ИЗУЧЕНИЕ ПОВЕДЕНИЯ СРЕДСТВ АВТОМАТИЗИРОВАННОГО СБОРА ИНФОРМАЦИИ С ВЕБ-РЕСУРСОВ

Менщиков А.А.[1], Комарова А.В.[2], Гатчин Ю.А.[3]

В данной статье производится анализ поведения веб–роботов на различных сайтах. Приводится классификация средств автоматизированного сбора информации, изучаются методы и подходы к обнаружению таких средств. Рассматриваются основные проблемы, стоящие перед разработчиками веб–роботов, которые могут быть положены в основу способов противодействия несанкционированному сбору информации. Отдельное внимание уделяется изучению типовых шаблонов поведения веб–роботов. Выделяются несколько базовых типов характеристик поведения, основанных на структуре, времени, поведении, типе контента и ошибочных запросах. В статье приводится сравнение структурных параметров, характерных для средств сбора информации и для обычных пользователей. Наблюдалось значительное различие значений данных характеристик. Производилось изучение поведения известных веб–роботов на наборе веб–логов крупного веб–ресурса. Было сделано заключение о принципиальной возможности различения пользовательского и автоматизированного поведения на веб–ресурсе. Теоретическая и практическая значимости полученных результатов заключаются в получении новых вычислительных результатов, которые послужат базой для разработки комплексной системы обнаружения и противодействия автоматизированному сбору информации с веб–ресурсов.

***Ключевые слова**: веб–роботы, парсинг, сбор информации, обнаружение веб–роботов, информационная безопасность, защита информации.*

DOI: 10.21681/2311-3456-2017-3-49-54

Введение

На сегодняшний день все более остро встает проблема защиты информации, содержащейся на веб-ресурсе. Услуги и критические данные все больше переносятся в интернет пространство, где подвергаются постоянным угрозам несанкционированного и автоматизированного сбора и обработки специальными средствами, называемыми веб-роботами [1, 2]. В России на сегодняшний день развивается рынок электронной коммерции, обороты которого, согласно различным исследованиям, растут до 15% в год и составляют около 850 млрд. рублей на 2016 год [3]. В связи с данными факторами, встает задача обеспечения целостности, конфиденциальности и доступности данных, расположенных на веб-ресурсах [2].

Для того чтобы собирать информацию с ресурсов в интернете существуют специализированные средства. К таким средствам относятся программы, называемые веб-роботами или парсерами (краулерами). Их можно условно разделить на две категории: роботы, используемые для законных целей (анализ контента сайта, индексирование для улучшения работы поисковых систем или создание «зеркал» веб-сайтов) и используемые злоумышленниками [4]. Веб-роботы могут не только собирать и обрабатывать информацию, но и выполнять активные действия на веб-ресурсе, такие как покупка товаров и услуг, написание рекламных текстов, рассылка спама и эксплуатация уязвимостей. Кроме того, работа веб-роботов приводит к увеличению нагрузки на сервер и уменьшению пропускной способности, а также проблемам доступа к ресурсу у обычных пользователей [5].

Для управления поведением легитимных веб-роботов существуют специальные механизмы высказывания пожелания администраторов веб-ресурсов и установления правил поведения краулеров в файле robots.txt. Однако, данные правила являются лишь рекомендациями и не ограничивают работу неэтичных веб-роботов [6].

Объемы веб-парсинга растут год от года (рис.1) [2, 3], по это причине актуальной задачей становится поиск методов различения обычных поль-

1 Менщиков Александр Алексеевич, аспирант, Университет ИТМО, Санкт-Петербург, Russia. E-mail: menshikov@corp.ifmo.ru

2 Комарова Антонина Владиславовна, аспирант, Университет ИТМО, Санкт-Петербург, Russia. E-mail: piter-ton@mail.ru

3 Гатчин Юрий Арменакович, доктор технических наук, профессор, Университет ИТМО, Санкт-Петербург, Russia. E-mail: gatchin@mail.ifmo.ru

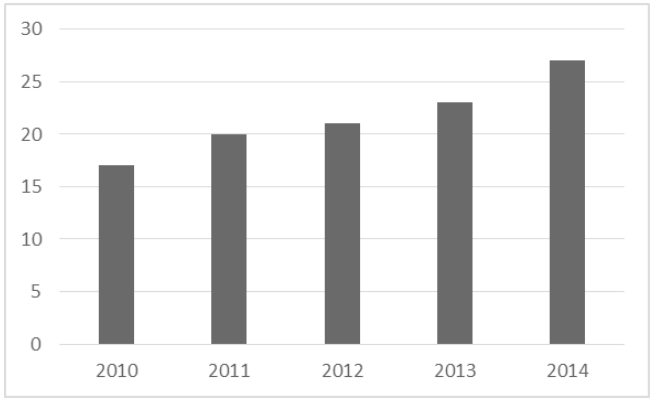

Рис. 1. Доля роботизированного трафика на веб ресурсе (в %)

зователей и вредоносных с целью дальнейшей блокировки неэтичных посетителей [4, 7].

Виды веб-роботов

Важной особенностью веб-роботов является следование целям и задачам получения актуальной информации, что включает в себя стремление к минимизации стоимости и времени сбора информации за счет исключения некорректного поведения и лишних запросов. Данное поведение характерно как для легитимных, так и для неэтичных веб-роботов, что позволяет прослеживать взаимосвязь их шаблонов обработки информации на веб-ресурсе, а также отличить их от обычных пользователей. Веб-роботов принято разделять на три основные категории (рис.2) [7, 8]:

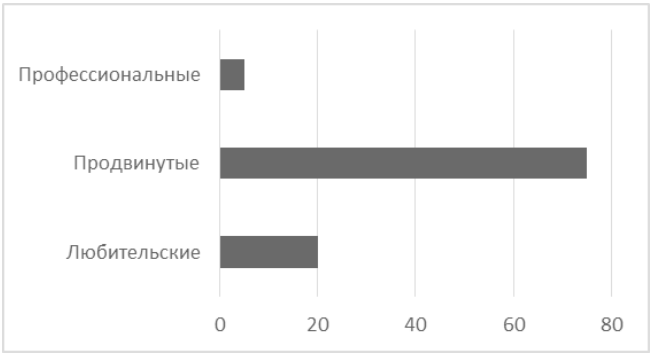

***Рис. 2.** Распространенность видов веб-роботов (в %)*

1. Любительские веб-роботы, использующие прямой перебор страниц и выполняющие только простейшие запросы к веб-серверу;
2. Продвинутые веб-роботы, изменяющих свое поведение и повторяющих шаблоны поведения обычных пользователей;
3. Профессиональные веб-роботы, использующие сложные алгоритмы поведения, и настраиваемые вручную под каждый веб-ресурс.

Обнаружение веб-роботов

Методы обнаружения веб-роботов можно классифицировать по принципу работы, по стратегии запуска и по используемым приемам. По первому критерию выделяют четыре основные категории (рис.3), по второму – подразделить на активные, которые действуют сразу во время обработки запроса, и на отложенные, выполняемые ретроспективно. Используемые приемы включают в себя обычную фильтрацию, методы машинного обучения: деревья принятия решений, классификаторы, обучение с использованием нейронных сетей и т.д.

Для улучшения производительности и точности результатов методы можно комбинировать между собой, например, используя результаты отложенного анализа для уточнения фильтров активного обнаружения и наоборот.

Проблема сбора информации

При разработке систем автоматизированного сбора информации с веб-ресурсов необходимо учитывать различные ограничительные факторы и основные проблемы, стоящие перед разработчиками таких программ. К таковым относятся:

1. Необходимость ручной настройки и отладки системы сбора информации для сайтов, имеющих сложную структуру;
2. Системы сбора информации должны уметь обрабатывать большие объемы данных за короткое время;
3. Структура отображения информации на сайте часто изменяется, за этим необходимо следить в автоматическом режиме.

***Рис. 3**. Классификация методов обнаружения веб-роботов*

При разработке системы противодействия автоматизированному сбору информации с веб-ресурсов важно понимать, какие сложности, стоящие перед разработчиками парсеров, и использовать их для защиты веб-ресурсов [9].

Характеристики поведения веб-роботов

Поведение известных веб-роботов похоже на поведение веб-роботов, используемых злоумышленниками. Основное различие состоит в целях сбора информации, типах загружаемого контента и соблюдении правил и пожеланий администраторов ресурса, описанных в файле robots.txt. Такие параметры как адреса источников запросов и HTTP заголовок User-agent позволяют идентифицировать легитимных веб-роботов и отличать их от обычных пользователей [10-12]. Мы изучили их характеристики, для того, чтобы обнаруживать неизвестных веб-роботов, скрывающих свое присутствие, используя тот факт, что паттерны роботизированного поведения характерны как для легитимных, так и для неизвестных веб-роботов в соответствии с проблемами сбора информации, описанными выше [13].

Мы выделяем следующие категории таких характеристик [14, 15]:

1. Временные – параметры, основанные на интервалах между запросами в рамках одной или нескольких сессий;
2. Структурные – параметры, зависящие от структуры HTTP пакета, корректности определенных полей и протоколов;
3. Основанные на типе контента – параметры, включающие тип и содержание загружаемого контента;
4. Основанные на ошибках – параметры, учитывающие количество и доли ошибок в запросах;
5. Поведенческие – параметры, изучающие поведение веб-робота в динамике, изменение путей его перемещения и характера действий.

Изучение поведения веб-роботов

В данном исследовании мы использовали набор логов запросов за два дня к крупному веб-ресурсу. Мы изучали выборку, содержащую 831 тысячу запросов, среди которых 9751 запроса относились к известным веб-роботам. Мы разделили их на 413 независимых сессий. Сессии, относящиеся к известным веб-роботам, определялись на основе совокупности следующих признаков: факт обращения к robots.txt, IP адрес, подсеть, User-agent. Период жизни сессии при разбиении был установлен длительностью в 30 минут. Алгоритм идентификации сессий можно представить в виде следующего приблизительного псевдокода:

```
for request in Requests:
    for session in ActiveSessions:
        if (request.time – session.lastTime > delta):
            session.close()
        else:
            if (session.containsIP(request.ip) and \\
        session.containsUserAgent(request.userAgent)):
                session.add(request)
            else:
                newSession = new Session()
                newSession.add(request)
```

Для изучения поведения веб-роботов были выбраны структурные параметры поведения, которые классифицировались характеристиками, представленными в (табл.1).

Таблица 1.
Основные характеристики обнаружения веб-роботов на базе содержимого HTTP пакета

№	Название	Описание
1	totalPages	Общее количество запросов.
2	nonStaticRequests	Число запросов к HTML документу.
3	staticRequests	Число запросов к статическим файлам и мультимедиа контенту .css, .js, .jpg, .png, .gif, .pdf.
4	robotsTXTRequest	Обращение к файлу robots.txt.
5	errorCodes3xx	Число ошибок с кодом 3XX в сессии.
6	errorCodes4xx	Число ошибок с кодом 4XX в сессии.
7	HEADRequests	Число HTTP HEAD запросов.
8	imagesCount	Число запросов на загрузку файлов .png, .jpg, .gif.
9	scriptsCount	Число запросов на загрузку файлов .css, .js.
10	unassignedReferer	Число запросов с пустым или равным «-» реферрером.

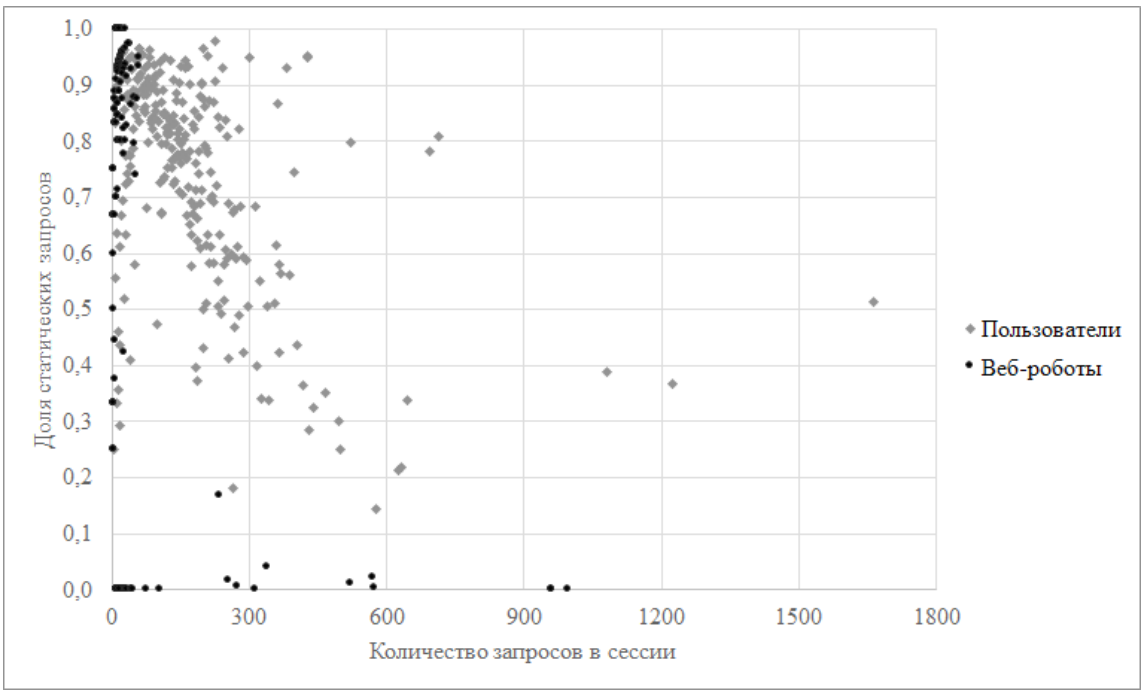

***Рис. 4**. Зависимость доли статических запросов от длины сессии*

Логи были предварительно очищены от пользовательских сессии и неизвестных веб-роботов. Запросы, которые остались в результаты были разделены на сессии, для которых были рассчитаны характеристики поведения. Затем из сессий были исключены нерелевантные и малоинформативные вхождения, в результате было произведено сравнение 56 сессий.

Характеристики веб-роботов были сравнены с пользовательскими. Данное сравнение позволило сделать заключение о значительном различии. Наиболее значимыми в контексте структурных характеристик оказались соотношения типов загружаемых файлов. На (рис.4) представлена зависимость доли запросов к статическим файлам от длины сессии для известных краулеров и обычных пользователей. Стоит отметить, что среди обычных пользователей были обнаружены веб-роботы, скрывающие свое присутствие.

Можно сделать вывод о том, что структурные методы обнаружения веб-роботов позволяют обнаружить определенные виды краулеров, но не покрывают всего многообразия средств автоматизированного сбора информации в отличии от временных и поведенческих методов. Данный вопрос требует дополнительных исследований.

Была разработана система, анализирующая поведение веб-роботов в автоматизированном режиме с целью более плотного изучения их характеристик и уточнения моделей поиска вредоносных посещений, а также для обеспечения блокировки явных веб-роботов в режиме онлайн. Примерная архитектура данной системы (рис.5) состоит из пяти модулей:

***Рис. 5**. Схематичное представление системы изучения поведения веб-роботов*

1. Обработчик запросов – компонент, получающий параметры поведения от веб-сервера;

2. Анализатор сессий – компонент, выделяющий сессии пользователей и веб-роботов и помещающий их в очередь;

3. Компонент, осуществляющий расчёт характеристик поведения для каждой из сессий;

4. Компонент принятия решений с использованием деревьев решений и пороговых значений показателей;

5. Анализатор результатов – компонент, представляющий характеристики для оператора системы с возможностью редактирования.

С использованием данной системы были рассчитаны и уточнены характеристики поведения веб-роботов и легитимных пользователей. Для каждой характеристики из рассматриваемого набора были составлены пороговые значения, в соответствии с которыми осуществлялась классификация. Набор данных был разделен на обучающий, на основе которого формировались пороговые значения, и тестовый, который был размечен вручную. Результаты классификации показали точность обнаружения на уровне 0,83 при полноте 0,92. Данные значения могут быть уточнены при исследовании большего набора данных при учете всех категорий характеристик обнаружения, также дополнительный интерес представляют результаты классификации с использованием методов машинного обучения, что будет являться предметом дальнейших исследований.

Выводы

Для решения проблемы обнаружения веб-роботов требуется целый комплекс инструментов. Во-первых, необходимы работающие методы, позволяющие детектировать веб-роботов на основе определенных параметров запросов и информации об их активности. Во-вторых, требуется разработать систему, позволяющую данные методы применять, собрать всю необходимую информацию, осуществлять ее препроцессинг, обработку и принятие решения. В-третьих, необходим фреймворк для настройки системы обнаружения и мониторинга ее работы.

Значимость результатов данного исследования заключается в разработке новых методических подходов и инструментов, которые могут быть использованы для защиты веб-ресурсов от автоматизированного сбора информации. Мы изучили набор логов веб-сервера и выявили роботизированные источники путем сравнения характеристик поведения посетителей. Полученные результаты позволяют автоматически выявлять активность веб-роботов на сайте и блокировать их деятельность. Данное исследование послужит заделом для построения комплексного подхода к обеспечению безопасности веб-ресурсов и формирования репрезентативных наборов данных, которые будут использованы для машинного обучения применительно к задаче обнаружения и противодействия автоматизированному сбору информации с веб-ресурсов.

Рецензент: *Безруков Вячеслав Алексеевич, кандидат технических наук, доцент кафедры проектирования и безопасности компьютерных систем, bezrukov@mail.ifmo.ru*

Литература

1. Отчет East-West Digital News [Электронный ресурс] – Режим доступа: http://www.ewdn.com/files/ecom-rus-download.pdf/, свободный (дата обращения: 27.10.2016).
2. Отчет компании scrapesentry [Электронный ресурс] – Режим доступа: https://www.scrapesentry.com/scrapesentry-scraping-threat-report-2015/, свободный (дата обращения: 27. 10.2016).
3. Отчет ассоциации компаний интернет-торговли [Электронный ресурс] – Режим доступа: http://www.akit.ru/wp-content/uploads/2016/05/E-commerce_1Q2016-FINAL.pdf, свободный (дата обращения: 01.11.2016).
4. Менщиков А.А., Гатчин Ю.А. Методы обнаружения автоматизированного сбора информации с веб-ресурсов // Кибернетика и программирование. – 2015. – № 5. – С. 136-157
5. Junsup Lee, Sungdeok Cha, Dongkun Lee, Hyungkyu Lee, Classification of web robots: An empirical study based on over one billion requests // Computers & Security. – 2009. – V. 28. – № 8. – P. 795-802.
6. Robots Exclusion Protocol Guide [Электронный ресурс]. – Режим доступа: http://www.bruceclay.com/seo/robots-exclusion-guide.pdf, свободный (дата обращения: 01.11.2016).
7. Менщиков А.А., Гатчин Ю.А. Построение системы обнаружения автоматизированного сбора информации с веб-ресурсов // Инженерные кадры - будущее инновационной экономики России: Материалы Всероссийской студенческой конференции: в 8 ч. – 2015. – Т. 4. – С. 58-61
8. D. Derek, S. Gokhale A Classification Framework for Web Robots // Journal of American Society of Information Science and Technology. – 2012. – V. 63. – P. 2549–2554.
9. G. Jacob, E. Kirda, C. Kruegel, G. Vigna PUB CRAWL: Protecting Users and Businesses from CRAWLers // Proceeding Security'12 Proceedings of the 21st USENIX conference on Security symposium. – 2012. – P. 25–36.
10. S. Kwon, YG. Kim, S. Cha Web robot detection based on pattern-matching technique // Journal of Information Science. – 2012. – V. 38(2). – P. 118–126.
11. Tan, Pang-Ning, and Vipin Kumar Discovery of web robot sessions based on their navigational patterns // Intelligent Technologies for Information Analysis. Springer Berlin Heidelberg. – 2004. – P. 193-222.

12. Stassopoulou, Athena, and Marios D. Dikaiakos Web robot detection: A probabilistic reasoning approach // Computer Networks. – V. 53. – № 3. – 2009. – P. 265-278.
13. Lu, Wei-Zhou, and Shun-zheng Yu Web robot detection based on hidden Markov model // 2006 International Conference on Communications, Circuits and Systems. – 2006.
14. T. H. Sardar, Z. Ansari Detection and Confirmation of Web Robot Requests for Cleaning the Voluminous Web Log Data // Proceeding International Conference on the IMpact of E-Technology on US. – 2014. – V. 28. – P. 795–802.
15. D. S. Sisodia, S. Verma, O. P. Vyas Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors // Journal of Data Analysis and Information Processing. – 2015. – V. 3. – P. 1–10.

A STUDY OF WEB-CRAWLERS BEHAVIOUR

A. Menshchikov[4], A. Komarova[5], U. Gatchin[6]

In this paper, we present a study of web–crawlers behaviour on different websites. We provide a classification of web–crawling tools and analyze web–crawling detection methods. Based on a large web–server logs set, we study behaviour of well–known scrapers. We conclude that humans can be distinguished from web–robots based on several features. Our results and observations can be used as a basis of comprehensive intrusion detection and prevention system development.

Keywords*: web robots, scraping, information gathering, web robots detection, information security, data protection.*

References

1. Report East-West Digital News [Electronic resource] – Mode of access: http://www.ewdn.com/files/ecom-rus-download.pdf/, free (date accessed: 27.10.2016).
2. Report of the company scrapesentry [Electronic resource] – Mode of access: https://www.scrapesentry.com/scrapesentry-scraping-threat-report-2015/, free (date accessed: 27. 10.2016).
3. Report of the Association of companies of Internet trade [Electronic resource] – Mode of access: http://www.akit.ru/wp-content/uploads/2016/05/E-commerce_1Q2016-FINAL.pdf free (reference date: 01.11.2016).
4. Menshchikov A.A., Gatchin YU.A. Metody obnaruzheniya avtomatizirovannogo sbora informacii s veb-resursov // Kibernetika i programmirovanie [Cybernetics and programming]. – 2015. – № 5. – S. 136-157
5. Junsup Lee, Sungdeok Cha, Dongkun Lee, Hyungkyu Lee, Classification of web robots: An empirical study based on over one billion requests // Computers & Security. – 2009. – V. 28. – № 8. – P. 795-802.
6. Robots Exclusion Protocol Guide [Electronic resource]. – Mode of access: http://www.bruceclay.com/seo/robots-exclusion-guide.pdf free (reference date: 01.11.2016).
7. Menshchikov A.A., Gatchin YU.A. Postroenie sistemy obnaruzheniya avtomatizirovannogo sbora informacii s veb-resursov // Inzhenernye kadry - budushchee innovacionnoj ehkonomiki Rossii: Materialy Vserossijskoj studencheskoj konferencii: v 8 ch. – 2015. – T. 4. – S. 58-61
8. D. Derek, S. Gokhale A Classification Framework for Web Robots // Journal of American Society of Information Science and Technology. – 2012. – V. 63. – P. 2549–2554.
9. G. Jacob, E. Kirda, C. Kruegel, G. Vigna PUB CRAWL: Protecting Users and Businesses from CRAWLers // Proceeding Security'12 Proceedings of the 21st USENIX conference on Security symposium. – 2012. – P. 25–36.
10. S. Kwon, YG. Kim, S. Cha Web robot detection based on pattern-matching technique // Journal of Information Science. – 2012. – V. 38(2). – P. 118–126.
11. Tan, Pang-Ning, and Vipin Kumar Discovery of web robot sessions based on their navigational patterns // Intelligent Technologies for Information Analysis. Springer Berlin Heidelberg. – 2004. – P. 193-222.
12. Stassopoulou, Athena, and Marios D. Dikaiakos Web robot detection: A probabilistic reasoning approach // Computer Networks. – V. 53. – № 3. – 2009. – P. 265-278.
13. Lu, Wei-Zhou, and Shun-zheng Yu Web robot detection based on hidden Markov model // 2006 International Conference on Communications, Circuits and Systems. – 2006.
14. T. H. Sardar, Z. Ansari Detection and Confirmation of Web Robot Requests for Cleaning the Voluminous Web Log Data // Proceeding International Conference on the IMpact of E-Technology on US. – 2014. – V. 28. – P. 795–802.
15. D. S. Sisodia, S. Verma, O. P. Vyas Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors // Journal of Data Analysis and Information Processing. – 2015. – V. 3. – P. 1–10.

4 Aleksandr Menshchikov, postgraduate student, St. Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia. E-mail: menshikov@corp.ifmo.ru

5 Antonina Komarova, postgraduate student, St. Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia. E-mail: piter-ton@mail.ru

6 Yurij Gatchin, Dr.Sc., Professor, St. Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia. E-mail: gatchin@mail.ifmo.ru