УДК 007.51+004.021

С.В. Минаков, О.А. Финько

ПОВЫШЕНИЕ ДОСТОВЕРНОСТИ ХРАНЕНИЯ И ПЕРЕДАЧИ ПЕРВИЧНЫХ ТЕКСТОВ НА ОСНОВЕ ГИБРИДНОЙ СЕМАНТИКО-КОДОВОЙ ИЗБЫТОЧНОСТИ

Предложена методика повышения достоверности и информационной живучести текстовых естественноязыковых данных путем использования гибридной семантико-кодовой избыточности применительно к семантическим единицам текста, наиболее уязвимым к смысловым ошибкам.

Достоверность текста; достоверная передача шифрованной информации избыточное кодирование; расстояние Левенштейна.

S.V. Minakov, O.A. Finko

INCREASING TO VALIDITY OF KEEPING AND ISSUES PRIMARY TEXT ON BASE HYBRID SEMANTICS-CODE REDUNDANCY

The Offered methods of increasing to validity and information vitality text data by by use hybrid semantics-code redundancy to semantic unit of the text with reference to, the most vulnerable to semantic mistake.

Validity of the text; the reliable transmission to encoded information surplus coding; distance Levenshteyna

К системам передачи шифрованной информации так же, как и к другим системам передачи, предъявляется ряд требований по обеспечению достоверности сообщений. Для этого шифрованные данные обычно подвергаются избыточному кодированию на основе известных методов помехоустойчивого кодирования. Однако, не исключая применения известных методов, могут использоваться и другие специальные методы, учитывающие специфику используемых шифров (шифров на основе помехоустойчивых кодов [1], шифров, не размножающих искажений [2, 3] и пр.) и конкретные особенности передаваемых данных (алфавитно-цифровые сообщения, изображения, звуковые данные и пр.).

Так, алфавитно-цифровые данные, представленные на естественном языке, имеют ряд ярко выраженных особенностей, связанных со статистической окраской следования символов и словарных величин и неравномерностью смыслового содержания, которые могут быть использованы для повышения достоверности шифрованной связи на этапе обработки первичного текста с целью его подготовки к процедуре зашифрования. Первичным текстом будем называть результат преобразования внутренней речи в письменную.

Примем следующее допущение: в тексте всегда присутствуют словарные величины (СВ – слова, фразы, цифровые последовательности), которые более чем другие СВ определяют смысловое содержание открытого текста. К таким СВ можно отнести, например, координаты (цели, пункта назначения), указание времени, наименования объектов, должностных лиц, пароли-ключи и пр. Назовем такие СВ критическими (КСВ).

Следовательно, достоверность передачи КСВ должна быть выше основного текста. При этом можно говорить о неравномерности избыточности СВ, естественным образом присутствующей в тексте, которая может быть использована для обеспечения достоверности его передачи. При этом важное значение приобретают свойства шифра по размножению искажений.

Примем второе допущение – применяется шифр, не размножающий искажений [2, 3] (например, шифр гаммирования).

Методы коррекции естественного текста широко известны. Однако для повышения достоверности КСВ можно применять специальные методы, например методы избыточного (разделимого) кодирования.

Определение критических словарных величин. При использовании методов помехоустойчивого кодирования в качестве метрики применяют расстояние Хэмминга, под которым понимают количество различающихся символов между сравниваемыми кодовыми комбинациями [2]. Для оценки текстовых данных известна метрика, основанная на понятии расстояния редактирования (Левенштейна) [4]. Расстояние редактирования вычисляется методом динамического программирования, согласно которому последовательно, по предыдущим значениям, вычисляются расстояния между все более и более длинными префиксами двух строк – до получения окончательного результата. Величиной расстояния между строками x и y, длины которых равны соответственно i и j, является $d_{i,j}$, т.е. $d_{i,j} = d\left(x(1,i), y(1,j)\right)$.

Цена преобразования символа a в символ b обозначается через $w(a,b)$. Таким образом, $w(a,b)$ – это цена замены одного символа на другой, когда $a \to b$, $w(a,\varepsilon)$ – цена удаления a, а $w(\varepsilon,b)$ – цена вставки b, где ε – специально введенный, для расширения алфавита, пустой символ. В случае, когда выполнены нижеследующие условия, $d_{i,j}$ является расстоянием Левенштейна:

$$w(a_i \to \varepsilon) = 1;$$
$$w(\varepsilon \to b_j) = 1;$$
$$w(a_i \to b_j) = 1, \quad если \quad a_i \neq b_j;$$
$$w(a_i \to b_j) = 0, \quad если \quad a_i = b_j.$$

В процессе вычислений значения $d_{i,j}$ определяются с помощью рекуррентного соотношения

$$d_{i,j} = \min \begin{cases} d_{i-1,j} + w(a_i \to \varepsilon), \\ d_{i,j-1} + w(\varepsilon \to b_j), \\ d_{i-1,j-1} + w(a_i \to b_j), \end{cases}$$

где $d_{i,j}$ – цена трансформации элементов СВ текста в элементы СВ словаря, a_i – элементы СВ текста, b_j – элементы СВ словаря, ε – пустой символ, $w(a_i \to b_j)$ – цена замены символа a_i на символ b_j, $w(a_i \to \varepsilon)$ – цена замены символа a_i на символ ε (удаление символа a_i), $w(\varepsilon \to b_j)$ – цена замены символа ε на символ b_j (вставка символа b_j), и записываются в массив размерностью $(m+1)(n+1)$. КСВ будут являться словарные величины, для которых $d_{\min}$ меньше либо равно заданному минимальному расстоянию редактирования.

Далее приведен массив (табл. 1), полученный при вычислении расстояния Левенштейна между строками «белый» и «серый». Из него видно, что расстояние между этими строками, т.е. $d_{\min}$, равно двум.

Таблица 1

Пример вычисления расстояния Левенштейна

	j	0	1	2	3	4	5
i			**б**	**е**	**л**	**ы**	**й**
0		0	1	2	3	4	5
1	**с**	1	1	2	3	4	5
2	**е**	2	2	1	2	3	4
3	**р**	3	3	2	2	3	4
4	**ы**	4	4	3	3	2	3
5	**й**	5	5	4	4	3	**2**

Для ограничения множества КСВ и, следовательно, избыточной информации предложено формировать базу данных СВ, сгруппированную в онтологические ряды [5], и контролировать критерий соответствия КСВ минимальному расстоянию редактирования в подмножествах таких словарных величин. Элементами базы (табл. 2) онтологических рядов является подмножество КСВ всего множества сообщений, передаваемых в некоторой организованной системе связи с максимальной метрикой между КСВ.

Таблица 2

Пример базы онтологических рядов для анализируемого текста

Тематические признаки	Элементы онтологических рядов
…	…
Годы	…, 1939, 1940, 1941, 1942, 1943, 1944, 1945 …, 2008, 2009, 2010, …
Месяцы года	март, май, июнь, июль, сентябрь, октябрь, ноябрь, декабрь
Даты	01, 02, 03, 04, 05, … , 29, 30, 31
Время	00:00, 00:01, … 01:00, 01:01, 01:02, … , 23:58, 23:59
Вооружение (танки)	Т-34, Т-34-57, ОТ-34, ТО-34, КВ-1с, КВ-85, ИС-1, ИС-3, ИСУ-152, ИСУ-122, ИСУ-122С, …
Вооружение (самолеты)	Су-20, Су-17М, Су-17М1, Су-22, Су-17М2, Су-22М, Су-17М3, Су-22М3, Су-17М3, Су-22М4, Су-17М4, Су-22УМ, Су-17УМ, Су-22УМ3, Су-17УМ3, Су-22УМ3К, …
Географические наименования (города России)	Аскино, Асино, Белово, Бековo, Елово, Бирск, Бийск, Булаево, Бураево, Ванино, Панино, Вельск, Вольск, Ельск, Волжск, Вольск, Ильск, Казаченское, Казачинское, Каргополь, Каргополье, Кола, Ола, Оха, Костанай, Кустанай, Кочки, Кошки, Кунья, Курья, Луга, Луза, Майма, Майна, Морки, Горки, Нема, Неман, Нея, Бея, Нолинск, Долинск, Ногинск, Омск, Томск, Орск, Сорск, Тужа, Тула, Убинское, Уинское, Яр, Уяр, Яя, Кесово, Кетово, Руза, Луза, Луга, Кубинка-1, Кубинка-10, Кубинка-2, Кубинка-8, Калуга, Калга, Тула, Тума, Кимовск, Кировск, Плавск, Славск…

Окончание табл. 2

Тематические признаки	Элементы онтологических рядов
Цвета	Алый, Бежевый, Беж, Белый, Бирюзовый, Бурый, Голубой, Желтый, Зеленый, Коричневый, Красный, Малиновый, Оливковый, Оранжевый, Пурпурный, Пунцовый, Розовый, Серый, Синий, Фиолетовый, Хаки, Чёрный, …
…	…

Достоинство метода применения метрики Левенштейна – математическая точность выявления КСВ и наилучшее соответствие решаемой задаче.

Оператор-эксперт может принимать решение о включении дополнительных СВ в множество КСВ на основе эмпирической (экспертной) значимости.

Показателем оценки является факт выявления КСВ. Достоинство метода экспертной оценки – возможность учета важности смыслового содержания СВ, пока не всегда доступная точным методам.

Учитывая известную неопределенность понятия «смысл» текста и КСВ, оказывающие на него существенное влияние, методика определения КСВ, скорее всего, должна быть комплексной, т.е. включать в себя метод, основанный на вычислении расстояния редактирования, и экспертный (рис. 1).

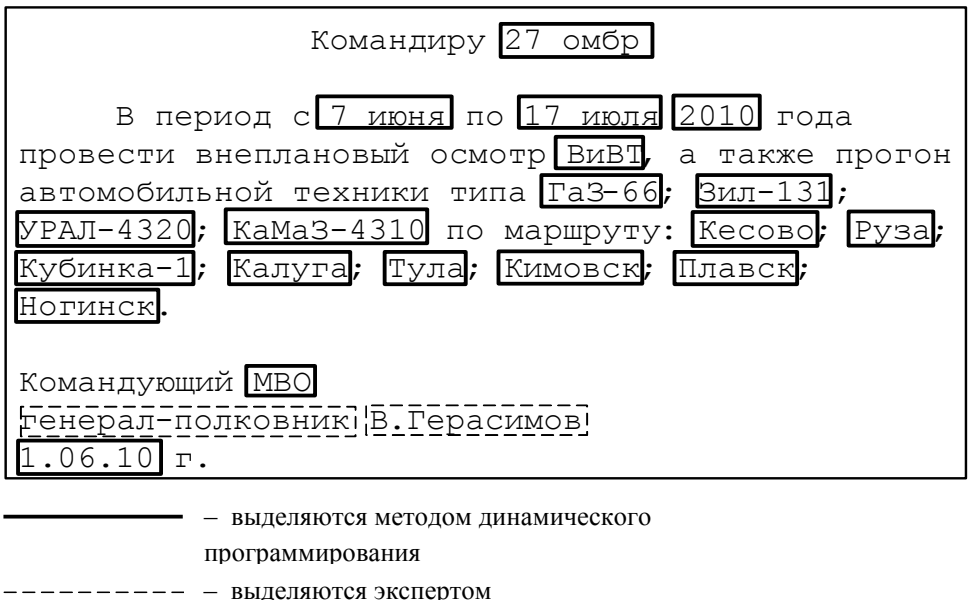

Рис. 1. Демонстрация метода комплексного выделения КСВ

Избыточное кодирование семантических единиц текста. При применении методов избыточного кодирования, увеличивается кодовое расстояние и тем самым повышается помехоустойчивость кода которым, по сути, является текст.

Текст выстраивается в виде матрицы. Элементами данной матрицы являются отдельные СВ. Далее строится маска КСВ в виде бинарной матрицы размерностью, равной размерности матрицы текста, в которой обычным СВ будут соответствовать «0», а КСВ – «1». КСВ посимвольно выстраиваются в отдельную квад-

ратную матрицу, подряд без пробела. Символы КСВ представлены кодом (в нашем случае – ASCII). Применяя избыточный код, вычисляют контрольные символы.

В результате получается маска КСВ и контрольные символы, которые могут храниться в служебной части сообщения либо отдельно от него.

В такой форме первичный текст может быть записан как на электронный, так и на бумажный носитель, может быть передан по каналам связи в открытом или шифрованном виде. Принимающая сторона определяет по полученной маске КСВ и выполняет процедуру декодирования. По результатам проверки принимается решение о достоверности принятого сообщения.

На рис. 2 представлена схема передачи шифрованной информации, использующая предлагаемую методику.

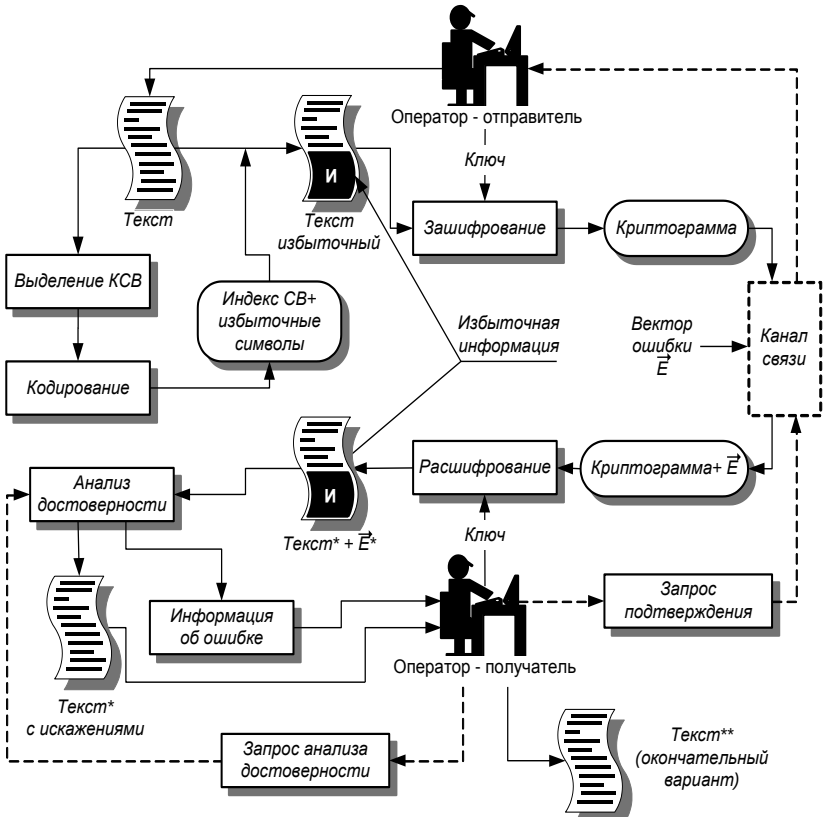

Рис. 2. Предлагаемая структурная схема помехоустойчивой передачи шифрованной информации

На рис. 3 и 4 представлены блок-схемы алгоритмов для предлагаемой гибридной методики определения КСВ на примере использования текстового анализатора.

Одним из путей повышения достоверности текстовых данных является предлагаемое применение гибридной семантической и кодовой (семантико-кодовая) избыточности. Это позволит избирательно защищать от искажений наиболее уязвимые СВ текста. Введение семантико-кодовой избыточности совершенствует процедуру выявления и устранения искажений, вызванных действиями оператора, шумов в канале связи, отказами аппаратуры, возникающими при обработке и передаче документированной информации.

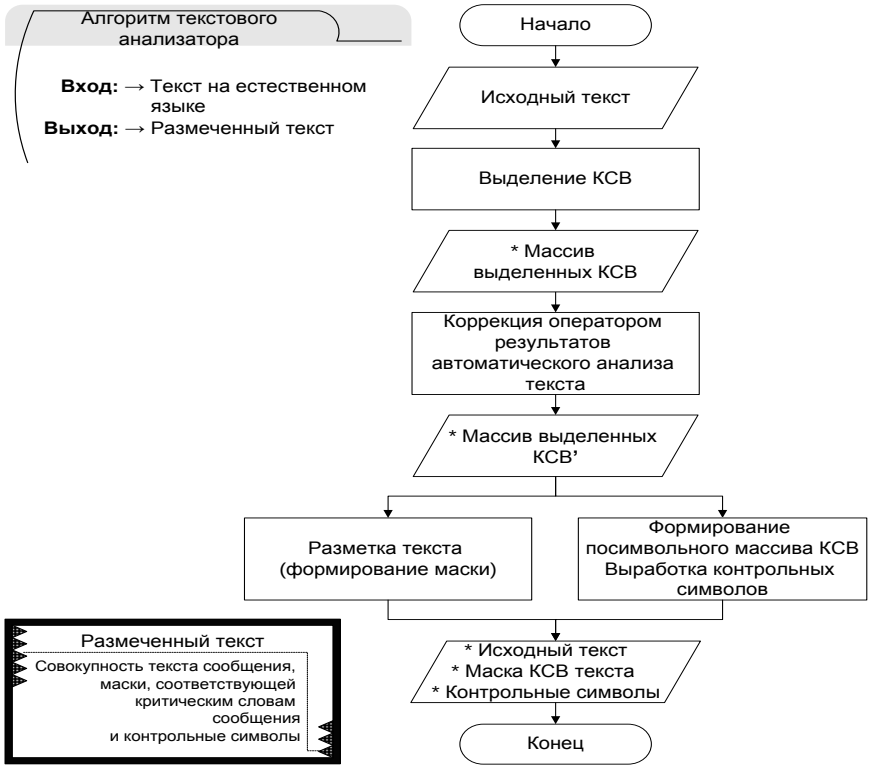

Рис. 3. Пояснение гибридной методики кодирования КСВ при отправке сообщения

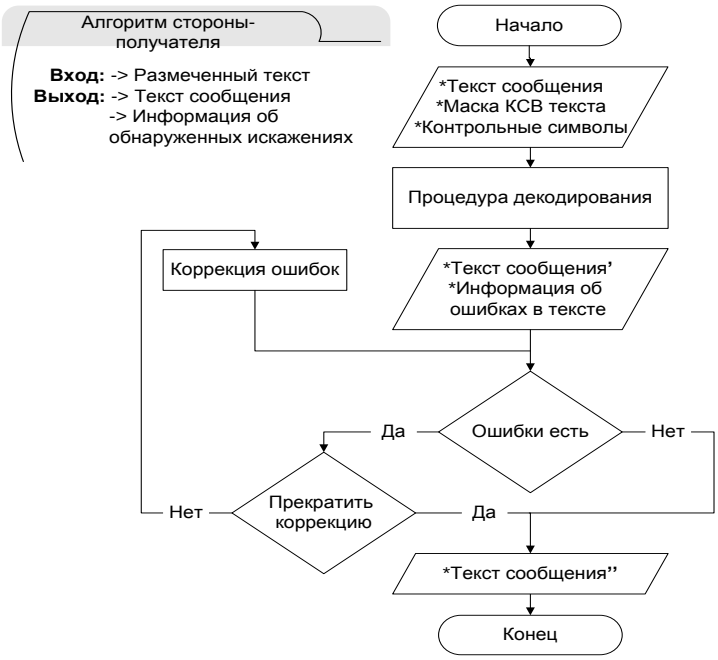

Рис. 4. Пояснение гибридной методики кодирования КСВ при получении сообщения

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ernst M. Gabidulin*, Public-key cryptosystems based on linear codes, Report DUT-TWI-95-30, Delft University of Technology, Delft, The Netherlads, 1995.
2. *Алфёров А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В.* Основы криптографии. – М., 2002. – 480 с.
3. *Бабаш А. В., Глухов М.М., Шанкин Г.П.* О преобразованиях множества слов в конечном алфавите, не размножающих искажений // Дискретная математика. – М., 1997. – Т. 9. – Вып. 3. – С. 3-19.
4. *Levenshtein V. I.* Binary codes capable of correcting deletions, insertions, and reversals // Sov. Phys. Dokl., February, 1966. – P. 707-710.
5. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем: Учебник для вузов. – СПб., 2000. – 384 с.

Минаков Сергей Викторович
Краснодарское высшее военное училище (ВИ).
E-mail: MinakovSergey1973@yandex.ru.
350035, г. Краснодар, ул. Красина, 4.
Тел.: +79184102497.

Финько Олег Анатольевич
Кубанский государственный технологический университет.
Институт информационных технологий и безопасности.
E-mail: ofinko@yandex.ru.
350072, г. Краснодар, ул. Московская, 2.
Тел.: +79615874848.

Minakov Sergey Viktorovich
Krasnodar higher military school (MI).
E-mail: MinakovSergey1973@yandex.ru.
4, Krasina, Krasnodar, 350035, Russia.
Phone: 79184102497.

Finko Oleg Anatol'evich
Kuban state technological university.
Institute of information technologies and safety.
E-mail: ofinko@yandex.ru.
2, Moscow, Krasnodar, 350072, Russia.
Phone: +79615874848.

УДК 511+519.719.2

Д.В. Самойленко, О.А. Финько

ПОМЕХОУСТОЙЧИВАЯ КРИПТОСИСТЕМА, ОСНОВАННАЯ НА КИТАЙСКОЙ ТЕОРЕМЕ ОБ ОСТАТКАХ, ДЛЯ *N* КАНАЛОВ С ШУМОМ И ИМИТИРУЮЩИМ ЗЛОУМЫШЛЕННИКОМ

Рассматривается помехоустойчивая модулярная криптографическая система, функционирующая в кольце Z_p положительных целых чисел по модулю p. Предложен алгоритм расширения системы оснований криптографической системы. Представлена оценка помехоустойчивости предложенной криптографической системы по отношению к традиционным (раздельным) методам помехо- и криптозащиты.

Достоверность; избыточное кодирование; Китайская теорема об остатках; криптоаналитик; криптосистема; модулярная арифметика; помехоустойчивость.