УДК 519.7

МЕТОД КОДИРОВАНИЯ ИНФОРМАЦИИ В КОМПЬЮТЕРНЫХ ТЕКСТАХ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

Ларионова К.Е., Губенко Н.Е.
Донецкий национальный технический университет

Лингвистическая стеганография – это наука, которая занимается скрытым внедрением кодированной произвольной информации в текстах, опираясь на лингвистические ресурсы. При этом требуется сохранить внешнюю «безобидность» и осмысленность несущего текста.

Целью данной работы является исследование лингвистических особенностей языков для повышения эффективности скрытого взаимодействия пользователей на основе моделирования соответствующих систем с использованием синонимических словарей.

Кодирование происходит путем замены одного синонима другим внутри синонимических групп, в которые входят слова исходного текста. Если это абсолютные синонимы, замены осуществляются независимо от контекста. В случае относительных синонимов все возможные синонимы и омонимы текстового слова до замены проверяются на совместимость с контекстом. Совместимость – это возможность вхождения в те же словосочетания, что и заменяемое слово. Опорными лингвистическими ресурсами являются специально подготовленные синонимические словари и обширная база языковых словосочетаний.

Ряду разработчиков подобных стегоалгоритмов удалось решить большинство проблем за счёт реализации идеи сочетания машинного и ручного способов стеговставки, объединяя их со стойкой криптографией. При этом они опирались на теорию лингвистической стеганографии из работ Bergmair и Katzenbeisser (2004) по проблемам машинного распознавания стеготекстов

и использования кодов Хаффмана для противодействия статистическому стегоанализу [1].

Однако на современном этапе развития информационного сообщества особо остро стоит проблема несанкционированного использования информации и несоблюдения авторских прав, что предъявляет особые требования к подобным алгоритмам.

Подтверждением тому является проведение ряда конференций на эту тематику, например в июле 2007 года прошла ECIW 2007 – шестая Европейская конференция по вопросам информационных войн и информационной безопасности.

Активное развитие лингвистической стеганографии обусловливает желание многих специалистов создать систему защиты, которая не была бы заметна.

Данная работа позволит разработать систему скрытого взаимодействия пользователей, реализованного методом кодирования информации в компьютерных текстах на основе лингвистических ресурсов, которая может быть использована для сокрытия факта передачи информации в виртуальных чатах, электронной почте, форумах и т.д. Данный программный продукт может быть использован при необходимости внедрить произвольную информацию в текст, сохранив при этом его контекст.

Разрабатываемая модель базируется на идеи использования словарей синонимов, а также на идеи «скрытого общения» двух объектов с использованием кодов Хаффмана и Диффи-Хелмана, идея которых рассмотрена в работе авторов K. Wouters, B. Wyseur и B. Preneel «Lexical Natural Language Steganography Systems with Human Interaction» – «Стеганографические системы, основанные на лексически естественных языках, работающие при взаимодействии с человеком» [2].

В качестве модели противника авторы данной работы предлагают не только программу-детектор, но и подготовленного человека (например, лингвиста), который пытается уловить все подозрительные и неестественные диалоги собеседников, которые бы указывали на наличие стегоканала.

В качестве среды для испытания протокола был выбран IRC-чат: в нём может одновременно общаться большое число людей, а условные пользователи Алиса и Боб могут не отправлять сообщения непосредственно друг другу, а обращаться только к другим пользователям. Это не позволит установить наличие прямого контакта между ними за один сеанс, кроме того, они могут быть более анонимными, используя чаты в сети.

Однако предложенная система имеет ряд ограничений, связанных не только с низкой пропускной способностью, но и с тем, что статистический анализ способен выявить употребление синонимов, нехарактерных для речи данного человека, если его личность установлена. Кроме того, бездумный выбор некоторых синонимов может привести к грамматическим ошибкам и потребует дополнительной внимательности пользователя к исправлению получившегося текста.

Предлагаемая нами модель скрытого взаимодействия на базе стеганографического алгоритма позволяет упростить процедуру взаимодействия пользователей за счет автоматического подбора синонимов с учетом особенностей кодируемой информации.

Исходными данными системы являются:

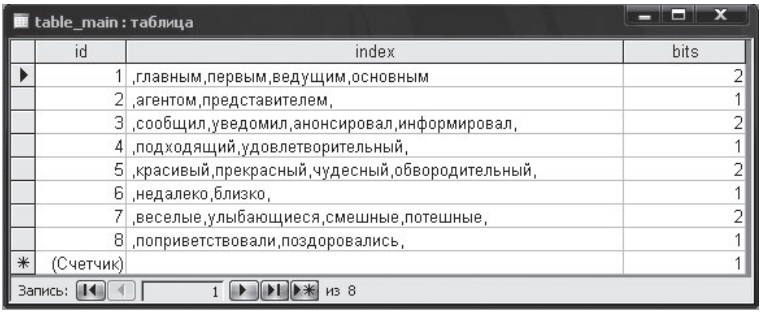
- информация, предназначенная для скрытой передачи;
- исходный несущий текст на русском языке на порядок превышающий объем исходной информации.

Алгоритм состоит из следующих шагов.

Шаг 1. Подготовка синонимического словаря и кодируемой информации.

Словарь представляет собой базу данных слов, которым соответствуют группы синонимов, пронумерованные двоичными числами (рис. 1, 2, 3). Он составляется вручную и является уникальным.

Информацию, которую необходимо закодировать, пользователь изначально вводит в специально отведенное окно в программе. Данная информация представляется в двоичном виде с помощью ASCII таблицы.



table_main : таблица

	id	index	bits
▶	1	,главным,первым,ведущим,основным	2
	2	,агентом,представителем,	1
	3	,сообщил,уведомил,анонсировал,информировал,	2
	4	,подходящий,удовлетворительный,	1
	5	,красивый,прекрасный,чудесный,обворотительный,	2
	6	,недалеко,близко,	1
	7	,веселые,улыбающиеся,смешные,потешные,	2
	8	,поприветствовали,поздоровались,	1
*	(Счетчик)		1

Запись: 1 из 8

Рисунок 1 – Главная таблица

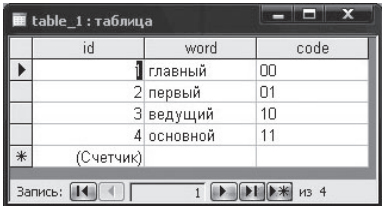


table_1 : таблица

	id	word	code
▶	1	главный	00
	2	первый	01
	3	ведущий	10
	4	основной	11
*	(Счетчик)		

Запись: 1 из 4

Рисунок 2 – Таблица синонимов двухбитового слова

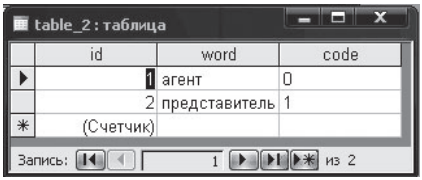


table_2 : таблица

	id	word	code
▶	1	агент	0
	2	представитель	1
*	(Счетчик)		

Запись: 1 из 2

Рисунок 3 – Таблица синонимов однобитового слова

Шаг 2. Поиск синонимичных слов.

Текст, вводимый пользователем, сканируется, и в нем выделяются те отдельные слова, которые являются статьями системного словаря и имеют синонимы.

Шаг 3. Кодирование.

Последовательность профильтрованных синонимических групп сканируется слева направо. При выявлении в тексте подходящих слов согласно информации, которую необходимо закодировать, программа производит автоматическую замену на синоним, который стоит под нужным порядковым номером в таблице подстановки (рис. 4).

Стеганографическая плотность, которая определяется отношением кодируемого фрагмента текста к контейнеру, при таком подходе невелика, но зато в достаточно протяженном тексте

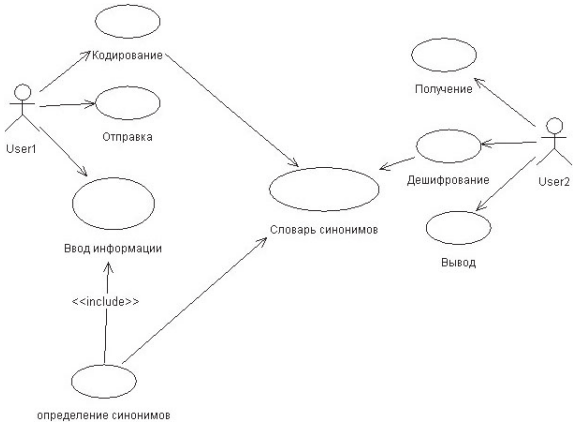


Рисунок 4 – Диаграмма прецедентов разработанного алгоритма

можно скрыть вполне содержательный объем информации.

Разработанная на данном этапе модель кодирования информации позволяет повысить эффективность процесса шифрования текстов за счет выбранного синонимического словаря. Анализ показывает, что алгоритм требует оптимизации за счет разработки модулей, учитывающих особенности падежных окончаний слов в русском языке, а также добавления ветви, позволяющей кодировать английские сообщения. Это позволит провести анализ результатов кодирования и сравнить какой из языков более подвержен искажениям при использовании алгоритмов с синонимической заменой.

Литература

[1] Разработана эффективная система стеганографии через чат [Электронный ресурс]. Режим доступа: http://www.itsec.ru/newstext.php?news_id=38490

[2] Большаков И.А. Использование синонимов, ограниченных контекстными словосочетаниями, для целей лингвистической стеганографии // Информационные процессы и системы, 2004, №5, С. 23-30.