СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2020. – № 1–2 (97).** – 99–112

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.6 DOI: 10.17212/2307-6879-2020-1-2-99-112

СРЕДСТВА АНАЛИЗА ТЕКСТОВ НА ОСНОВЕ КРИПТОАНАЛИЗА ПРОСТОЙ ЗАМЕНЫ*

Ю.А. КОТОВ[1], Д.И. МАКАРСКАЯ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат физико-математических наук, доцент, доцент кафедры защиты информации. E-mail: kotov@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: dashamakarskaya96@mail.ru*

В статье рассматриваются средства анализа текстов на основе числовой оценки, полученной с использованием криптографического подхода. Он заключается в применении методов частотного криптоанализа простой замены к открытому тексту. В этом случае данные методы сводятся к методам идентификации букв текста, а задача криптоанализа – к задаче идентификации букв текста. Для методов частотного криптоанализа простой замены определяется ошибка идентификации как выборочная статистика ряда ошибок криптоанализа для определенных объемов текстов. Эта числовая оценка, называемая добротностью метода, может быть перенесена с методов на тексты. Анализ текстов на основе добротности текстов включает в себя выбор криптографического метода, с помощью которого дается количественная оценка текста, и последующее вычисление добротности вектора текстов. В целях проведения анализа текстов на основе криптоанализа простой замены определяются основные этапы такого анализа и необходимые для него средства. Эти средства реализованы в графической оболочке MS Access и включают в себя 30 вкладок навигации, на которых расположены 143 информационно-управляющих элемента. Приведены примеры реализации некоторых из рассмотренных инструментов.

Ключевые слова: анализ, тексты, криптоанализ, идентификация, добротность, простая замена

ВВЕДЕНИЕ

Задача анализа текстов тесно связана с их классификацией. Возможны различные подходы к классификации текстов, обусловленные целями исследований или практического приложения: библиографическая классификация, тематическая классификация, универсальная десятичная классификация

* Статья получена 11 мая 2020 г.

(УДК) и т. д. Например, задачами библиографии являются: стандартизация библиографической деятельности, составление библиографических указателей и индексов цитирования, классификация документов [1]. При этом имеется ряд стандартов, применяемых к библиографическим записям (например, ISBD [2]). Для стандартов важными параметрами являются области вида содержания и типа средств.

Проблема классификации текстов по содержанию может быть решена с помощью использования субъективной классификации, например, универсальной десятичной классификации [3]. Более объективную классификацию текстов по содержанию можно получить, используя методы тематического моделирования. Тематическая модель – это модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции [4, 5].

Таким образом, в рассмотренных подходах к классификации текстов в одних случаях классификационные признаки извлекаются непосредственно из текста (объем, частота слов и словосочетаний) или его стандартизованного (в том числе библиографического) описания (автор, название, источник, объем), в других – приписываются ему автором или другими уполномоченными лицами (УДК, ISBD и др.). Однако в целях исследования и решения криптографических задач [6, 7], для которых предназначена система Tbase [8–10], необходима более точная числовая оценка текста, полностью свободная от какого-либо субъективизма. Она может быть сформирована на основе криптографического подхода, как показано в работе [11]. Для проведения анализа текстов на ее основе требуется определить основные этапы такого анализа и реализовать необходимые для него инструменты, которые и рассматриваются в настоящей работе.

1. МЕТОДОЛОГИЯ АНАЛИЗА И ПОСТАНОВКА ЗАДАЧИ

Задача идентификации букв текста заключается в сопоставлении знака произвольного текста на некотором языке с определенной буквой соответствующего алфавита на основе числовых характеристик, получаемых из данного текста [11]. В работе [11] была определена выборочная статистика, названная добротностью метода, вычисляемая по формуле

$$Q(x) = 1 - O_1(x)O_2(x)O_3(x), \qquad (1)$$

где x – объем текста в знаках; O_1 – ошибка, определяемая как отношение количества текстов, в которых обнаружена хотя бы одна ошибка идентификации

букв, к общему количеству обработанных текстов; O_2 – ошибка, определяемая отношением количества неправильно идентифицированных букв к общему количеству букв в отдельном обрабатываемом тексте; O_3 – ошибка, определяемая отношением суммарного количества появлений в отдельном тексте неправильно идентифицированных букв к общему объему данного текста.

Аналогичным образом с (1) может быть определена добротность текста, вычисляемая по формуле

$$Q(x, y, z) = Q_z(x), \tag{2}$$

где x – объем страницы в знаках; y – номер (идентификатор) текста; z – номер (идентификатор) метода.

В этом случае некоторый текст y последовательно разделен на страницы объемом x с применением к этому множеству страниц метода идентификации z. Для определенности примем следующие замечания.

1. При разделении текста на страницы объемом x необходимо решить проблему не полностью заполненных (последних) страниц. При определении добротности текста такие страницы отбрасываются.

2. Объем любого текста в знаках всегда известен.

Пример изменения добротности Q 12 текстов при x = 10 000 методом DAT [12] в зависимости от текста представлен на рис. 1.

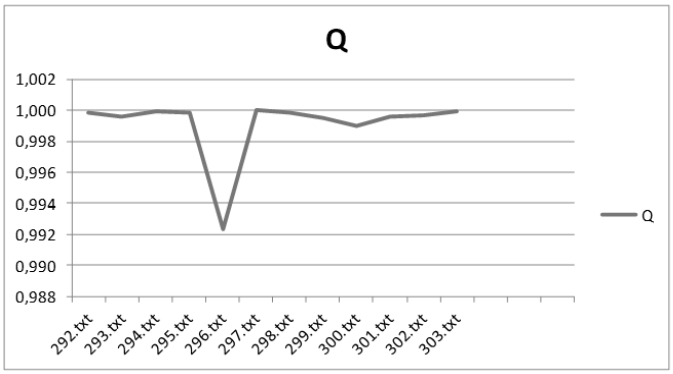

Рис. 1. График изменения значений Q

Данные, используемые для формирования графика, получены из таблицы.

Значения добротности текстов *Q*

x = 10 000	O_1	O_2	O_3	Q
292.txt	0.14	0.071	0.0133	0.999868
293.txt	0.2	0.0871	0.0237	0.999587
294.txt	0.05	0.0737	0.0149	0.999945
295.txt	0.08	0.086	0.0254	0.999825
296.txt	0.17	0.192	0.2359	0.992300
297.txt	0	0	0	1
298.txt	0.37	0.0655	0.0082	0.999801
299.txt	0.27	0.0806	0.0239	0.999480
300.txt	0.27	0.097	0.0373	0.999023
301.txt	0.2	0.0806	0.0251	0.999595
302.txt	0.32	0.0645	0.0154	0.999682
303.txt	0.33	0.0597	0.0051	0.999899

Можно заметить, что значения Q изменяются в диапазоне от нуля до единицы $0 \le Q \le 1$. Примем 0.5 в качестве пороговой погрешности для ошибок O_1, O_2, O_3. Если одна из ошибок принимает подобное значение, можно сделать вывод о том, что половина текста повреждена или текст имеет определенные особенности, которые не распознаются программой. Тогда можно оценить нижнюю границу добротности текста как $Q = 1 - 0.5 \cdot 0.5 \cdot 0.5 = 0.875$. В случаях, когда добротность текста имеет значение ниже 0.875, следует повторно проанализировать текст и принять решение о необходимости редактирования или удаления подобного текста.

Примем 0.1 в качестве пороговой погрешности для ошибок O_1, O_2, O_3. В результате можно оценить добротность текста, близкую к максимальной:

$$Q = 1 - 0.01 \cdot 0.01 \cdot 0.01 = 0.999999.$$

То есть при значениях ошибок O_1, O_2, O_3, равных 0.1, значение Q будет отличаться на 0.000001 от максимума (единицы). В этой связи более удобной может оказаться процентная форма добротности текста, $Q \cdot 100\ \%$.

Пусть для одного текста O_1 и O_2 равны 0.3 и 0.5 соответственно, для другого текста O_1 и O_2 равны 0.5 и 0.3 соответственно.

Тогда добротности этих текстов будут равны:

$$Q_1 = 1 - 0.3 \cdot 0.5 \cdot 0.1 = 0.985;$$

$$Q_2 = 1 - 0.5 \cdot 0.3 \cdot 0.1 = 0.985;$$

$$Q_1 = Q_2.$$

При этом ошибки O_1 и O_2 имеют различное значение: O_1 – количество страниц, содержащих хотя бы одну ошибку; O_2 – количество неверно идентифицированных букв по отношению к буквам алфавита. Данную особенность необходимо учитывать при сравнении текстов с одинаковыми значениями Q.

Оценку (2) можно обобщить на множество размеров страниц, текстов и методов (3):

$$Q(X,\ Y,\ Z), \tag{3}$$

где X – множество объемов страницы в знаках; Y – множество номеров (идентификаторов) текстов; Z – множество номеров (идентификаторов) методов.

Тогда для множества текстов Y значение добротности Q_x всех текстов множества формируется как статистическое объединение добротностей каждого текста, входящего во множество, при определенном объеме страницы; для множества объемов страниц X – как среднее $\widetilde{Q_x}$ для всех входящих во множество страниц, а для множества Z – как отдельные значения Q_x для каждого метода или как минимальные, максимальные или средние значения из полученных значений Q_i каждым методом для каждого текста.

Все рассмотренные особенности определения добротности текстов (3) необходимо учитывать при определении этапов анализа и реализации соответствующих инструментов.

2. ЭТАПЫ АНАЛИЗА ТЕКСТОВ НА ОСНОВЕ ДОБРОТНОСТИ

Анализ текстов на основе криптоанализа простой замены [13] может включать в себя 6 этапов:

1) подготовка первичных данных, включающая в себя выбор множеств текстов, объемов страниц и методов для анализа;

2) упорядочивание множества текстов по добротности;

3) классификация множества текстов по добротности;

4) сопоставление добротности, частотных и атрибутивных характеристик текстов в зависимости от объемов текстов и используемых методов;

5) возврат к этапу 1 или 2 при необходимости;

6) фиксация результатов анализа.

По завершении подготовки первичных данных для анализа формируется вектор (множество) текстов. На входе отдельно указываются размерность страниц и методов для анализа. Далее осуществляется упорядочивание множества текстов по добротности. За счет деления текстов на группы по значениям Q проводится классификация множества текстов по добротности. Полученные результаты сопоставляются. В случае, если полученных результатов недостаточно, возможен повтор алгоритма.

Для некоторых текстов после получения и фиксации результатов может быть принято решение: оставить текст без изменений, отредактировать текст или удалить текст [14].

Результатами анализа добротности текстов являются:

1) определение стандартного размера страницы и метода для вычисления добротности текстов;

2) вычисление и сохранение значений добротности каждого текста;

3) разбиение диапазона значений добротности на стандартные поддиапазоны;

4) определение текстов, значения параметров которых не соответствуют заданным граничным условиям.

Стандартный размер страницы и метода подразумевает, что при заданных условиях результаты анализа будут наиболее точными. Значения добротности текстов, полученные для стандартного размера страниц и метода, сохраняются в базе. В результате анализа формируется выборка текстов с определенными значениями добротности для каждого текста. Далее определяются диапазоны и поддиапазоны значений добротности этих текстов. Тексты, значения добротности которых больше или меньше установленных норм (границ), могут быть помещены в карантин. Такие тексты исключаются из любой выборки и не влияют на результаты анализа.

3. РЕАЛИЗАЦИЯ АНАЛИЗА ТЕКСТОВ НА ОСНОВЕ ДОБРОТНОСТИ

Реализация средств анализа текстов на основе добротности выполнена в графической оболочке MS Access и включает в себя 30 вкладок навигации, на которых расположены 143 информационно-управляющих элемента (рис. 2).

Эти средства позволяют выполнять анализ в соответствии с указанными выше этапами, проводить классификацию текстов по добротности и соотно-

сить ее с основными частотными характеристиками текстов [15] (рис. 3–5), работать с выделенными текстами (рис. 6) и фиксировать результаты (рис. 7).

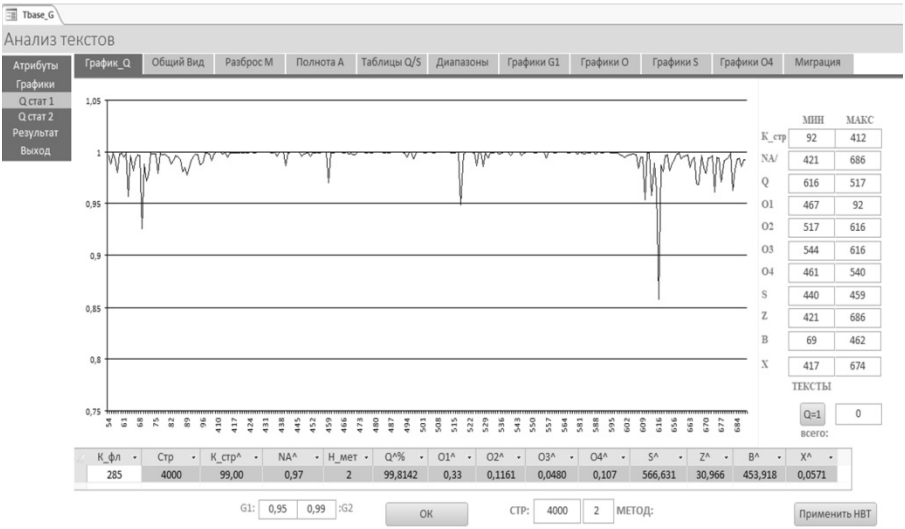

Рис. 2. Значения *Q* выбранного множества текстов

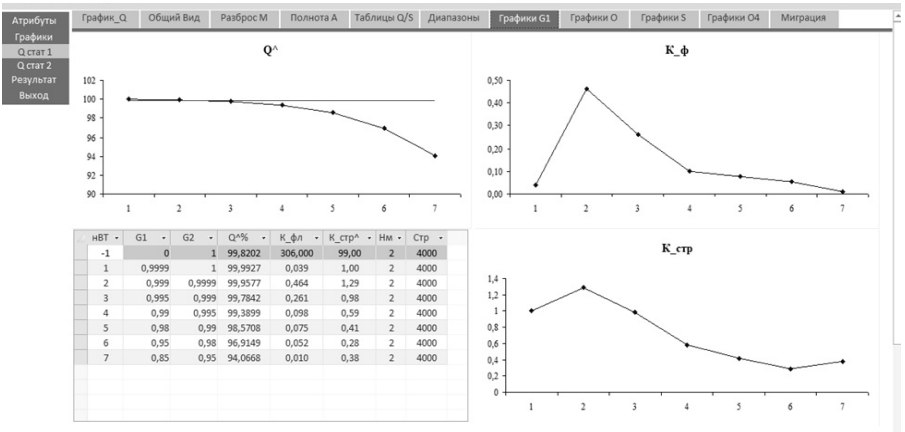

Рис. 3. Распределения средних значений *Q* по диапазонам

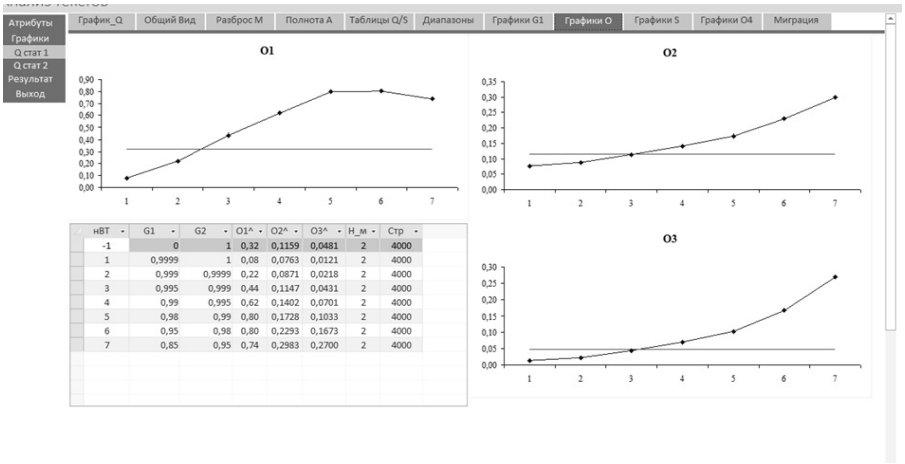

Рис. 4. Распределения средних значений ошибок O_1, O_2, O_3 по диапазонам

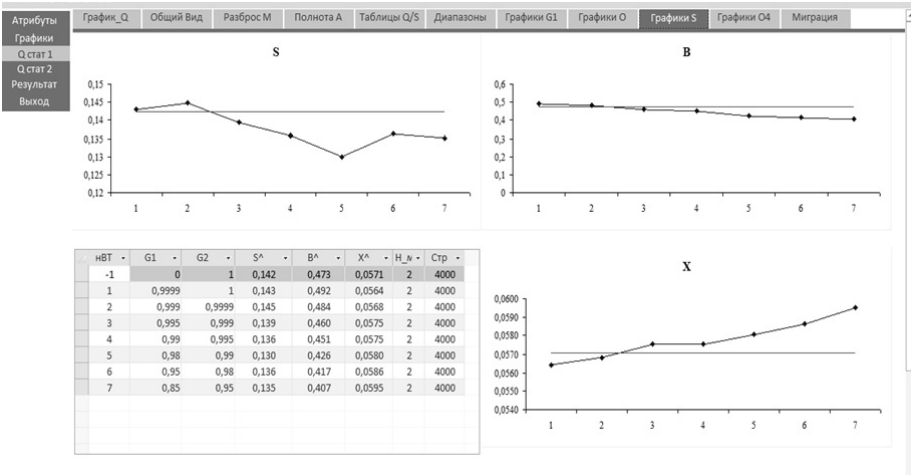

Рис. 5. Распределения средней частоты встречаемости пробела *S*, количества биграмм *B*, индекса совпадения *X*

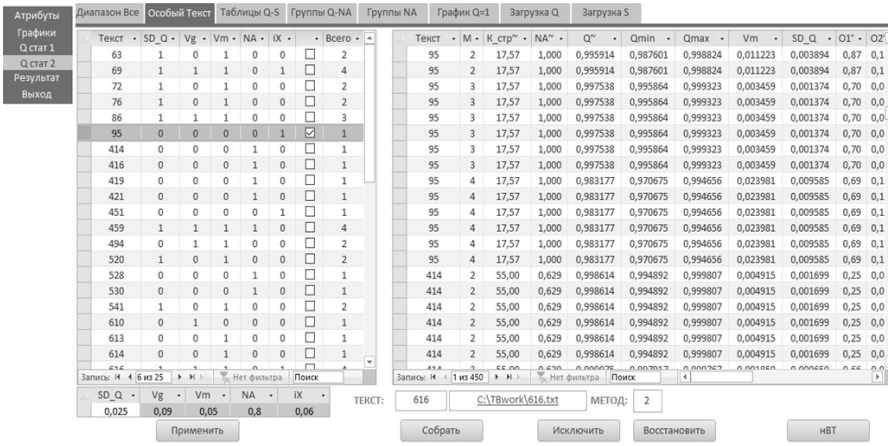

Рис. 6. Обработка выделенных текстов

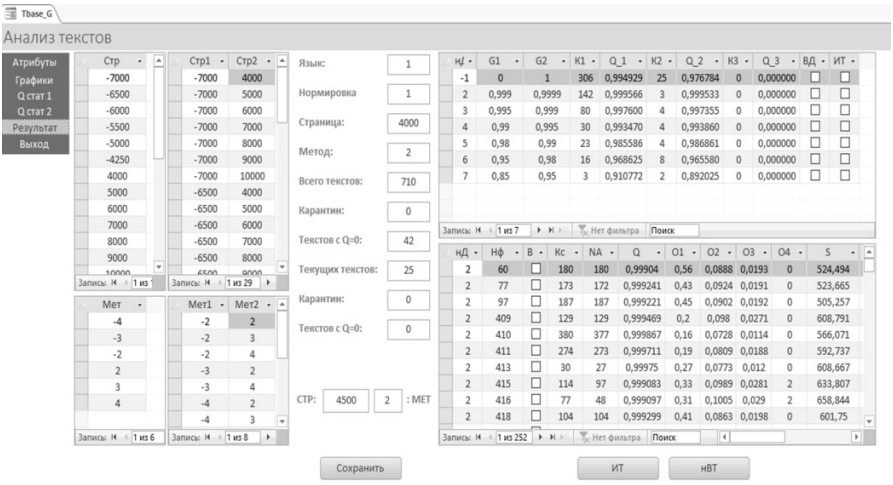

Рис. 7. Фиксация результатов

ЗАКЛЮЧЕНИЕ

В работе рассмотрены средства анализа текстов, основанные на числовой оценке, полученной с помощью применения методов частотного криптоанализа простой замены к открытому тексту. Эти средства позволяют проводить точную и единообразную оценку различных текстов и планировать вычислительный эксперимент с их использованием, что особенно важно в области исследования криптографических методов, методов криптоанализа и других формальных методов исследования текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Нормализация тестовой шкалы: Human Technologies. – URL: https://old.ht-lab.ru/cms/component/content/article/3-dictionary/1267-2009-10-09-15-52-01 (дата обращения: 08.07.2020).

2. ISBD: Международное стандартное библиографическое описание / Международная федерация библиотечных ассоциаций и учреждений и др.; пер. с англ. Н.В. Шпановой; науч. ред. пер.: Т.А. Бахтурина, Н.Н. Каспарова. – Консолидированное изд. – М., 2014. – 325 с.

3. УДК классификатор. – URL: https://www.triumph.ru/udk-klassifikator.html (дата обращения: 08.07.2020).

4. *Sebastiani F.* Machine learning in automated text categorization // ACM Computing Surveys. – 2002. – Vol. 34. – P. 1–47.

5. Тематическое моделирование. – URL: http://www.machinelearning.ru/wiki/index.php?title=Тематическое_моделирование (дата обращения: 08.07.2020).

6. Развитие криптографических методов и средств защиты информации / Л.К. Бабенко, Е.А. Ищукова, Е.А. Маро, И.Д. Сидоров, П.П. Кравченко // Известия ЮФУ. Технические науки. – 2012. – № 4. – С. 40–50.

7. *Бабенко Л.К., Ищукова Е.А.* Анализ симметричных криптосистем // Известия ЮФУ. Технические науки. – 2012. – № 12. – С. 136–147.

8. *Барановская А.О.* Взаимодействие интерфейса textLab и базы данных Tbase // Наука, образование и культура. – 2018. – № 4 (28). – С. 9–13.

9. *Коломец Н.В.* Структура системы TextLab для частотного анализа текста / науч. рук. Ю.А. Котов // Материалы 54 международной научной студенческой конференции (МНСК-2016). Информационные технологии = Proceedings of the 54 international students scientific conference (ISSC-2016). Information technologies, 16–20 апр. 2016 г. – Новосибирск: Изд-во НГУ, 2016. – С. 173.

10. *Коломец Н.В.* Графическая оболочка для учебного практикума в интегрированной среде анализа текстов TexTLab / науч. рук. Ю.А. Котов // Науч-

ное сообщество студентов XXI столетия. Технические науки. – Новосибирск, 2017. – № 4 (51). – С. 112–119.

11. *Котов Ю.А.* Методика и результаты сравнительного анализа четырех методов идентификации букв текстов // Информационные технологии и вычислительные системы. – 2019. – № 3. – С. 41–56.

12. *Котов Ю.А*. Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. – 2016. – Вып. 1. – С. 181–197.

13. *Шеннон К.* Теория связи в секретных системах // Шеннон К. Работы по теории информации и кибернетике. – М.: Иностранная литература, 1963. – С. 333–369.

14. *Makarskaya D.I.* Text analysis for solving cryptographical problems // Youth Con-tributions to the Breakthroughs into the Future: труды Всероссийской научно-практической конференции студентов бакалавриата, магистрантов и аспирантов / отв. ред. М.Н. Гордеева. – Новосибирск: Изд-во НГТУ, 2019. – С. 49–51.

15. *Котов Ю.А.* Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // Труды СПИИРАН. – 2017. – Вып. 1 (50). – С. 190–208.

Котов Юрий Алексеевич, кандидат физико-математических наук, доцент, доцент кафедры защиты информации Новосибирского государственного технического университета. Основные направления научных исследований: информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Имеет более 35 публикаций. E-mail: kotov@corp.nstu.ru

Макарская Дарья Игоревна, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Основное направление научных исследований – информационная безопасность. Имеет 2 публикации. E-mail: dashamakarskaya96@mail.ru

DOI: 10.17212/2307-6879-2020-1-2-99-112

Tools for text analysis based on the cryptanalysis of one-to-one substitution*

Yu.A. Kotov[1], D.I. Makarskaya[2]

[1] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, candidate of physical and mathematical sciences, docent, docent of the information security department. E-mail: kotov@corp.nstu.ru*
[2] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, master's student of the computer engineering department. E-mail: dashamakarskaya96@mail.ru*

The article considers text analysis tools based on a numerical estimate obtained using the cryptographic approach. It contains applying frequency cryptanalysis methods of one-to-one substitution to plaintext. In this case, these methods tend to methods for identifying letters of the text, and the task of cryptanalysis is to the task of identifying letters of text. For frequency cryptanalysis methods of one-to-one substitution, the identification error is determined as a statistic of a number of cryptanalysis errors for certain volumes of texts. The numerical score called the quality factor of a method can be transferred from methods to texts. Text analysis based on the quality factor of texts includes the selection of a cryptographic method with the help of which a quantitative assessment of the text is given, and the subsequent calculation of the quality factor of the text vector. In order to analyze texts based on cryptanalysis of a simple substitution the main stages of this analysis and necessary means are determined. These tools are implemented in the graphical user interface of MS Access and include 30 navigation tabs, on which 143 information and control elements are located. The article includes examples of implementation of some tools.

Keywords: analysis, texts, cryptanalysis, identification, quality factor, one-to-one substitution

REFERENCES

1. *Normalizatsiya testovoi shkaly: Human Technologies* [Scale normalization: Human Technologies]. Available at: https://old.ht-lab.ru/cms/component/content/article/3-dictionary/1267-2009-10-09-15-52-01 (accessed 08.07.2020).

2. *ISBD: International Standard Bibliographic Description*. Consolidated ed. München, K.G. Saur, 2011 (Russ. ed.: *ISBD: Mezhdunarodnoe standartnoe bibliograficheskoe opisanie*. Moscow, 2014. 325 p.).

3. *UDK klassifikator* [UDC classifier]. Available at: https://www.triumph.ru/udk-klassifikator.html (accessed 08.07.2020).

4. Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002, vol. 34, pp. 1–47.

* *Received 19 May 2020.*

5. Tematicheskoe modelirovanie [Topic modeling]. Available at: http://www.machinelearning.ru/wiki/index.php?title=Tematicheskoe_modelirovanie (accessed 08.07.2020).

6. Babenko L.K., Ishchukova E.A., Maro E.A., Sidorov I.D., Kravchenko P.P. Razvitie kriptograficheskikh metodov i sredstv zashchity informatsii [Development of cryptographic techniques and means of information security]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki = Izvestiya Southem Federal University. Engineering sciences*, 2012, no. 4, pp. 40–50.

7. Babenko L.K., Ishchukova E.A. Analiz simmetrichnykh kriptosistem [Analysis of simmetric cryptosystems]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki = Izvestiya Southem Federal University. Engineering sciences*, 2012, no. 12, pp. 136–147.

8. Baranovskaya A.O. Vzaimodejstvie interfejsa textLab i bazy dannyh Tbase [Interaction of the textLab interface and Tbase database]. *Nauka, obrazovanie i kul'tura = Science, Education and Culture*, 2018, no. 4 (28), pp. 9–13.

9. Kolomets N.V. [TexTLab system structure for text frequency analysis]. *Materialy 54 mezhdunarodnoi nauchnoi studencheskoi konferentsii (MNSK-2016). Informatsionnye tekhnologii* [Proceedings of the 54 international students scientific conference (ISSC-2016). Information technologies], 16–20 April 2016. Novosibirsk, 2016, p. 173. (In Russian).

10. Kolomets N.V. Graficheskaya obolochka dlya uchebnogo praktikuma v integrirovannoi srede analiza tekstov TexTLab [Graphical shell for a training workshop in an integrated text analysis environment TexTLab / scientific. hands]. *Nauchnoe soobshchestvo studentov XXI stoletiya. Tekhnicheskie nauki* [Scientific community of students of the XXI century. Technical science]. Novosibirsk, 2017, no. 4 (51), pp. 112–119.

11. Kotov Yu.A. Metodika i rezul'taty sravnitel'nogo analiza chetyrekh metodov identifikatsii bukv tekstov [Comparative analysis of four methods for identifying letters of texts]. *Informatsionnye tekhnologii i vychislitel'nye sistemy = Journal of Information technologies and computing systems*, 2019, no. 3, pp. 41–56.

12. Kotov Yu.A. Determinirovannaya identifikatsiya bukvennykh bigramm v russkoyazychnykh tekstakh [Determinate identification of Russian text letter bigrams]. *Trudy SPIIRAN = SPIIRAS Proceedings*, 2016, iss. 1, pp. 181–197.

13. Shennon K. Teoriya svyazi v sekretnykh sistemakh [Communication Theory of Secrecy Systems]. Shennon K. *Raboty po teorii informatsii i kibernetike* [Works on information theory and cybernetics]. Moscow, Inostrannaya literatura Publ., 1963, pp. 333–369. (In Russian).

14. Makarskaya D.I. Text analysis for solving cryptographical problems. *Youth Contributions to the Breakthroughs into the Future*. *Proceedings 2019 All-Russian*

Academic and Research Conference of Graduate and Postgraduate Students, Novosibirsk, NSTU Publ., 2019, pp. 49–51.

15. Kotov Yu.A. Approksimatsiya raspredelenii chastot bukvennykh bigramm teksta dlya identifikatsii bukv [Approximation of distributions of text characters bigrams frequencies for alphabetic characters identification]. *Trudy SPIIRAN = SPIIRAS Proceedings*, 2017, iss. 1 (50), pp. 190–208.

Для цитирования:

Котов Ю.А., Макарская Д.И. Средства анализа текстов на основе криптоанализа простой замены // Сборник научных трудов НГТУ. – 2020 – № 1–2 (97). – С. 99–112. – DOI: 10.17212/2307-6879-2020-1-2-99-112.

For citation:

Kotov Yu.A., Makarskaya D.I. Sredstva analiza tekstov na osnove kriptoanaliza prostoi zameny [Tools for text analysis based on the cryptanalysis of one-to-one substitution]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Transaction of scientific papers of the Novosibirsk state technical university*, 2020, no. 1–2 (97), pp. 99–112. DOI: 10.17212/2307-6879-2020-1-2-99-112.