УДК: 004.8

О новом методе обеспечения безопасности семантических вычислений

A.Yu. Shcherbakov

On a New Method for Securing Sematic Computations

Abstract. *In this article a new method for the separate exchange of keys and information between subscribers interested in the joint processing of undisclosed information is proposed, which includes set-theoretic operations on encrypted data. A new concept of using symmetric cryptographic algorithms for these purposes is considered.*

Keywords: *homomorphic encryption, set-theoretic comparison, intruder, separate key management.*

А.Ю. Щербаков
Доктор технических наук, профессор кафедры комплексной безопасности критически важных объектов РГУ нефти и газа (НИУ) имени И.М. Губкина, заведующий кафедрой когнитивно-аналитических и нейро-прикладных технологий РГСУ, ведущий научный сотрудник Государственного университета управления, главный научный сотрудник РАН (ИТМиВТ им.С.А.Лебедева).
E-mail: x509@ras.ru

Аннотация. *В статье предлагается новый способ раздельного обмена ключами и информацией между абонентами, заинтересованными в совместной обработке нераскрываемой информации, включая теоретико-множественные операции над зашифрованными данными. Рассматривается новая концепция применения для данных целей симметричных криптографических алгоритмов.*

Ключевые слова: *гомоморфное шифрование, теоретико-множественное сравнение, нарушитель, раздельное управление ключами.*

ВВЕДЕНИЕ. МОДЕЛЬ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ

В современной научной литературе в качестве конструктивной модели семантической обработки информации утвердилась модель, предполагающая первичное рассмотрение некоторых текстов, понимаемых как последовательности слов [1].

Для решения различных задач обработки текстов и их оптимизации используется процедура индексации – небиективное (неоднозначное) преобразование, заменяющее каждое слово текста двоичным вектором фиксированной длины.

В качестве примера индексации рассмотрим процедуру, в которой слова длины меньшей, чем MAX_WORD, заменяются вектором длины 8 байт. При этом преобразуемое слово устанавливается в качестве ключа в криптографическом алгоритме ГОСТ 28147-89 и происходит одна итерация этого алгоритма [2], при этом исходным вектором является фиксированный вектор, получаемый по процедуре for(i=0; i<8; i++) x[i]=0x88^(char)i.

```
#define MAX_WORD 512
int xb(char *wd, unsigned char *x)
 {
int i, j, len, part, ost;
unsigned char wd1[32];

len=strlen(wd);
if(len>MAX_WORD) return(-1);
part=len/32;
ost =len%32;
for(i=0;i<32 ;i++) wd1[i]=0;
for(i=0;i<ost;i++) wd1[i]=wd[i];

for(i=0;i<part;i++)
  for(j=0;j<32;j++) wd1[j]=wd1[j]^wd[i*32+j];

for(i=0; i<8; i++) x[i]=0x88^(char)i;
 imit_fast((unsigned long *)x,(unsigned long *)
wd1);
return(0);
 }
```

В данной статье не рассматриваются свойства описанного небиективного преобразования. Отметим только, что в приведенном примере совпадение индексов x для различных слов зависит от числа байт, которые выбираются для формирования индекса из восьми байт массива x. Это связано с тем, что рассматриваемый алгоритм в большой степени статистически совпадает со случайным отображением. На-

пример, если выбрать 4 байта, то вероятность совпадения векторов x для различных слов (вероятность коллизий) составит порядка 2^{-32}.

Таким образом, предлагается заменить тексты со словами различной длины массивами индексов одинаковой длины (практически длинными числами) и рассматривать различные операции уже с индексированными текстами.

СРАВНЕНИЕ ТЕКСТОВ

В самом простом случае рассмотрим два текста: T и R. Задача ставится следующим образом: необходимо конструктивно и вычислительно нетрудоёмко сравнить тем или иным образом эти тексты. Например, теоретико-множественное сравнение выполняется по «облаку» слов, входящих в эти тексты.

Используя диаграммы Эйлера для условно пересекающихся множеств, рассмотрим три множества: 1 – множество слов, входящих только в текст T, 2 – множество слов, входящих только в текст R и 3 – пересечение текстов T и R. Заметим, что множество 3 может быть и пустым. Объединение множеств 1, 2 и 3 совпадает с объединением текстов T и R, понимаемых как множества слов (упорядоченные или неупорядоченные).

Принимая априорно, что чем больше мощность множества 3, тем более возможно говорить о том, что тексты «сходны» между собой, конструктивно для оценки сходства текстов ввести следующие меры [3].

Обозначим m(i) – мощность множества i.

“Нулевая” мера (исторически введенная первой)

M0= 2m(3)/((m(1)+m(2))

«Верхняя» мера

M=0.5(m(3)/m(1)+m(3)/m(2))

«Нижняя” мера

M=m(3)/(m(1)+m(2)+m(3))

Несмотря на простоту, эта конструкция достаточно универсально работает для решения различных семантических задач.

Задавая в качестве R некоторые эталонные тексты и оценивая значения введенных мер на множествах 1-3, возможно решать задачи не только простого, но и расширенного поиска (по длинному нечеткому произвольному словесному описанию), определения принадлежности текста к некоторой тематике, делать выводы об авторстве текста, а также оценивать по множеству 1 (возможно, даже по его мощности в первом приближении) оригинальность текста и его новизну.

ЗАКРЫТЫЕ ВЫЧИСЛЕНИЯ НА ТЕКСТАХ

В настоящее время весьма актуальной является задача закрытых вычислений, когда зашифрованные данные обрабатывается неким методом без их расшифрования. В этом смысле уместно привести пример с системой электронного голосования, когда подаваемый в закрытом (зашифрованном) виде голос суммируется без его раскрытия с другими голосами, также без их раскрытия (расшифрования), для получения итогов голосования. В этом случае весьма уместно использовать различные методы гомоморфного шифрования [4].

Согласно этому источнику под гомоморфным шифрованием понимается криптографический примитив, представляющий собой функцию шифрования, удовлетворяющую дополнительному требованию гомоморфности относительно каких-либо алгебраических операций (op) над открытыми текстами.

Пусть *E(k,m)* — функция шифрования, где *m* — открытый текст, *k* — ключ шифрования. Заметим, что для данных фиксированных *k* и *m* криптограмма (зашифрованный текст) *E(k,m)* может быть, вообще говоря, случайной величиной. В таких случаях говорят о вероятностном шифровании. Функция *E* гомоморфна относительно операции op над открытыми текстами, если существует эффективный алгоритм *M*, который, получив на вход любую пару криптограмм вида *E(k,m1)*, *E(k,m2)*, выдает такую криптограмму *c*, что при расшифровании *c* будет получен открытый текст *m1opm2*.

Как правило, рассматривается следующий важнейший частный случай гомоморфного шифрования. Для данной функции шифрования *E* и операции *op1* над открытыми текстами существует операция *op2* над криптограммами такая, что из криптограммы *E(k,m1)op2E(k,m2)* при расшифровании извлекается открытый

текст *m1op1m2*.

В случае сравнения текстов методы вычисления «под шифром» не подходят, поскольку необходимо производить теоретико-множественное сравнение.

ТЕОРЕТИКО-МНОЖЕСТВЕННОЕ СРАВНЕНИЕ ЗАКРЫТЫХ ДАННЫХ

Рассмотрим формальную постановку задачи, связанную с работой по сравнению текстов и расширенному поиску. Под расширенным поиском как раз будем понимать поиск по максимальной величине одной из мер ранее введенного сходства текстов – поискового запроса и массива информации, в которой ведется поиск.

Введем владельца информации (целевых данных) V достаточно большого объема и текстового вида. Это могут быть списки покупок (от торговой сети), списки клиентов, перечень медицинских симптомов и показаний и т.д. Важно, что в любом случае у владельца есть персональные данные, нуждающиеся в защите по закону, либо конфиденциальная информация. Пусть также есть потенциальный потребитель (интересант) I, который нуждается в некоторой выборке этих данных. Потребитель I также может рассматриваться в качестве нарушителя (может проявить интерес к раскрытию или переносу в свои хранилища данных владельца V).

Рассмотрим также посредника P (владельца вычислительных ресурсов), принимающего поисковый запрос от I и обрабатывающий данные от владельца V. В нашей модели он также является потенциальным нарушителем, поскольку может проявить интерес как к содержанию запроса I, так и информации, которой владеет V. Посредник не имеет ключа K и не имеет доступа к индексам.

Напомним также, что в классической криптографии для установления связи между абонентами необходимо обменяться ключами (или иметь одинаковые или ассиметричные ключи для шифрования и расшифрования информации), а также обменяться зашифрованной на этих ключах информацией.

Предлагается «разорвать» эту парадигму отдельно на обмен ключами и обмен информацией. Очевидно, что обменявшись ключами, но не обменявшись никакой информацией, стороны не могут нарушить конфиденциальность друг друга.

Соответственно, предлагается проиндексировать текстовые массивы владельца V и получить вместо слов или текстовых данных массив B индексов. Аналогично, поисковый запрос X от I также индексируется в Y. После этого стороны зашифровывают каждый индекс на общем для I и V ключе K.

Сформулируем важное утверждение.

При реализации шифрующего преобразования y=E(k, x) мощности множеств 1, 2 и 3 (введенные выше), используемые для сравнения текстов B и Y, будут совпадать с соответствующими мощностями множеств 1, 2 и 3 для E(K, B) и E(K, Y).

Верность этого утверждения следует из однозначности алгоритма шифрования при фиксации каждого ключа K.

Тогда зашифрованные владельцем информации и интересантом индексы от массива, по которому ведется поиск и поискового запроса могут быть направлены посреднику, где над ними будет произведено сравнение в теоретико-множественном смысле и результат сходства может быть сообщен одной или обеим сторонам.

Поскольку каждый индекс шифруется отдельно, то посредник, рассматриваемый в качестве нарушителя, может статистически анализировать шифр простой замены, заданный на отдельных словах (полагаем, что язык текстов ему известен).

При этом очевидно, что поисковый запрос имеет не очень большую длину и такая атака неконструктивна. Опасение представляет массив B владельца, который может быть достаточно большим.

ЗАКЛЮЧЕНИЕ

Предложенный алгоритм раздельного обмена ключами и информацией, а также теоретико-множественные операции над зашифрованными данными позволяют решить задачу обработки текстов без нарушения их конфиденциальности.

СПИСОК ЛИТЕРАТУРЫ

1. Рязанова А.А., Анисимова А.Э. О методике сравнительного квалификационного анализа требований к профессиональным навыкам с целью коррекции национальных образовательных программ // Научно-технический сборник "Научно-техническая информация", сер. 2 Информационный процессы и системы, 2019. № 2. С. 29-35.

2. ГОСТ Р 34.12 – 2015. Информационная технология. Криптографическая защита информации. Блочные шифры. Национальный стандарт РФ: утвержден и введен в действие Приказом Федерального агенства по техническому регулированию и метрологии от 19 июня 2015 г. № 749-ст.: введен впервые: дата введения 01.01.2016. – Москва: Стандартинформ, 2018.- 25 с.

3. Рязанова А.А., Щербаков А.Ю. К вопросу о метриках сходства текстов для методов их автоматизированного сравнения // Приоритетные задачи и стратегии развития технических наук. Выпуск II. Сборник научных трудов по итогам международной научно-практической конференции (25 мая 2017 г.), г. Тольятти. С. 66-69.

4. Варновский Н.П., Шокуров А.В. Гомоморфное шифрование // Сборник трудов ИСП РАН №12. Т.27. 2007. С. 27-35.