сложных систем. – 2017. – № 1 (17). – С. 77–84. DOI 10.21685/2307-4205-2017-1-10.