ТЕХНОЛОГИЧЕСКИЕ ОСНОВЫ ПОВЫШЕНИЯ НАДЕЖНОСТИ И КАЧЕСТВА ИЗДЕЛИЙ

УДК 004.93

DOI 10.21685/2307-4205-2016-3-5

ОСНОВЫ РАСПОЗНАВАНИЯ ОБРАЗОВ В ТЕКСТОВОЙ ИНФОРМАЦИИ

А. В. Полтавский, Е. Ю. Русяева, А. А. Бурба

Введение

В современном мире наблюдается переизбыток неструктурированной текстовой информации, поэтому так актуален вопрос о возможностях ее адекватного анализа. Само понятие «информация» имеет настолько широкий спектр определений, что сделалось, по сути, лишь зонтичным брендом. Поэтому мы изначально обратились к концептуальным истокам и, совместив философско-культурологический и технический подходы, выяснили, что информацией можно считать лишь тот феномен [1], который содержит следующие компоненты: автор, адресат, канал передачи и некий код [2], к примеру, вербальный или математический язык в устной или письменной форме (сам текст). А при таком раскладе становится очевидным, что для определения авторства текстовой информации необходимо выработать определенный подход и разработать инновационный механизм для ее анализа. При разработке такого инструментария мы решили применить современные 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 Антиплагиат[1], поскольку предлагаемый алгоритм дает возможность повысить уровень достоверности определения авторства текста по вероятностным критериям.

Изначально, мы задались вопросом: каковы инновационные данные и технические возможности анализа текстовой информации? В результате поисков ответа на этот вопрос впервые было создано вычислительное устройство [3], которое может быть использовано при расчетах, связанных с идентификацией текстовой информации и в случаях необходимости определения ее авторства. Такая техническая задача еще не решалась, тем более, подобным образом. Данный факт объясняется еще и тем, что пока отсутствовали комбинации методов содержательного анализа различных частей текстовой информации. Новизна подхода заключается в том, что в него положены формальные логические и числовые методы как основы для принципа работы вычислительного устройства. Данное обстоятельство, в итоге, повышает возможность принимать более объективные решения при определении и в случае необходимости при защите авторских прав создателей текстовой информации.

1. Алгоритм анализа текстовой информации

Числовой и логико-содержательный анализ текстовой информации осуществляется следующим образом. Два отрывка текстовой информации сравниваются путем сопоставления инфор-

[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.

мации о вероятностях появления какой-либо буквы в двух различных отрывках текста. Среднее значение разности $\Delta P_{\text{ср}}$ между вероятностью появления *i*-й буквы для отрывка «*а*» $P_{\text{a}i}$ и вероятностью появления *j*-й буквы из отрывка «*б*» $P_{\text{б}j}$ оценивается по следующей формуле [4]:

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

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

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 $P_{\text{a}i}$ и $P_{\text{б}j}$ определяются по следующим формулам [5]:

$$P_{\text{a}i} = \frac{m_{\text{a}i}}{N_{\text{a}}},\ i = 1, \ldots,\ m_{\text{a}}, \tag{2}$$

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

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

2. Апробация работы алгоритма

Для подтверждения работы инновационного алгоритма и иллюстрации разработанного подхода целесообразно рассмотреть определенный пример. Так, в качестве объектов для содержательного анализа 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 $n = 33$. Количество i-й буквы $m_{\text{а}i}$ в отрывке «*а*» и j-й буквы $m_{\text{б}j}$ в отрывке «*б*» приведены в 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 по формулам, соответственно (2) и (3), представлены также в табл. 1.

Таблица 1

Вероятности появления различных букв в двух стихотворениях

Параметр	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

Окончание табл. 1

	Буквы русского алфавита										
	К	Л	М	Н	О	П	Р	С	Т	У	Ф
m_{ai}	33	30	34	46	83	15	33	45	54	12	0
P_{ai}	0,044	0,040	0,046	0,062	0,111	0,020	0,044	0,060	0,072	0,016	0
$m_{бj}$	23	35	18	34	61	20	30	33	34	24	1
$P_{бj}$	0,038	0,057	0,029	0,056	0,100	0,033	0,049	0,054	0,056	0,039	0,002
$\lvert P_{ai} - P_{бj} \rvert$	0,006	0,017	0,017	0,006	0,011	0,013	0,005	0,006	0,016	0,023	0,002
	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 величины модуля разности для 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 отрывка «*а*» P_{ai} и вероятностью появления *i*-й буквы для отрывка «*б*» $P_{бj}$ оценивается по формуле (1)

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

Если принять (в качестве критерия оценки) допустимое значение этой вероятности $\Delta P_{д} = 0{,}02$, то можно делать вывод о том, что отрывки «*а*» и «*б*» принадлежат одному автору соответственно в принятых ограничениях.

3. Описание технического результата инновационного решения

Технический результат непосредственно достигается тем, что вычислительное устройство для 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 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 элементов 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 входных регистров являются входами для задания исходной информации. На них поступает, 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, элементов первой и второй групп входных регистров соединены с входами делимого каждого соответствующего элемента, соответственно, первой и второй групп блоков деления. Их входы делителя подключены к выходам, соответственно, первого и второго входных регистров, а выходы подключены к выходам уменьшаемого и к входам вычитаемого группы блоков-вычитания по модулю. Их выходы соединены с входами с первого по *n*-й накопительного сумматора, а его выход подключен к входу делимого блока деления. Вход делителя соединен с выходом третьего входного регистра, а выход – с информационным входом блока сравнения. Его пороговый вход подключен к выходу четвертого входного регистра, а выход – к входу блока индикации.

4. Принцип работы вычислительного устройства

Устройство для содержательного анализа текстовой информации работает следующим образом (рис. 1). На информационные входы с первого по *n*-й элемент 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 устройства задается генератором 14 тактовых импульсов (ГТИ).

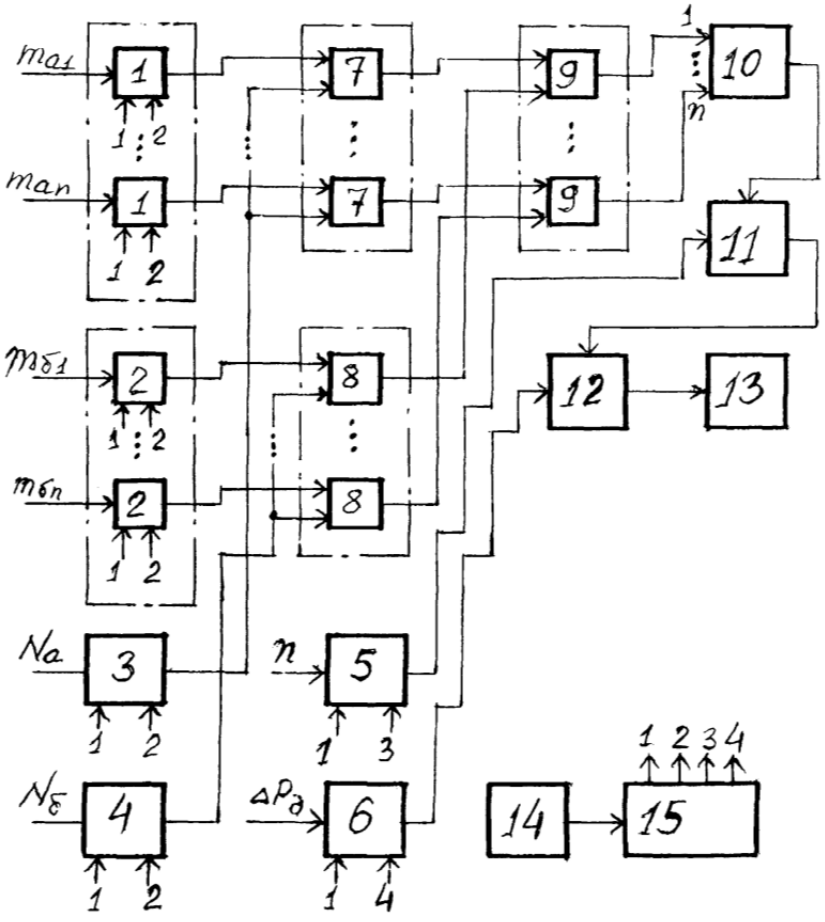

Рис. 1. Схематизация вычислительного устройства анализа текстов

По сигналу со второго выхода вычислительного устройства РИ15 на входы считывания первой *1* и второй *2* групп входных регистров величины m_{ai} 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_{ai} 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 $\left|P_{ai} - P_{бj}\right|_k$ засылаются на входы модели **накопительного сумматора** *10*, с выхода которого значение $\frac{1}{n}\sum_{k=1}^{n}\left|P_{ai} - P_{бj}\right|_k$ подается на вход делимого блока *11* 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 $\Delta P_{ср}$, 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 $\Delta P_{ср} \le \Delta P_д$ (отрывки текстов «*а*» и «*б*» принадлежат одному автору) на выходе блока *12* сравнения появится управляющий сигнал, который приведет к загоранию 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, разработанная новая информационная технология и инновационные составляющие предлагаемой разработки для практического использования вычислительного устройства заключаются в следующем:

1) впервые предложен инновационный подход к анализу текстовой информации с описанием принципа работы вычислительного устройства для реализации разработанного алгоритма;

2) разработан новый оригинальный инновационный механизм анализа текстовой информации, основанный на формально-числовых и логико-содержательных методах исследования;

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

4) расширен арсенал новых информационных технологий (НИТ) и технических средств, с помощью которых можно осуществлять оценку и определение авторства текстовой информации;

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

В итоге отметим также и то, что промышленная применимость данной инновационной разработки обосновывается и тем, что она может быть использована в разных областях (отраслях) управления знаниями, например, при расчетах, связанных с идентификацией текстовой информации в случаях необходимости определения ее автора с целью принятия объективных решений при защите авторских прав для создателей текстовой информации; в лингвистической области при построении графовых моделей, в системах с искусственным интеллектом (ИИ) при принятии решений и т.п.

Список литературы

1. Пелипенко, А. А. Постижение культуры : моногр. : в 2 ч. 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 Российская Федерация. Устройство для содержательного анализа текстовой информации / Полтавский А. В., Русяева Е. Ю., Бурба А. А. Зарег. 16.10.2015 г.

4. Садыков, С. С. Оценка возможности распознавания отдельных реальных плоских объектов на основе их безразмерных контурных признаков / С. С. Садыков, Я. Ю. Кульков // Надежность и качество сложных систем. – 2015. – № 4 (12). – С. 101–109.
5. Федотов, Н. Г. Вопросы построения алгоритмов сокращения признакового пространства на основе селекции информативных признаков / Н. Г. Федотов, А. А. Семов, А. В. Моисеев // Труды международного симпозиума Надежность и качество. – 2016. – № 1 (13). – С. 299–301.

Полтавский Александр Васильевич
доктор технических наук,
ведущий научный сотрудник,
Институт проблем управления
Российской академии наук им. В. А. Трапезникова
(117997, Россия, г. Москва, ул. Профсоюзная, 65)
E-mail: avp57avp@yandex.ru

Русяева Елена Юрьевна
кандидат философских наук,
старший научный сотрудник,
Институт проблем управления
Российской академии наук им. В. А. Трапезникова
(117997, Россия, г. Москва, ул. Профсоюзная, 65)
E-mail: 1779624@mail.ru

Бурба Александр Алексеевич
кандидат технических наук, научный сотрудник,
Институт проблем управления
Российской академии наук им. В. А. Трапезникова
(117997, Россия, г. Москва, ул. Профсоюзная, 65)
E-mail: lab-54@bk.ru

Аннотация. Предложена новая информационная технология и подход к анализу текстовой информации на основе применения формальных числовых методов. Описан алгоритм и принцип работы запатентованного авторами вычислительного устройства, которое может быть использовано при расчетах, связанных с идентификацией текстовой информации в случаях определения ее авторства.

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

Poltavskiy Aleksandr Vasil'evich
doctor of technical sciences, leading researcher,
Institute of management problems
of Russian Academy of Sciences
named after V. A. Trapeznikov
(117997, 65 Profsoyuznaya street, Moscow, Russia)

Rusyaeva Elena Yur'evna
candidate of philosophical sciences,
senior stuff scientist,
Institute of management problems
of Russian Academy of Sciences
named after V. A. Trapeznikov
(117997, 65 Profsoyuznaya street, Moscow, Russia)

Burba Aleksandr Alekseevich
candidate of technical science, stuff scientist,
Institute of management problems
of Russian Academy of Sciences
named after V. A. Trapeznikov
(117997, 65 Profsoyuznaya street, Moscow, Russia)

Abstract. Applying an innovative approach to the analysis of textual information based on the use of formal numerical methods. The mechanism and principle of operation of the patented authors computing device that can be used in the calculations relating to the identification of textual information in the case of certain of its authorship.

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

УДК 004.93
Полтавский, А. В.
Основы распознавания образов в текстовой информации / А. В. Полтавский, Е. Ю. Русяева, А. А. Бурба // Надежность и качество сложных систем. – 2016. – № 3 (15). – С. 28–34. DOI 10.21685/2307-4205-2016-3-5.