УДК 004.738
Doi: 10.31772/2712-8970-2022-23-2-148-155

Для цитирования: Донцов Д. Ю., Исаев С. В. Применение методов тематического моделирования для идентификации групп интернет-ресурсов с целью снижения риска киберугроз // Сибирский аэрокосмический журнал. 2022. Т. 23, № 2. С. 148–155. Doi: 10.31772/2712-8970-2022-23-2-148-155.

For citation: Dontsov D. Y., Isaev S. V. [Application of topic modeling methods to identify groups of internet resources in order to reduce the risk of cyber threats]. *Siberian Aerospace Journal*. 2022, Vol. 23, No. 2, P. 148–155. Doi: 10.31772/2712-8970-2022-23-2-148-155.

Применение методов тематического моделирования для идентификации групп интернет-ресурсов с целью снижения риска киберугроз

Д. Ю. Донцов*, С. В. Исаев

Институт вычислительного моделирования СО РАН
Российская Федерация, 660036, Красноярск, Академгородок, 50, стр. 44
*E-mail: denis.dontsov96@gmail.com

Безопасность внутренней сети является важным аспектом успешного предприятия. Существуют различные средства для предотвращения киберугроз и анализа посещаемых интернет-ресурсов, но их быстродействие и возможность применения сильно зависит от объема входных данных. В статье рассматриваются существующие методы определения сетевых угроз с помощью анализа журналов прокси-сервера и предлагается метод кластеризации интернет-ресурсов, направленный на снижение объема входных данных путем исключения групп безопасных интернет-ресурсов или выбором только подозрительных интернет-ресурсов. Предложенный метод состоит из 3-х этапов: предобработка данных, анализ данных и интерпретация полученных результатов. Исходными данными для него являются записи журнала прокси-сервера. На первом этапе из исходных данных выбираются полезные для анализа данные, после чего непрерывный поток данных делится на небольшие сессии при помощи метода ядерной оценки плотности. На втором этапе выполняется мягкая кластеризация посещенных интернет-ресурсов путем применения метода тематического моделирования. Результатом второго этапа являются неразмеченные группы интернет-ресурсов. На третьем этапе, с помощью эксперта, происходит интерпретация полученных результатов путем анализа наиболее популярных интернет-ресурсов в каждой группе. Метод имеет множество настроек на каждом этапе, что позволяет сконфигурировать его под любой формат и специфику входных данных. Его область применения никак не ограничивается. Полученный метод может быть использован в качестве дополнительного шага предобработки с целью снижения количества входных данных.

Ключевые слова: тематическое моделирование, кибербезопасность, анализ данных.

Application of topic modeling methods to identify groups of internet resources in order to reduce the risk of cyber threats

D. Y. Dontsov*, S. V. Isaev

Institute of Computational Modeling SB RAS
50, b. 44, Academgorodok St., Krasnoyarsk, 660036, Russian Federation
*E-mail: denis.dontsov96@gmail.com

Internal network security is an important aspect of a successful enterprise. There are various means to prevent cyber threats and analyze visited Internet resources, but their speed and the possibility of application strongly depends on the volume of input data. This article discusses the existing methods for determining network threats by analyzing proxy server logs, and proposes a method for clustering Internet resources aimed at reducing the volume of input data by excluding groups of secure Internet resources or selecting only suspicious Internet resources. The proposed method consists of 3 stages: data preprocessing, data analysis and interpretation of the results obtained. The initial data for the method are the proxy server log entries. At the first stage, data useful for analysis is selected from the source data, after which the continuous data stream is divided into small sessions using the nuclear density estimation method. At the second stage, soft clustering of visited Internet resources is performed by applying the thematic modeling method. The result of the second stage are unmarked groups of Internet resources. At the third stage, with the help of an expert, the results are interpreted by analyzing the most popular Internet resources in each group. The method has many settings at each stage, which allows you to configure it for any format and specifics of the input data. The scope of the method is not limited in any way. The resulting method can be used as an additional preprocessing step in order to reduce the amount of input data.

Keywords: topic-modeling, cyber security, data analysis.

Введение

С каждым днем информационные технологии все глубже внедряются в жизни людей, в связи с чем вопросы обеспечения кибербезопасности становится все более важным.

Существует три класса источников киберугроз – человеческий, технологический и форс-мажорный [1]. Человек является причиной большинства киберугроз [2], в связи с чем разработка решений, позволяющих снизить число вторжений по вине человека, является перспективным направлением.

Для предотвращения посещения вредоносных ресурсов, на больших предприятиях используется технология фильтрации интернет-трафика [3]. Данное решение значительно снижает риск кибератак, но не дает 100 % защиты, поэтому необходимо использовать дополнительные средства защиты внутренней сети.

Безопасность внутренней сети включает в себя захват, сохранение и анализ данных использования сети. Результаты анализа позволяют выявлять изменения в шаблонах поведения пользователей, тем самым предоставляя возможность своевременно реагировать и предотвращать сетевые угрозы [4–6]. Процесс анализа данных, генерируемых пользователями внутренней сети, занимает некоторое время, и снижение затрачиваемого времени на анализ данных напрямую влияет на безопасность сети.

Пользователи сети ежедневно генерируют сотни тысяч запросов к различным интернет-ресурсам, в связи с чем, снижение объема анализируемых данных является одним из наиболее значимых подходов к снижению времени анализа.

Распределение посещаемых ресурсов на группы и выявление групп безопасных и потенциально опасных ресурсов может снизить объем анализируемых данных и дать значительный прирост к скорости обнаружения аномалий в поведении пользователей

В данной статье предложен подход разделения посещаемых интернет-ресурсов на группы со схожей тематикой при помощи анализа журналов доступа прокси-сервера. Основная цель предложенного метода в разделении ресурсов на группы с целью снижения объема анализируемых данных через исключение «безопасных» групп ресурсов.

Входные данные

Входными данными являются файлы журнала прокси-сервера, который является посредником между пользователем и интернет-ресурсами. Журнал (лог-файл) содержит информацию по всем запросам пользователей, совершенных в течение суток.

Каждая строка лог-файла содержит следующую информацию:

– время совершения запроса;

– адрес посещенного интернет-ресурса;
– уникальный идентификатор пользователя, совершившего запрос;
– тип запроса (get, post, put, delete и т. д.);
– тип запрашиваемого контента (image, html, css, js, …);
– объем переданных данных.
Для дальнейшего анализа нужны только некоторые из этих полей, такие как:
– время посещения – момент времени, в который пользователь отправил запрос;
– адрес посещаемого ресурса;
– идентификатор пользователя – уникальный идентификатор-пользователя, используемый для выделения ресурсов, посещенных одним пользователем.

Предложенный подход

В работе используется подход выделения и установления связей между посещаемыми интернет-ресурсами через анализ их совместной встречаемости в пределах некоторых сессий (рис. 1).

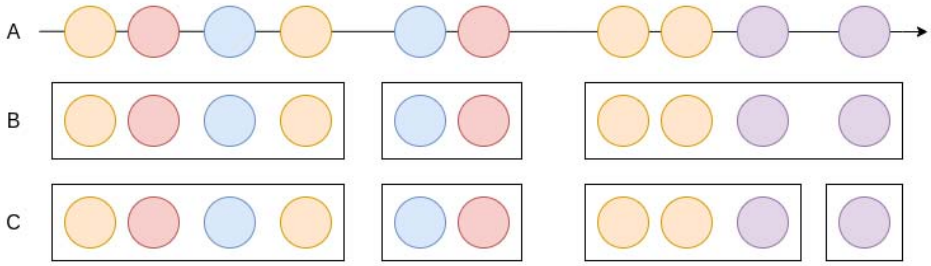

Рис. 1. Схема разбиения лог-файла на сессии:
A – исходный временной ряд; B – первый вариант разбиения; C – второй вариант разбиения

Fig. 1. The scheme of splitting the log file into sessions:
A – is the original time series; B – is the first variant of the partition; C – is the second variant of the partition

Под сессией подразумевается совокупность интернет-ресурсов, посещенных за некоторый промежуток времени. В самом простом случае, сессией можно считать одни сутки, однако для повышения качества работы метода необходимо рассмотреть другие варианты выделения сессий.

Для анализа совместной встречаемости ресурсов в пределах одной сессии используется вероятностное тематическое моделирование [7]. Тематическое моделирование выполняет мягкую кластеризацию «документов», опираясь на совместную встречаемость «термов» в этих документах. В качестве документов в данном случае используются ресурсы, посещенные в пределах одной сессии, а в качестве термов – сами ресурсы. Результатом работы тематического моделирования являются интернет-ресурсы, сгруппированные на определенное число не именованных групп (рис. 2).

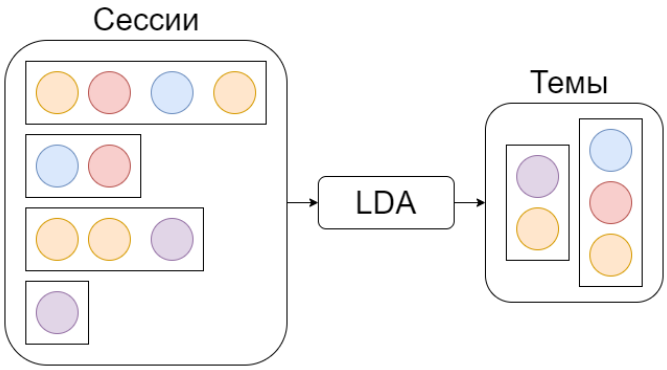

Рис. 2. Схема работы предложенного метода

Fig. 2. The scheme of the proposed method

Ручной анализ наиболее популярных интернет-ресурсов, попавших в каждую группу, позволит определить название каждой группы и выявить группы безопасных и опасных интернет-ресурсов.

Предобработка лог-файлов

Пользователи сети интернет ежедневно генерируют тысячи записей в лог-файлах (рис. 3), и большинство записей в этих файлах не несет полезной информации. При посещении одной интернет-страницы, браузер совершает в среднем 10–20 запросов, и каждый из этих запросов фиксируются в журнале прокси-сервера. Основная цель предобработки – снижение числа обрабатываемых данных, что позволит ускорить процесс анализа и повысить качество результатов [8].

Для дальнейшего анализа разумно исключить записи, удовлетворяющие одному из требований:

– запрашиваемый ресурс имеет тип css/js/image;

– запрос совершен анонимным пользователем.

В среднем такая фильтрация снижает объем данных примерно в 5 раз. Опционально для большего снижения числа обрабатываемых данных можно учитывать только get-запросы.

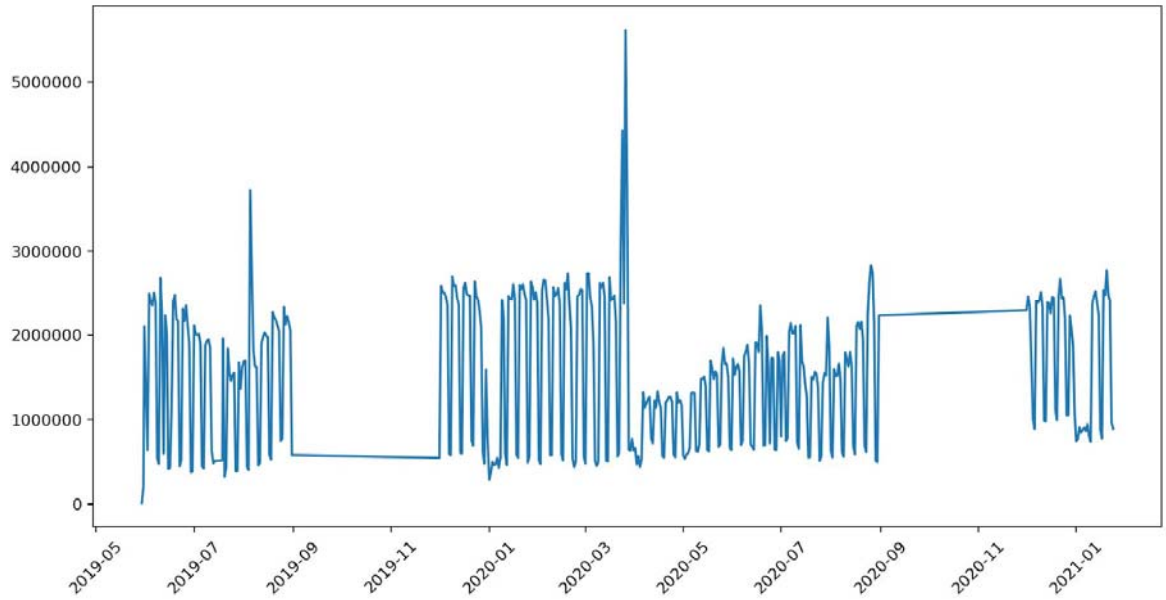

Рис. 3. Ежедневное число запросов, генерируемых 700 пользователями внутренней сети

Fig. 3. Daily number of requests generated by 700 users of the internal network

Вторым этапом предобработки является выделение доменов (или IP-адресов) посещенных ресурсов, чтобы учитывать посещение двух страниц одного сайта как посещение одного и того же ресурса дважды.

Разделение лог файлов на сессии

На данном этапе требуется разделить записи журнала прокси-сервера каждого пользователя на короткие сессии. Возможны различные варианты выделения сессий. Дальше рассмотрены некоторые из них.

Сессии фиксированной длины. Для получения сессий фиксированной длины достаточно задать некоторый временной интервал, например 1 день, и разбить все множество записей через выбранный интервал. Данный подход плох тем, что он объединяет сессии небольшого размера. Например, пользователь мог пользоваться интернетом дважды – утром и вечером, однако для данного подхода это будет считаться одной сессией.

Использование периода неактивности пользователя позволяет порождать сессии разной длины, разделенные некоторым промежутком времени, в который не было никакой активности (например, 1 ч). Этот метод имеет один существенный недостаток – он не будет выявлять сессии, если у пользователя есть фоновые процессы, постоянно генерирующие запросы (например, 1 раз в 10 мин).

Недостатки рассмотренных подходов можно устранить, используя метод KDE (Kernel Density Estimation) [9–11]. Данный метод позволяет оценивать плотность распределения одномерного набора данных и определять локальные точки экстремума. Использование таких точек для разделения непрерывного набора данных на отрезки позволит генерировать сессии различной длины, близкие к реальному поведению пользователя (рис. 4). Метод KDE имеет два настраиваемых параметра – ядро и ширину канала. Эти параметры значительно влияют на результат, и их нужно подбирать, вручную анализируя размеры получаемых сессий.

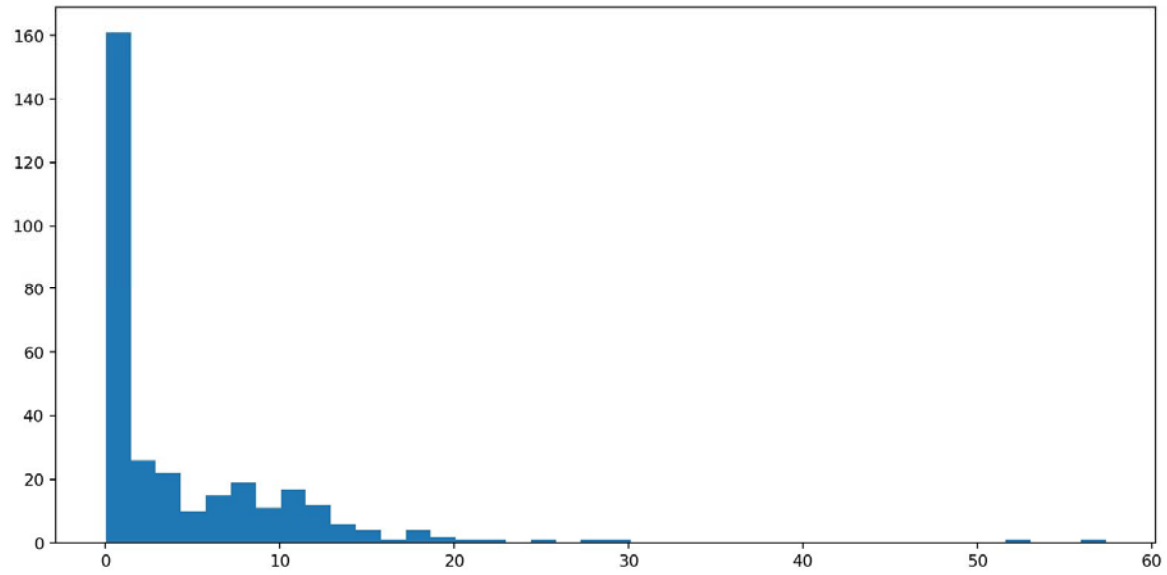

Рис. 4. Гистограмма распределения длин сессий за 1 день. По горизонтальной оси указано время в минутах, а по вертикальной – количество сессий. Средняя длина сессии – 4,5 мин

Fig. 4. Histogram of the distribution of session lengths for 1 day. The horizontal axis shows the time in minutes, and the vertical axis shows the number of sessions. The average session length is 4.5 minutes

Тематическое моделирование

Тематическое моделирование используется для строгой или мягкой кластеризации документов, состоящих из термов. Существует множество различных методов тематического моделирования [12–13], однако в данной статье используется метод LDA [14–15].

Для использования тематического моделирования необходимо определить документы и термы. Термом является домен интернет-ресурса, посещенного пользователем, а документом является множество доменов (термов), посещенных одним пользователем в пределах одной сессий.

Применение любой готовой реализации метода LDA для полученных документов позволяет мягко сгруппировать все домены интернет-ресурсов на фиксированное число групп. Количество групп задается пользователем и определяется опытным путем. В таблице представлен результат моделирования 5 групп. Чем выше ресурс расположен в группе, тем сильнее его принадлежность к этой группе.

Анализ наиболее популярных интернет-ресурсов в каждой группе позволяет определить тему каждой группы и решить, является ли группа «безопасной». В случае, если темы групп определить не удается, следует попробовать изменить число искомых тем.

Результат моделирования данных за февраль 2020 г. на 5 групп

1	2	3	4	5
newslab.ru	nowa.cc	update.eset.com	apps.webofknowledge.com	fitohobby.ru
4pda.ru	ugadalki.ru	law-college-sfu.ru	packages.linuxmint.com	ib.adnxs.com
sfkras.ru	scask.ru	kinoaction.ru	http.debian.net	allrefs.net
edu.sfu-kras.ru	forum.rcmir.com	kiwt.ru	urod.ru	ckp-rf.ru
worldcrisis.ru	2baksa.net	dostavka-krasnoyarsk.ru	fips.ru	teammodels.no
libgen.is	autoopt.ru	kinoaction.ru	mc.corel.com	profinance.ru

Для более точных результатов следует выбирать большее число групп. На рис. 5 представлена проекция 30 тем на две главные компоненты.

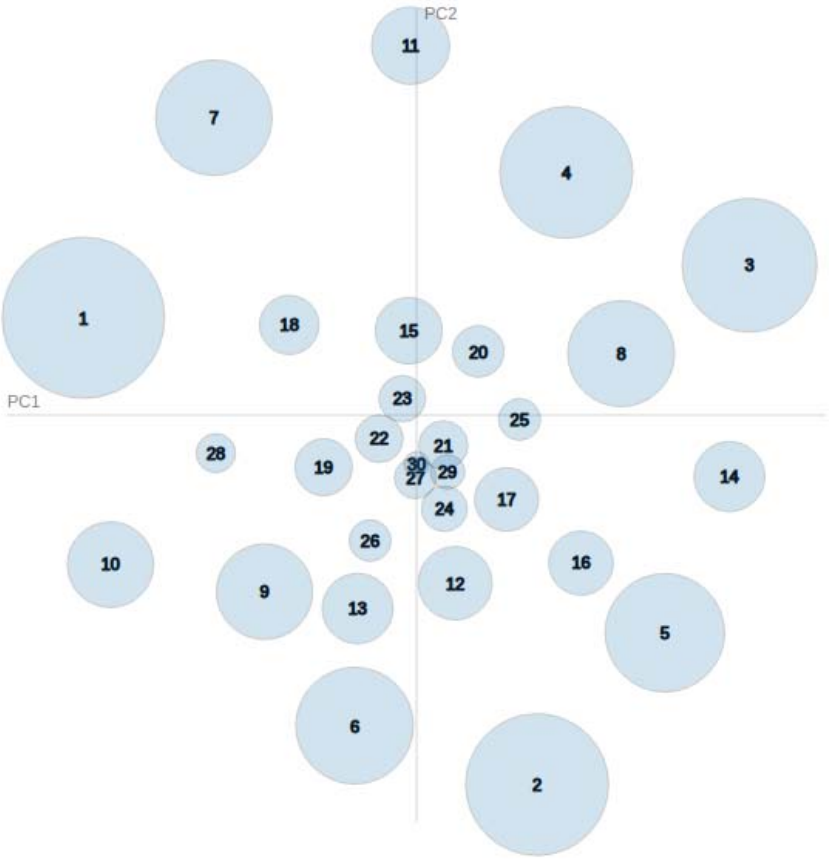

Рис. 5. Проекция 30 групп, полученных путем моделирования данных за февраль 2020 г., на две главные компоненты

Fig. 5. Projection of 30 groups obtained by modeling data for February 2020 into two main components

Заключение

Предложенный в статье метод имеет много настраиваемых параметров, позволяющих точно настроить его под разные источники данных, будь то небольшая внутренняя сеть или высоконагруженный узел масштабной сети.

Группировка интернет-ресурсов по схожей теме может быть использована в различных задачах, таких как:

– определение интересов пользователя;

– определение безопасных и опасных сайтов для снижения числа анализируемых данных;

– определение тематики интернет-ресурса.

В качестве дальнейших исследований планируется рассмотреть использование различных метаданных, таких как тип запрашиваемого контента и время совершения сессии. Выявление и отсеивание рекламных сервисов также может быть направлением дальнейшей разработки.

Библиографические ссылки

1. Mouna J., Latifa B., Latifa B. R., Anis A. Classification of security threats in information systems. // Procedia Computer Science. 2014. Vol. 32. P. 489–496.

2. Дерендяев Д. А., Гатчин Ю. А., Безруков В. А. Определение влияния человеческого фактора на основные характеристики угроз безопасности // Кибернетика и программирование. 2019, № 3. С. 38–42.

3. Gyorodi R., Cornelia G., Pecherle G., Radu L. Network Security Using Firewalls // Journal of Computer Science and Control Systems, 2008. Vol. 1.

4. Kao D. Y., Wang S. J., Huang F. Dataset Analysis of Proxy Logs Detecting to Curb Propagations in Network Attacks // Intelligence and Security Informatics. 2008. P. 245–250.

5. Marshall B., Chen, H. Using Importance Flooding to Identify Interesting Networks of Criminal Activity. // Lecture Notes in Computer Science. 2006. Vol. 3975. P. 14–25.

6. Mukkamala S., Sung A. Identifying significant features fornetwork forensic analysis using artificial techniques // InternationalJournal of Digital Evidence. 2003. Vol. 1, no 4. P. 67–74.

7. Blei D. M. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55, No. 4. P. 77–84.

8. Analysis of Web Proxy Logs / B. Fei, J. Eloff, M. Oliver, H. Venter // IFIP International Conference on Digital Forensics. Orlando, 2006. Vol. 222. P. 247–258.

9. Scott D. W. Multivariate Density Estimation. Theory. Practice and Visualization: Second edition. New York, 2015.

10. Using kernel density estimation to understand the influence of neighbourhood destinations on BMI / T. L. King, R. J. Bentley, L. E. Thornton et al. // BMJ Open, 2016, Vol. 6.

11. Kalinic M., Krisp J. Kernel Density Estimation (KDE) vs. Hot-Spot Analysis – Detecting Criminal Hot Spots in the City of San Francisco // Lund, Sweden, 2018.

12. Воронцов К. В. Вероятностное математическое моделирование: теория, модели, алгоритмы и проект BigFRTM. Москва : МАИ, 2021. 112 с.

13. Albalawi R., Yeap T., Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. // Frontiers in Artificial Intelligence. 2020. Vol. 3.

14. Jelodar H., Wang Y., Yuan, Ch., Xia, F. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. 2017.

15. Tharwat A., Gaber T., Ibrahim A., Hassanien A. E. Linear discriminant analysis: A detailed tutorial // Ai Communications. 2017. Vol. 30. P. 169–190.

References

1. Mouna J., Latifa B., Latifa B. R., Anis A. Classification of security threats in information systems. // Procedia Computer Science. 2014. Vol. 32. P. 489–496.

2. Derendyaev D. A., Gatchin Yu. A., Bezrukov V. A. [Determining the influence of the human factor on the main characteristics of security threats]. *Cybernetics and programming*. 2019, No. 3, P. 38–42 (In Russ.).

3. Gyorodi R., Cornelia G., Pecherle G., Radu L. Network Security Using Firewalls. *Journal of Computer Science and Control Systems*. 2008, Vol. 1.

4. Kao D. Y., Wang S. J., Huang F. Dataset Analysis of Proxy Logs Detecting to Curb Propagations in Network Attacks. *Intelligence and Security Informatics*. 2008, P. 245–250.

5. Marshall B., Chen, H. Using Importance Flooding to Identify Interesting Networks of Criminal Activity. *Lecture Notes in Computer Science*. 2006, Vol. 3975, P. 14–25.

6. Mukkamala S., Sung A. Identifying significant features fornetwork forensic analysis using artificial techniques. *InternationalJournal of Digital Evidence*. 2003, Vol. 1, No 4.

7. Blei D. M. Probabilistic topic models. *Communications of the ACM*. 2012, Vol. 55, No. 4, P. 77–84.

8. Fei B., Eloff J., Oliver M., Venter H. Analysis of Web Proxy Logs. *IFIP International Conference on Digital Forensics*. Orlando, 2006, Vol. 222, P. 247–258.

9. Scott D. W. Multivariate Density Estimation. Theory. Practice and Visualization: Second edition. New York, 2015.

10. King T. L., Bentley R. J., Thornton L. E. et al. Using kernel density estimation to understand the influence of neighbourhood destinations on BMI. *BMJ Open*. 2016, Vol. 6.

11. Kalinic M., Krisp J. Kernel Density Estimation (KDE) vs. Hot-Spot Analysis – Detecting Criminal Hot Spots in the City of San Francisco. *Lund, Sweden*, 2018.

12. Vorontsov K. V. *Obzor veroyatnostnykh tematicheskikh modelei* [Overview of probabilistic thematic models]. Moscow, 2021. 112 p.

13. Albalawi R., Yeap T., Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*. 2020, Vol. 3.

14. Jelodar H., Wang Y., Yuan, Ch., Xia, F. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. 2017.

15. Tharwat A., Gaber T., Ibrahim A., Hassanien A. E. Linear discriminant analysis: A detailed tutorial. *Ai Communications*. 2017, Vol. 30, P. 169–190.

Донцов Денис Юрьевич – аспирант, Институт вычислительного моделирования СО РАН. E-mail: denis.dontsov96@gmail.com.

Исаев Сергей Владиславович – кандидат технических наук, доцент, заведующий отделом информационно-телекоммуникационных технологий; Институт вычислительного моделирования СО РАН. E-mail: si@icm.krasn.ru.

Dontsov Denis Yurievich – postgraduate, Institute of Computational Modeling SB RAS. E-mail: denis.dontsov96@gmail.com.

Isaev Sergey Vladislavovich – Cand. Sc., docent, head of the department of information and telecommunication technologies; Institute of Computational Modeling SB RAS. E-mail: si@icm.krasn.ru.