С. Д. Кулик, К. И. Ткаченко, Д. А. Никонец

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ВЫЯВЛЕНИЯ ИСКАЖЕНИЙ ИНФОРМАЦИИ В ДОКУМЕНТАХ

Введение

В настоящее время под документом принято понимать не только обычный бумажный документ, но и электронный документ. Документ, содержащий текст на некотором языке, например итальянском, может быть представлен не только на бумажном носителе, но и в электронном виде, к примеру, в виде сигнала при передаче информации по каналам связи, которая может быть искажена злоумышленником. Некоторые *лекарственные средства* (ЛР) могут содержать в упаковке к ним специальные электронные устройства, в которых хранится значимая для идентификации их подлинности информация, которая также может быть искажена. Вексель или ценная бумага (ЦБ) имеет не только материальное представление в виде бумажного документа, но и может быть представлен в электронном виде, информация о котором также может быть искажена. Далее основное внимание будет уделено не самому материальному объекту (ЛР, ЦБ, векселю), а именно информации, содержащейся в этих объектах или передаваемой по каналам связи. Это связано с тем, что искажение именно этой информации, например, в инструкции к лекарственному средству позволяет идентифицировать эти объекты (документы) именно как поддельные объекты.

Анализ статистических данных о регистрации преступлений в России, согласно ГИЦ МВД России [1], показывает следующее. В 2002 г. было зарегистрировано 69348 преступлений в сфере мошенничества (ст. 159 УК РФ), из которых было раскрыто только 46594. Таким образом, осталось не раскрыто почти 20 тысяч преступлений.

Практика показывает, что мошенничество порой связано с различными поддельными объектами, такими, например, как векселя, лекарственные средства [2], коммерческие (финансовые) документы и тому подобные различные объекты (документы).

Для решения проблемы, например, связанной с поддельными документами, необходимо автоматизировать саму работу эксперта по выявлению искаженной информации для последующего распознавания фальшивых объектов. АСОИБ имеет 2 режима работы. Основные (типовые) случаи искажения информации выявляются АСОИБ в автоматическом режиме (т. е. автоматически) с помощью специально реализованных алгоритмов. Сложные (нетиповые) случаи искажения информации выявляются АСОИБ в автоматизированном режиме с помощью специально обученного человека-оператора.

Одним из важных этапов при разработке АСОИБ является обучение (переобучение) человека-оператора: курсанта, эксперта (или стажера) — выявлять в сложных случаях искаженную информацию, характеризующую поддельные объекты. Для эффективного обучения эксперта необходимо иметь в большом количестве различные образцы поддельных объектов («фальшивок»). На практике реальные коллекции поддельных объектов иногда не в полной мере удовлетворяют этому. Для решения этой проблемы предлагается разработать *генератор* искаженной информации, характеризующей *образцы поддельных строго формализованных объектов* (документов).

Другим важным этапом является текущая работа эксперта, связанная с самим процессом выявления искаженной информации для последующего распознавания поддельных объектов. Для повышения эффективности работы эксперта предлагается разработать [3] *автоматизированное рабочее место* (АРМ) как одной из подсистем АСОИБ. В рамках этой подсистемы необходимо разработать алгоритм и его программную реализацию для определения языка представления информации (в том числе и искаженной).

1. Разработка специального генератора искаженной информации

Документы (в том числе и электронные), содержащие печатные и рукописные фрагменты [4], все чаще становятся объектами фальсификации. В список подделок часто попадают не только платежные документы, ценные бумаги (например, пользующиеся стабильным спросом векселя), но и инструкций на лекарственные средства, которые можно представить как *строго формализованные объекты* (документы). Эти объекты характеризуются тем, что имеют в своем составе *строго формализованные тексты*, содержащие информацию, которая может быть искажена.

На начальном этапе был выполнен краткий анализ возможных искажений информации строго формализованных документов (объектов). Эти искажения представлены в таблице 1. На их основе были выявлены возможные изменения строго формализованных текстовых документов. Основные возможные искажения строго формализованных текстовых документов представлены в таблице 2.

Таблица 1. Искажения информации в документах

№ п/п	Возможные группы искажений	Возможные искажения информации
1	дефект содержания	изменено содержание текста документа (по сравнению с эталонным текстом)
		изменен синтаксис языка текста документа
		изменена орфография текста документа
		изменен язык текста документа
		изменен текст документа
		добавлен текст документа
		изменен стиль текста документа
		добавлены дополнительные элементы содержания
		другие дефекты содержания
2	дефект формы представления информации	изменено общее форматирование текста
		изменен размер шрифта
		изменен тип шрифта
		изменены графические элементы
		удалены графические элементы
		изменен формат заголовков
		изменен формат абзацев
		изменен формат листа (бумаги)
		изменены поля листа (бумаги)
		изменено число абзацев
		изменены элементы текста, выделенные курсивом
		изменены элементы текста, выделенные полужирным
		изменены элементы текста, выделенные подчеркиванием
		изменен цвет (рисунок) фона документа
		отсутствуют водяные знаки (рисунок) документа
		изменены водяные знаки (рисунок) документа
		добавлены дополнительные элементы формы
		другие дефекты формы

Таблица 2. Искажения в формализованных текстовых документах

№ п/п	Возможные искажения	Примечание
1	выделение полужирным	реализация этих искажений возможна на практике и не требует специальных дополнительных алгоритмов
2	выделение курсивом	
3	выделение подчеркиванием	
4	пропуск согласной буквы	
5	пропуск гласной буквы	
6	удаление точки	
7	добавление точки	
8	удаление запятой	
9	добавление запятой	
10	удвоение буквы	
11	удаление пробела	
12	добавление пробела	
13	удаление фрагмента текста	
14	добавление фрагмента текста	
15	замена фрагмента текста	необходим алгоритм определения (поиска) заданного фрагмента текста
16	внесение орфографических ошибок	требуется специальный алгоритм орфографического анализа текста
17	удаление орфографических ошибок	
18	внесение синтаксических ошибок	требуется специальный алгоритм синтаксического анализа текста
19	удаление синтаксических ошибок	
20	изменение общего форматирования текста	необходимо наличие алгоритма анализа форматирования текста
21	изменение размера шрифта	требуются средства анализа размера шрифта
22	изменение типа шрифта	требуются средства анализа типа шрифта

Затем были разработаны специальные шаблоны (рекомендации) для генерации искажений информации, связанной с поддельными документами, которые были успешно реализованы в алгоритме **ALG-T**.

В этом алгоритме для генератора реализована идея генерирования искажений информации в копии подлинного документа. Для внесения искажений (опираясь на опыт [5] разработки генераторов [5, 6] и др.) был разработан генератор **GCT** искажений текста [7]. Сама идея генерации фактографических данных заимствована из работы [5]. На вход генератора поступает задание на выполнение искажений информации, например [2, 7], сведений инструкции по применению лекарственного средства. Далее при помощи специального диалогового окна пользователь сообщает программе, какие он хотел бы видеть искажения информации. Затем эта программа в специальном формате выдает сведения о том, какие искажения были внесены в информацию.

Полученные сведения могут быть использованы для обучения персонала АСОИБ. Задачей персонала (человека-оператора) является выявление искажений информации в том случае, когда автоматические средства АСОИБ не могут этого сделать.

Следующим этапом (после внедрения) является этап разработки набора различных генераторов искажения информации для целой группы специальных документов.

Обобщенный алгоритм работы этих генераторов представлен на рис. 1.

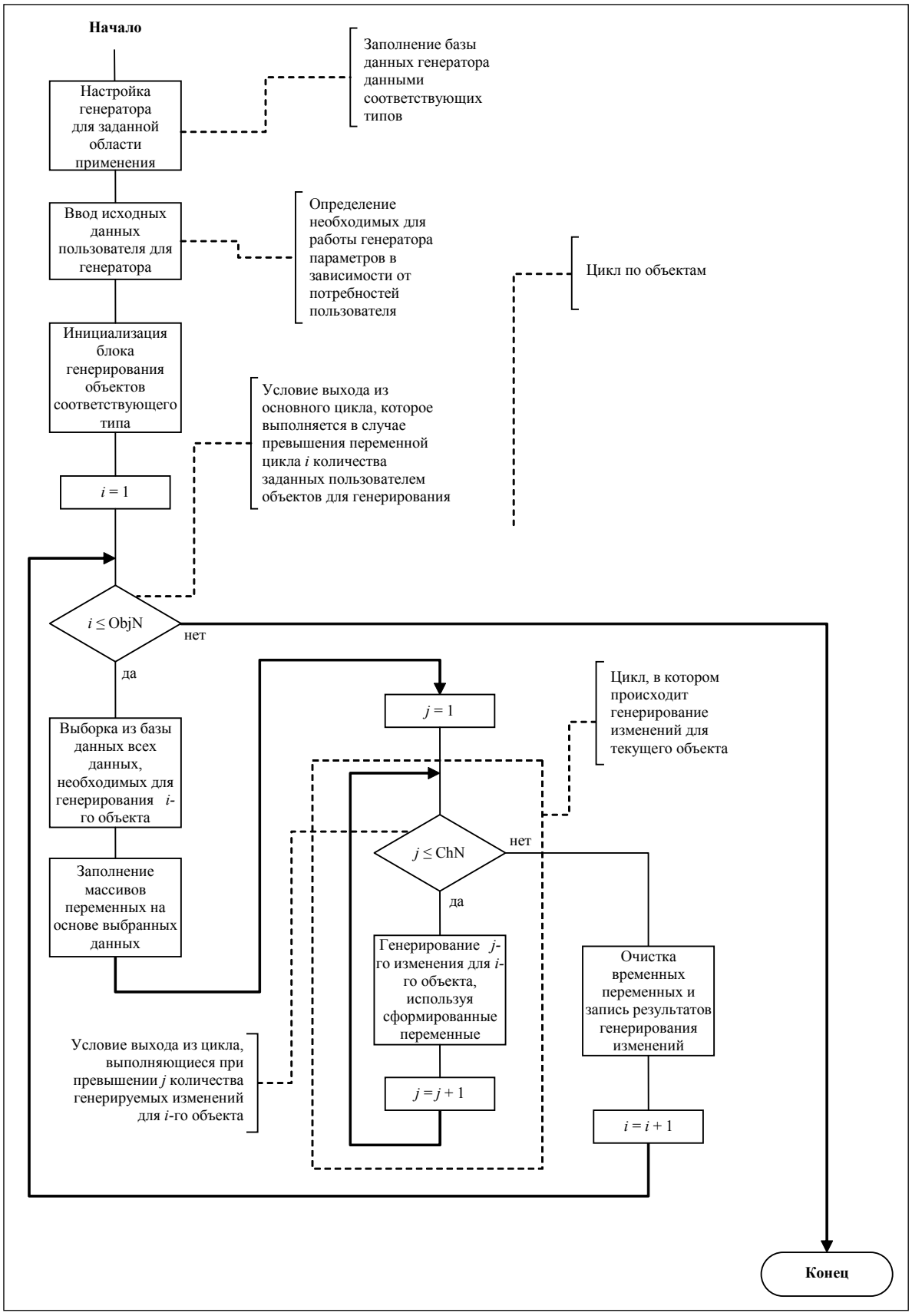

Рис. 1. Схема обобщенного алгоритма работы генератора АСОИБ

Их внедрение позволит автоматизировать процесс обучения (переобучения) персонала АСОИБ. Предложен еще один генератор, работоспособность которого пока проверяется на практике на реальных задачах. Разработана также подсистема **SGT** [6], решающая аналогичные задачи.

Полученные результаты позволили разработать специальное устройство [11], содержащее блок генерирования выборки. Это устройство позволяет выявлять искажения информации и тем самым идентифицировать фальшивые документы на русском языке. Разработанные генераторы достаточно универсальны и могут применяться не только в АСОИБ, но и в других областях, например в криминалистике, медицине и т. д.

2. Разработка алгоритма определения языка представления информации

Известно, что каждый язык имеет ряд отличительных характерных признаков, по которым этот язык возможно достаточно точно идентифицировать (распознать). Такими признаками могут быть символы алфавита языка, их уникальность, частота использования, а также факт принадлежности алфавита к определенному семейству языков, наличие специальных характерных слов (например, служебных слов), характерные (уникальные) сочетания символов и т. п. Практика показывает, что в некоторых языках алфавиты могут быть очень схожими (по изображению) или вообще одинаковыми, в самих алфавитах может не быть уникальных символов. В этих и подобных им случаях выделенные из текста признаки (символы) при экспертном исследовании текста могут неоднозначно идентифицировать язык представления информации. В текущей версии АСОИБ для определения национального языка текста представления информации реализовано пока только 10 языков, первые пять из которых используют алфавиты, относящиеся к **кириллице**: *русский*, *белорусский*, *сербский*, *болгарский*, *украинский*, а другие пять — к **латинице**: *испанский*, *немецкий*, *итальянский*, *французский*, *английский*. Рассмотрим подробнее некоторые очень важные признаки каждого из исследуемых 10 языков, используемых для представления информации. Опираясь на работу [8], были изучены перечисленные выше 10 языков (таблицы 3–12).

При передаче на письме текстов на русском языке (таблица 3) используется кириллица и состоит алфавит этого языка из 33 символов. Алфавит украинского языка (таблица 4) тоже относится к кириллице и состоит из 33 символов. Характерными символами этого языка являются: Ґґ, Єє, Її. Алфавит белорусского языка (таблица 5) относится к кириллице и состоит из 32 символов.

Таблица 3. Алфавит русского языка

А а	Б б	В в	Г г	Д д	Е е	Ё ё
Ж ж	З з	И и	Й й	К к	Л л	М м
Н н	О о	П п	Р р	С с	Т т	У у
Ф ф	Х х	Ц ц	Ч ч	Ш ш	Щ щ	Ъ ъ
Ы ы	Ь ь	Э э	Ю ю	Я я		

Таблица 4. Алфавит украинского языка

А а	Б б	В в	Г г	Ґ ґ	Д д	Е е
Є є	Ж ж	З з	И и	І і	Ї ї	Й й
К к	Л л	М м	Н н	О о	П п	Р р
С с	Т т	У у	Ф ф	Х х	Ц ц	Ч ч
Ш ш	Щ щ	Ь ь	Ю ю	Я я		

Таблица 5. Алфавит белорусского языка

А а	Б б	В в	Г г	Д д	Е е	Ё ё
Ж ж	З з	І і	Й й	К к	Л л	М м
Н н	О о	П п	Р р	С с	Т т	У у
Ў ў	Ф ф	Х х	Ц ц	Ч ч	Ш ш	Ы ы
Ь ь	Э э	Ю ю	Я я			

Таблица 6. Алфавит сербского языка

А а	Б б	В в	Г г	Д д	Ђ ђ
Е е	Ж ж	З з	И и	Ј ј	К к
Л л	Љ љ	М м	Н н	Њ њ	О о
П п	Р р	С с	Т т	Ћ ћ	У у
Ф ф	Х х	Ц ц	Ч ч	Џ џ	Ш ш

Таблица 7. Алфавит болгарского языка

А а	Б б	В в	Г г	Д д	Е е
Ж ж	З з	И и	Й й	К к	Л л
М м	Н н	О о	П п	Р р	С с
Т т	У у	Ф ф	Х х	Ц ц	Ч ч
Ш ш	Щ щ	Ъ ъ	Ь ь	Ю ю	Я я

Таблица 8. Алфавит английского языка

A a	B b	C c	D d	E e	F f	G g
H h	I i	J j	K k	L l	M m	N n
O o	P p	Q q	R r	S s	T t	U u
V v	W w	X x	Y y	Z z		

Таблица 9. Алфавит французского языка

A a	B b	C c	D d	E e	F f	G g
H h	I i	J j	K k	L l	M m	N n
O o	P p	Q q	R r	S s	T t	U u
V v	W w	X x	Y y	Z z	À à	Â â
Æ æ	Ç ç	È è	É é	Ê ê	Ë ë	Î î
Ï ï	Ô ô	Œ œ	Ù ù	Û û	Ü ü	Ÿ ÿ

Таблица 10. Алфавит немецкого языка

A a	Ä ä	B b	C c	D d	E e
F f	G g	H h	I i	J j	K k
L l	M m	N n	O o	Ö ö	P p
Q q	R r	S s	ß	T t	U u
Ü ü	V v	W w	X x	Y y	Z z

Таблица 11. Алфавит испанского языка

A a	B b	C c	D d	E e	F f	G g
H h	I i	J j	K k	L l	M m	N n
Ñ ñ	O o	P p	Q q	R r	S s	T t
U u	V v	W w	X x	Y y	Z z	

Таблица 12. Алфавит итальянского языка

A a	B b	C c	D d	E e	F f	G g
H h	I i	L l	M m	N n	O o	P p
Q q	R r	S s	T t	U u	V v	Z z
J j	K k	X x	W w	Y y		

Характерным символом этого языка является символ Ўў («У краткое» или «У неслоговое»). Алфавит сербского языка (таблица 6) относится к кириллице и состоит из 30 символов. Характерными символами этого языка являются: Ђђ (означает аффрикату /dђ/, русская транскрипция [дьже]), Jj (русская транскрипция [йе]), Љљ (лигатура «ль», русская транскрипция [ле]), Њњ (лигатура «нь», русская транскрипция [не]), Ћћ (русская транскрипция [че]), Џџ (означает аффрикату /dʒ/, русская транскрипция [дже]). Алфавит болгарского языка (таблица 7) относится к кириллице и состоит из 30 символов. Алфавит английского языка (таблица 8) относится к латинице и состоит из 26 символов. Алфавит французского языка (таблица 9) относится к латинице, в нем используется 26 латинских символов, а также еще 14 диакритических знаков и 2 лигатуры. Характерными символами этого языка являются: Àà, Ââ, Ææ, Çç, Èè, Éé, Êê, Ëë, Îî, Ïï, Ôô, Œœ, Ùù, Ûû, Ÿÿ. Алфавит немецкого (таблица 10) языка относится к латинице. Используются 26 латинских символов, а также еще 3 умляутированные буквы и лигатура ß. Характерными символами этого языка являются: Ää, Öö, ß («эсцет»). Алфавит испанского (таблица 11) языка относится к латинице и состоит из 27 символов. Характерным символом этого языка является Ññ («n мягкое»). Алфавит итальянского языка (таблица 12) относится к латинице и состоит из 21 символа, а также в языке используются еще 5 дополнительных символов для записи слов иностранного происхождения.

В процессе разработки подсистемы АСОИБ для принятия решения о национальном языке представления информации был успешно разработан алгоритм (см. таблицу 13), в основе которого лежат вышерассмотренные признаки, применяемые на практике.

Таблица 13. Алгоритм определения национального языка представления информации

№ шага	Описание шага (пункта) алгоритма
1	В исследуемом тексте выбрать символы, соответствующие заданным образцам возможных символов. Каждый выбранный символ является некоторым признаком j_i (для более точного определения языка текста рекомендуется выбрать максимально возможное количество признаков).
2	Подсчитать число найденных признаков и присвоить его переменной k.
3	Принимается решение в зависимости от количества k выбранных признаков: если не определен ни один признак (т. е. $k = 0$), выдается сообщение о предложении попробовать определить признаки еще раз, т. е. перейти к шагу 1; если же признаки выбраны (т. е. $k > 0$), то перейти к шагу 4.
4	Переменной L присваивается значение множества языков, заранее подготовленных для АСОИБ.
5	Установить $i = 1$ (это означает рассмотрение первого признака, т. е. символа текста).
6	Принимается решение в зависимости от числа обработанных признаков: если обработаны не все признаки ($i \le k$), то перейти к шагу 7; если обработаны все признаки (т. е. $i > k$), то перейти к шагу 10.
7	Присвоить переменной m значение множества языков, соответствующее признаку j_i.
8	Учитывая значение переменной m, переменной L присваивается новое значение — множества языков, соответствующее признаку j_i, с учетом выбранных на предыдущих этапах языков, содержащих признаки j_n, где $i > n$ ($L := L \cap m$).
9	Выполнить $i = i + 1$. Повторить шаги 6—8 (количество итераций равно числу выбранных признаков).
10	Составить отчет о принятом решении (т. е. о языке или языках представления информации).

Следует отметить, что при работе АСОИБ (как и для систем криминалистического назначения [11]) для принятия решения о национальном языке представления информации желательно иметь как можно больший ее объем (если не встречаются характерные символы). Однако в некоторых случаях возможна однозначная идентификация даже по одному символу. Алгоритм определения языка представления информации был успешно реализован в АСОИБ.

В качестве сервера БД была выбрана СУБД Firebird 2.0. Для определения национального языка представления информации применяется специальная процедура расчета, написанная на языке Firebird PSQL. Для программной реализации алгоритма была разработана и затем заполнена данными БД, содержащая данные о буквах алфавитов с их характеристиками. В дальнейшем следует рассмотреть возможность применения статистических методов и нейросетевого алгоритма [10].

Выводы

Выполнен краткий анализ возможных искажений информации строго формализованных документов (объектов). Разработаны специальные шаблоны (рекомендации) для генерации искажений информации, связанной с поддельными документами. Разработан алгоритм работы генератора искажений информации, связанной со строго формализованными документами заданного вида. Разработан обобщенный алгоритм работы генератора.

В рассмотренных 10 языках был выделен ряд признаков, по которым можно достаточно точно идентифицировать язык представления информации. Этими признаками являются символы используемого алфавита, наличие их уникальности и частота повторения, а также некоторые характерные слова этого языка и характерные сочетания символов.

Разработан обобщенный алгоритм определения языка представления информации. Выполнена реализация этого алгоритма в АСОИБ. Экспериментальная проверка показала, что алгоритм работоспособен и с его помощью можно реально определять язык представления информации и тем самым выявлять фальшивые документы, обеспечивая информационную безопасность. По результатам проведенных исследований была подана заявка на выдачу патента на изобретение в виде полезной модели и получено положительное решение о выдаче этого патента.

СПИСОК ЛИТЕРАТУРЫ:

1. Регистрация преступлений в России за январь – декабрь 2002 г. М.: ГИЦ МВД России, 2003. – 24 с.
2. *Кулик С. Д., Ткаченко К. И.* Инструментальные средства выявления поддельных лекарств // Научная сессия МИФИ – 2008. XV Всероссийская науч. конф. «Проблемы информационной безопасности в системе высшей школы». Сб. науч. трудов. М.: МИФИ, 2008. С. 89–91.
3. *Кулик С. Д., Никонец Д. А.* Автоматизация криминалистического исследования рукописных документов и вопросы безопасности // Научная сессия МИФИ – 2008. XV Всероссийская науч. конф. «Проблемы информационной безопасности в системе высшей школы». Сб. науч. трудов. М.: МИФИ, 2008. С. 88–89.
4. Эксперт. Руководство для экспертов органов внутренних дел. М.: КноРус, Право и закон, 2003. – 592 с.
5. *Кулик С. Д.* Свидетельство на программу № 2000610698, РФ, «Генератор программ с фактографическими данными о ценных бумагах РФ» (GEN–FAKT) / С. Д. Кулик (Россия). – Заявка № 2000610525; Заяв. 02.06.2000; Зарегистр. 01.08.2000. Бюл. № 4 (33). С. 101–102. (РОСПАТЕНТ).
6. *Кулик С. Д., Ткаченко К. И.* Подсистема генерирования задач // Научная сессия МИФИ – 2007. Сб. науч. трудов в 17 т. М.: МИФИ, 2007. Т. 12. С. 19–21.
7. *Кулик С. Д., Ткаченко К. И.* Генератор изменений текста // Научная сессия МИФИ – 2008. Сб. науч. трудов в 15 т. М.: МИФИ, 2008. Т. 13. С. 83–84.
8. *Гиляревский Р. С., Гривнин В. С.* Определитель языков мира по письменностям. М.: Наука, 1964.
9. *Кулик С. Д., Никонец Д. А., Воронкова М. М., Шелякова П. А.* Разработка подпрограмм определения пола по почерку и языка текста в АРМ эксперта-криминалиста «FHWE V. 1.0» // Научная сессия МИФИ – 2008. Сб. науч. трудов в 15 т. М.: МИФИ, 2008. Т. 13. С. 81–82.
10. *Никонец Д. А.* Автоматизация криминалистического исследования рукописных текстов при помощи нейронных сетей // Труды РНТОРЭС им. А. С. Попова. 10-я Международная конференция и выставка. Цифровая обработка сигналов и ее применение. Выпуск X – 2. М.: РНТОРЭС, 2008. С. 693–697.
11. *Кулик С. Д., Никонец Д. А., Ткаченко К. И., Жижилев А. В.* Заявка на выдачу Свидетельства на полезную модель, РФ (RU), кл. МПК[7] G 07 D 7/00. Устройство определения фальшивых рукописных документов на русском языке. Заявка № 2007147832/20; Заяв. 25.12.2007; Приоритет от 25.12.2007. (РОСПАТЕНТ).
12. *Кулик С. Д.* Проектирование АФИПС криминалистического назначения // Безопасность информационных технологий. 2002. № 1. С. 78–81.