УДК 004.912: 004.822
DOI: 10.17586/0021-3454-2023-66-12-1002-1010

ФЕНОМЕНОЛОГИЧЕСКОЕ ОПИСАНИЕ ПРОЦЕССОВ СБОРА И ОБРАБОТКИ ИНТЕРНЕТ-ДОКУМЕНТОВ

С. В. Кулешов, А. А. Зайцева*

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, Россия
**cher@iias.spb.su*

Аннотация. Проанализировано состояние сети Интернет как хранилища информационных ресурсов с точки зрения бота — программы, занимающейся сбором данных в целях мониторинга ресурсов, наполнения поисковой системы или других коммерческих или исследовательских целях. Предложен подход к описанию исследуемой проблемы через совокупность феноменов, возникающих при сборе документов в Интернете. Описанные феномены необходимо учитывать при построении систем мониторинга либо поисковых систем. Приведен ряд особенностей, возникающих при веб-скрейпинге, харвестинге и в других случаях использования ботов для сбора данных в сети Интернет. Описаны проблемы использования поддоменов, рекурсивных поддоменов, технологий динамически загружаемого контента, поисковой оптимизации текстового контента и других. Показано, что задача сбора данных с интернет-ресурсов является не только технологической, но и в большей степени наукоемкой, а поскольку исследования находятся в активной фазе, для них не существует „коробочного“ решения. Статья будет полезна исследователям в области развития Интернета, разработчикам поисковых систем, специалистам по дата-ретривингу и интернет-технологиям, а также специалистам в области создания и поддержки интернет-ресурсов и в области интернет-маркетинга.

Ключевые слова: *интернет-документы, технологии сбора данных, дата-ретривинг, поисковые системы, интернет-ресурсы*

Благодарности: работа поддержана Государственным заданием на 2023 г. № FFZF-2022-0005.

Ссылка для цитирования: *Кулешов С. В., Зайцева А. А.* Феноменологическое описание процессов сбора и обработки интернет-документов // Изв. вузов. Приборостроение. 2023. Т. 66, № 12. С. 1002—1010. DOI: 10.17586/0021-3454-2023-66-12-1002-1010.

PHENOMENOLOGICAL DESCRIPTION OF INTERNET DOCUMENTS COLLECTING AND PROCESSING

S. V. Kuleshov, A. A. Zaytseva*

St. Petersburg Federal Research Center of the RAS, St. Petersburg, Russia
**cher@iias.spb.su*

Abstract. The state of the Internet as a repository of information resources is analyzed from the point of view of a bot - a program that collects data for the purpose of monitoring resources, filling a search engine, or other commercial or research purposes. An approach is proposed to describe the problem under study through a set of phenomena that arise when collecting documents on the Internet. The described phenomena must be taken into account when developing monitoring systems or search engines. A number of features that arise during web scraping, harvesting and other cases of using bots to collect data on the Internet are given. The problems of using subdomains, recursive subdomains, dynamically loaded content technologies, search engine optimization of text content and others are described. It is shown that the task of collecting data from Internet resources is not only technological, but also to a greater extent knowledge-intensive, and since research is in an active phase, there is no “out-of-the-box” solution for it. The article will be useful to researchers in the field of Internet development, search engine developers, specialists in data retrieval and Internet technologies, as well as specialists in the field of creation and support of Internet resources and in the field of Internet marketing.

Keywords: Internet documents, data collection technologies, data retrieval, search engines, Internet resources

Acknowledgments: This work was supported by State Assignment for 2023 No. FFZF-2022-0005.

For citation: Kuleshov S. V., Zaytseva A. A. Phenomenological description of Internet documents collecting and processing. *Journal of Instrument Engineering.* 2023. Vol. 66, N 12. P. 1002—1010 (in Russian). DOI: 10.17586/0021-3454-2023-66-12-1002-1010.

Введение. С момента создания Тимом Бернерсом-Ли [1] протокола http (HyperText Transfer Protocol) в 1991 г. [2] и языка html (HyperText Markup Language), которые надолго определили наиболее удобный и доступный способ размещения данных в сети Интернет, а также способ работы с ним, прошло уже много лет, изменились некоторые базовые принципы, а также инструментарий. Возникает множество задач научного и технического характера, без решения которых уже невозможно в автоматическом режиме качественно и полноценно собирать данные из Интернета, например, при помощи программного бота.

Описанные в зарубежных книгах [3—9] технологии и методы извлечения данных из интернет-контента быстро устаревают и перестают отвечать новым требованиям. Это связано с тем, что в текущей фазе развития Интернет является сверхдинамичной* открытой системой, а методы сбора данных должны учитывать новые особенности, часть из которых будет рассмотрена ниже.

Изменение технологических особенностей представления данных на интернет-ресурсах приводит к созданию и развитию новых технологий получения данных, подобных веб-скрейпингу (web scraping) [10], который используется для синтаксического преобразования веб-страниц в более удобные для работы формы и может считаться технологией „очистки“ веб-контента [11].

К наиболее распространенным вариантам автоматической загрузки документов можно отнести: режим извлечения данных с использованием программного интерфейса (API) информационного ресурса; загрузку определенного набора страниц для извлечения данных из них (этот режим используется, например, в системах мониторинга погоды или курсов валют различных финансовых учреждений); полный сбор всех доступных страниц с ресурса с последующим их анализом и индексацией.

Первые два варианта автоматической загрузки документов достаточно широко используются и описаны как в научной литературе, так и в технической документации [12—15]. Рассмотрим режим сбора данных „харвестер“, когда (рис. 1) у бота априори нет информации о структуре информационного ресурса, и ботом используются те же механизмы, что использовал бы человек с помощью браузера, последовательно извлекая данные из загруженных страниц. При этом увеличивается объем данных об информационных ресурсах путем накопления ссылок из уже загруженных страниц.

Преимущество этого режима сбора данных состоит в том, что его работа возможна в отсутствие предварительного списка извлекаемых документов, схемы сайта и часто даже списка сайтов для обработки. Следовательно, множество обработанных ресурсов обычно не поддается прогнозированию.

Для того чтобы, с одной стороны, управлять процессом обработки только нужных информационных ресурсов, а с другой — не расходовать „впустую“ вычислительные ресурсы обработчика, можно выделить:

1) доверенные (определяемые при помощи „белого списка“) домены, имеющие наивысшие информационную ценность и приоритет в обработке;

* Все приведенные в статье фактические числовые и экспериментальные данные актуальны на начало 2023 года.

2) домены, созданные для целей „черного SEO“ (SEO — Search Engine Optimization), т.е. содержащие нелегальные, с точки зрения поисковых систем, методы продвижения сайта, могут определяться как автоматически во время обработки, так и при помощи „черного списка“;

3) определяемые при помощи „черного списка“ домены, которые должны быть исключены из обработки, поскольку содержат контент, противоречащий законодательству РФ [16];

4) обычные домены, не относящиеся к первым трем категориям.

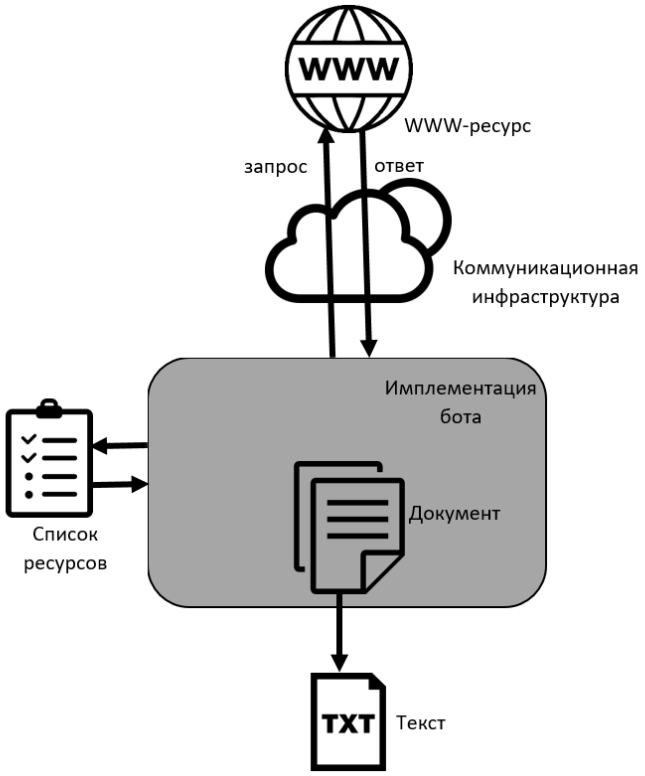

Рис. 1

В качестве „черного“ и „белого“ списков могут использоваться как локальная база данных, так и внешние информационные системы [16].

Множество документов, содержащихся внутри каждого домена, также может быть классифицировано (по уникальности: уникальные, дубли; по назначению: информационные, навигационные, служебные и т.д.), это позволяет оптимизировать работу бота в рамках одного домена. Примеры алгоритмов, предназначенных для построения оптимальной стратегии обхода доменов, рассматриваются, например, в [9].

Для повышения эффективности работы автоматических парсеров используются различные эвристические подходы: методы создания соответствующих значений веса; методы, учитывающие границы экрана и расположение элементов страницы в браузере, основанные на использовании различных характеристик; методы, использующие JavaScript и контент, доступный через веб-формы для имитации работы пользователя [17—25].

В настоящее время отдельную проблему при автоматическом сборе документов для построения систем мониторинга либо поисковых систем создают различные феномены — явления, возникающие в сети Интернет.

Можно выделить две большие группы феноменов: появившиеся вследствие развития SEO-технологий и явления, которые стали результатом развития технологий, обеспечивающих повышение безопасности пользователей и защиты размещаемых в сети Интернет данных.

К первой группе можно отнести следующие феномены:

Информационный ресурс — фабрика поддоменов. Поддомен (или субдомен), являющийся частью домена более высокого уровня. Традиционно поддомены используются для

структуризации информационных ресурсов, в том числе для связи сайтов различной тематики, для выделения отдельной мобильной версии сайта, для многорегионального продвижения сайтов с учетом регионального таргетинга и т.п.

Одновременно с этим в современном Интернете встречаются случаи, когда могут автоматически генерироваться сотни тысяч поддоменов, на которых располагается „сайт одной страницы“ для увеличения доли присутствия страниц ресурса в поисковом индексе. Реальный пример структуры таких поддоменов приведен на рис. 2, в которой экспериментально обнаружено более 181 тысячи таких поддоменов.

```
fotozona-svoimi-rukami-dlya-detey.infectology.spb.ru
kaspiyskiy-gruz-likiy-advayta-guantanamera-skachat.infectology.spb.ru
minusovka-moya-milaya-mama-svet-tvoih-glaz.infectology.spb.ru
nirvana-smells-like-spirit.infectology.spb.ru
honda-akkord-95-g-kak-snyat-balansirovochnye-valy.infectology.spb.ru
muzyka-krutaya-v-maynkrafte.infectology.spb.ru
white-town-your-woman-rington.infectology.spb.ru
noty-higo-de-la-luna.infectology.spb.ru
skachat-noty-novinku-britni-spirs.infectology.spb.ru
accidentally-in-love-noty-dlya-fortepiano.infectology.spb.ru
ya-nenavizhu-lyuboy-okean-potomu-chto-dazhe-on-tebya-hochet.infectology.spb.ru
pol-makkartni-yesterday-noty-dlya-sintezatara.infectology.spb.ru
treyler-dragon-age-4.infectology.spb.ru
noiz-ms-zhvachka-akkordy-dlya-pianino.infectology.spb.ru
skachat-noty-veter-gonit-zlye-tuchi.infectology.spb.ru
detskie-pesni-iz-multikov-sborniki-s-klipami.infectology.spb.ru
naverno-potomu-chto-vse-eto-moi-chuvstva-rington.infectology.spb.ru
uzhasy-2019-stivena-kinga.infectology.spb.ru
```

Рис. 2

В связи с тем, что для поисковой системы каждый поддомен — это отдельный сайт, система сбора данных должна анализировать их независимо, и „пессимизация“ одного поддомена не должна влиять на остальные.

Для борьбы с неконтролируемым возникновением поддоменов рекомендуется ограничивать число допустимых поддоменов в пределах одного домена, проверять число различных страниц одного поддомена или проверять, являются ли поддомены дубликатами друг друга. Известен, например, алгоритм, согласно которому в поисковой системе поддомен считается страницей второго уровня в пределах домена, если содержит менее 200 страниц [26].

Рекурсивные поддомены. Ситуация, когда все поддомены являются копией друг друга, может возникать при сочетании следующих факторов:

— используются настройки DNS для поддоменов с применением шаблонов (wildcards) [27], пример DNS-записи приведен на рис. 3;

— возникают (специально или случайно) ошибки при формировании ссылок внутри информационного ресурса, увеличивающие вложенность поддомена для каждого последующего перехода по ссылке (sub.example.com → sub.sub.example.com → sub.sub.sub.example.com → …).

```
*.example.com.   3600 IN  MX 10 host1.example.com.
```

Рис. 3

Вырожденным случаем рекурсивного поддомена является общепринятый префикс „www.“ в начале имени домена. Система сбора данных должна быть готова распознавать случай рекурсивных поддоменов путем анализа идентичности контента поддоменов и выполнить „склейку“ доменов-зеркал или проигнорировать ошибочные вложенные поддомены.

Расширение URL служебными полями для сбора статистики и управления рекламными кампаниями. Для гибкого управления сбором статистики и рекламными кампаниями существуют различные варианты добавления параметров в сам контент или в протоколы управления контентом. Широко используется метка UTM (Urchin Tracking Module; Urchin — компания, которая создала эти метки и начала их использовать, позднее на ее основе была

создана Google Analytics [28]). UTM-метки стали применяться в Яндекс.Метрике и других системах интернет-аналитики в связи с удобством их использования и универсальностью.

Технически UTM-метки — это дополнительные параметры в URL (utm_source, utm_medium, utm_campaign и другие), добавляемые к URL-адресу страницы, не влияющие на возвращаемый сервером http-контент. Другими словами, образуется множество различных URL, различающихся UTM-параметрами, но ведущих к одному документу на сервере.

Для предотвращения дублирования загружаемого харвестером контента необходимо удалять такие параметры из URL новых страниц до начала загрузки последних. Список параметров определяется стандартами сбора статистики (utmstat, openstat и др.).

С целью облегчения работы с такими параметрами разработчики предложили расширение для формата файла robots.txt [29], с помощью которого через директиву Clean-param перечисляются параметры, подлежащие удалению, но, к сожалению, этот стандарт используется очень редко и полагаться на наличие директивы при использовании UTM-меток нецелесообразно.

Спам-домены с дублирующимися документами. Дублирование документов в пределах домена может быть частичным или полным.

Полное дублирование обычно является причиной использования инструментов управления данными на сайте, что приводит к появлению документов-дублей, имеющих различный URL.

Частичное дублирование может быть вызвано как попыткой SEO-оптимизации, так и применением инструментов управления данными на сайте (использование фильтров и сортировок для данных), это наиболее сложный для выявления вид дублирования, особенно если дублирующиеся фрагменты текста перемешаны между собой или чередуются с фрагментами уникального текста.

Так, в результате проведенного эксперимента на 17 979 обработанных ботом в режиме харвестера доменов загружено 308 170 документов, из которых для 87 094 документов обнаружено совпадение более 75 % контента, а для 17 204 документов совпадение 100 %.

Основным препятствием для качественной фильтрации документов-дублей является высокая ресурсоемкость проверки степени идентичности каждого нового документа с уже имеющимися.

Разумным решением в таком случае будет сравнение документов в пределах домена с определением процента дублирования с последующим принятием решение о спам-ресурсе или частичной пессимизацией части страниц-дублей.

Поисковая оптимизация текстового контента. Наиболее сложна в решении задача предварительного анализа документов в случае использования поисковой оптимизации контента (SEO-оптимизации), результаты которой влияют на результаты работы поисковой системы в целом. Для решения используются статистические методы, методы, основанные на машинном обучении, а также большая группа эвристических методов определения „качества“ контента. Некоторые варианты эвристических решений были предложены в работах [30—32].

Группа феноменов, появившихся в результате развития технологий повышения безопасности пользователей и защиты данных:

Динамически загружаемый контент. Современные информационные ресурсы — это не просто набор статических страниц с гипертекстовой разметкой. Последние тенденции в веб-разработке приводят к тому, что сайты превращаются в сложные, большие javascript-приложения, по сути состоящие из одной html-страницы и подгружающие контент с помощью ajax-запросов к серверу [25]. Наиболее яркие примеры подобных ресурсов — dzen.ru, vk.com, aliexpress.ru. Среди преимуществ такого подхода к разработке — скорость работы сайтов, снижение трафика, перенос значительной части логики приложения с сервера на клиента, что приводит к снижению нагрузки на сервер. Однако, несмотря на очевидные преиму-

щества, подобный веб-сайт, фактически состоящий из одной страницы, практически не подлежит индексации при использовании харвестера без дополнительных инструментов по выполнению скрипта на клиентской стороне.

Для существующих на данный момент веб-ресурсов можно предложить следующие уровни, соответствующие степени использования динамического контента:

— Уровень 1. „Старые“ html-страницы: статический контент с html-разметкой и CSS-стилями;

— Уровень 2. html-контент с использованием скрипта на клиентской стороне, который не производит ajax-запросов, а манипулирует только тем контентом, который получен с сервера (например, динамическая сортировка таблиц или показ скрытого текста);

— Уровень 3. Onload-загрузчик: после загрузки страницы начинает работать скрипт и догружает дополнительный контент;

— Уровень 4. Клиентское приложение: пользовательский интерфейс модифицируется в соответствии с действиями пользователя; скрипт догружает данные недетерминировано. По такому принципу работают одностраничные приложения, „бесконечные“ списки, „умные догрузчики“.

Чем выше уровень динамичности, тем сложнее осуществлять загрузку документов в автоматическом режиме.

Защита от ботов. Некоторые интернет-ресурсы используют для защиты от действий ботов методы обнаружения, а также и блокировку от загрузки ботами страниц. Администраторы ресурсов могут настраивать блокировки программных ботов, используя следующие признаки:

— необычное поведение пользователя (например, десятки и сотни переходов на новую страницу сайта каждую секунду);

— повторяющиеся однотипные или безрезультатные действия (реальный пользователь не будет выполнять одни и те же задачи раз за разом);

— использование ссылок, которые содержатся только в коде веб-сайта и не видны обычным пользователям (ссылки-приманки) [33].

Способы блокировки, которые используют ресурсы после обнаружения программного бота:

— запрет доступа к ресурсу с определенного IP-адреса;

— выдача вместо страницы с контентом страницы с сообщением об ошибке;

— запрет идентификатора пользователя, являющегося, с точки зрения администратора сайта, злоумышленником, заходящим на сайт по аутентификации [33].

Отдельно необходимо отметить относительно новое явление, о котором впервые как о глобальной угрозе было заявлено с трибуны Форума ООН по вопросам управления Интернетом в 2019 году [34], — *фрагментацию пространства Интернет.* Глобальная сеть Интернет должна обеспечивать техническую возможность обмена данными каждого конечного устройства с любым другим конечным устройством, готовым эти данные получать. Одно и то же действие в глобальной сети должно приводить к одному и тому же результату, вне зависимости от географической локации и времени. Соответственно с технической точки зрения фрагментация Интернета проявляется там, где одни и те же действия могут приводить к разным результатам [35]. Технические, организационные, политические и иные причины могут вызывать ограничение связности сети Интернет. Это приводит к тому, что для пользователя оказываются недоступными множество URL, при этом конкретный набор недоступных документов зависит от текущей точки подключения (региона, провайдера). Так, могут оказаться недоступными как отдельные интернет-сервисы, так и целые группы адресов по географическому признаку.

Выводы. В статье показано, что задача сбора данных интернет-ресурсов является не просто технологической, но и наукоемкой, при этом для преодоления ряда проблем не существует готового „коробочного“ решения.

Приведен ряд особенностей, возникающих при веб-скрейпинге, харвестинге и в других случаях использования ботов для сбора данных в сети Интернет.

Проведен анализ феноменов — явлений, которые необходимо учитывать при автоматическом сборе документов в Интернете в задачах построения систем мониторинга либо поисковых систем. Большинство рассмотренных феноменов появилось вследствие развития SEO-технологий, однако необходимо также выделить группу явлений, которые стали результатом развития технологий, обеспечивающих повышение безопасности пользователей и защиты размещаемых в сети Интернет данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Berners-Lee T.* Information Management: A Proposal. CERN, March 1989, May 1990 [Электронный ресурс]: <https://www.dcs.gla.ac.uk/~wpc/grcs/bernerslee.pdf>.
2. RFC 1945 [Электронный ресурс]: <https://datatracker.ietf.org/doc/html/rfc1945>.
3. *Barnet B.* Memory Machines: The Evolution of Hypertext. Anthem Press, 2013.
4. *Olston C. and Najork M.* Web Crawling, Foundation and Trends // Information Retrieval. 2010. Vol. 4, N 3. P. 175—246.
5. *Najork M., Heydon A.* High-Performance Web Crawling // Handbook of Massive Data Sets. Massive Computing / Ed. by *J. Abello, P. M. Pardalos, M. G. C. Resende.* Springer, Boston, MA, 2002. Vol. 4. https://doi.org/10.1007/978-1-4615-0005-6_2.
6. *Laliwala Z., Shaikh A.* Web Crawling and Data Mining with Apache Nutch. Packt Publishing, 2013.
7. *Nasraoui O.* Web data mining: exploring hyperlinks, contents, and usage data // ACM SIGKDD Explorations Newsletter, 2008. DOI: https://doi.org/10.1145/1540276.1540281.
8. *Chakrabarti S.* Mining the Web: Discovering knowledge from hypertext data. Elsevier, 2003.
9. *Castillo C.* Effective web crawling // ACM SIGIR Forum. 2005. DOI: https://doi.org/10.1145/1067268.1067287.
10. *Boeing G., Waddell P.* New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings // Journal of Planning Education and Research. 2017. Vol. 37, N 4. DOI:10.2139/ssrn.2781297.
11. Practical Web Scraping for Data Science. Apress, Berkeley, CA, 2018. https://doi.org/10.1007/978-1-4842-3582-9_6.
12. *Bloch J.* How to design a good API and why it matters // Companion to the 21st ACM SIGPLAN Symp. on Object-oriented Programming Systems, Languages, and Applications. 2006. P. 506—507.
13. *Robillard M. P.* et al. Automated API property inference techniques // IEEE Transactions on Software Engineering. 2012. Vol. 39, N 5. P. 613—637.
14. *Ofoeda J., Boateng R., Effah J.* Application programming interface (API) research: A review of the past to inform the future // Intern. J. of Enterprise Information Systems (IJEIS). 2019. Vol. 15, N 3. P. 76—95.
15. *Qi L. et al.* Data-driven web APIs recommendation for building web applications // IEEE Transactions on Big Data. 2020. Vol. 8, N 3. P. 685—698.
16. Единый реестр доменных имен, указателей страниц сайтов в сети „Интернет“ и сетевых адресов, позволяющих идентифицировать сайты в сети „Интернет“, содержащие информацию, распространение которой в Российской Федерации запрещено [Электронный ресурс]: <https://eais.rkn.gov.ru/>.
17. HTML::LinkExtor - Extract links from an HTML document [Электронный ресурс]: <http://search.cpan.org/dist/HTML-Parser/lib/HTML/LinkExtor.pm>.
18. Немного на тему разработки веб-архивов [Электронный ресурс]: <http://habrahabr.ru/post/185816/>.

19. Насколько умны поисковые роботы? // Типичные ошибки внутренней оптимизации. Вып. 76 [Электронный ресурс]: <http://seopult.ru/subscribe.html?id=76>.

20. Google пытается проиндексировать Невидимую Сеть [Электронный ресурс]: <http://habrahabr.ru/post/23456/>.

21. Googlebot начал делать POST-запросы через Ajax [Электронный ресурс]: <http://habrahabr.ru/post/130258/>.

22. *Якушев А. В., Дейкстра Л.* Сетецентрические технологии сбора данных в Интернет [Электронный ресурс]: <http://socio.escience.ifmo.ru/content/files/file/network+centered.pdf>.

23. Поисковые технологии Яндекса [Электронный ресурс]: <http://download.yandex.ru/company/techno/YandexTech_1.pdf>.

24. Поисковые технологии или в чем загвоздка написать свой поисковик [Электронный ресурс]: <http://habrahabr.ru/post/123671/>.

25. HtmlUnit – JavaScript Tutorial [Электронный ресурс]: <https://htmlunit.sourceforge.io/javascript-howto.html>.

26. Поддомены: что это такое и зачем они нужны? [Электронный ресурс]: <https://timeweb.com/ru/community/articles/poddomeny-chto-eto-takoe-i-zachem-oni-nuzhny>.

27. RFC1035: Domain Names – Implementation And Specification. Network Working Group, November 1987 [Электронный ресурс]: <http://www.faqs.org/rfcs/rfc1035.html>.

28. Большой гайд по UTM-меткам: как узнать, откуда приходят пользователи [Электронный ресурс]: <https://habr.com/ru/company/click/blog/478758/>.

29. A Standard for Robot Exclusion [Электронный ресурс]: <http://www.robotstxt.org/orig.html>.

30. *Kuleshov S., Zaytseva A., Aksenov A*. Natural Language Search and Associative-Ontology Matching Algorithms Based on Graph Representation of Texts // Intelligent Systems Applications in Software Engineering. Advances in Intelligent Systems and Computing / Ed. by *R. Silhavy, P. Silhavy, Z. Prokopova.* Springer, Cham, 2019. Vol. 1046. P. 285—294. DOI 10.1007/978-3-030-30329-7_26.

31. *Михайлов С. Н., Кулешов С. В.* Экспертный мониторинг неструктурированных информационных ресурсов в интересах информационно-аналитического обеспечения космических исследований // Изв. Юго-Западного государственного университета. 2013. № 6-2(51). С. 40—43.

32. *Зайцева А. А., Кулешов С. В., Михайлов С. Н.* Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов // Тр. СПИИРАН. 2014. Вып. 37. С. 144—155.

33. *Москаленко А. А., Лапонина О. Р., Сухомлин В. А.* Разработка приложения веб-скрапинга с возможностями обхода блокировок // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 2. С. 413—420.

34. *Игнатьев А. Г., Линдре Ю. А*. Актуальные тренды регулирования Интернета: от открытого пространства безграничной свободы к региональной и страновой фрагментации. М.: Центр компетенций по глобальной ИТ-кооперации, 2023. 30 с. EDN EHZLLW.

35. *Куликова А. В.* О фрагментации интернета: старые вопросы и новые вызовы // Индекс безопасности. 2015. Т. 21, № 1(112). С. 115—120. EDN XBFPKZ.

Сведения об авторах

Сергей Викторович Кулешов — д-р техн. наук, профессор РАН; Санкт-Петербургский Федеральный исследовательский центр РАН, лаборатория автоматизации научных исследований, Санкт-Петербургский институт информатики и автоматизации РАН; гл. научный сотрудник; E-mail: kuleshov@iias.spb.su

Александра Алексеевна Зайцева — канд. техн. наук; Санкт-Петербургский Федеральный исследовательский центр РАН, лаборатория автоматизации научных исследований, Санкт-Петербургский институт информатики и автоматизации РАН; ст. научный сотрудник; E-mail: cher@iias.spb.su

Поступила в редакцию 28.08.2023; одобрена после рецензирования 07.09.2023; принята к публикации 27.10.2023.

REFERENCES

1. Berners-Lee T. *Information Management: A Proposal*, CERN, March 1989, May 1990.
2. *RFC 1945*, https://datatracker.ietf.org/doc/html/rfc1945.
3. Barnet B. *Memory Machines: The Evolution of Hypertext*, Anthem Press, 2013.
4. Olston C. and Najork M. *Information Retrieval*, 2010, no. 3(4), pp. 175–246.
5. Najork M., Heydon A. *High-Performance Web Crawling* in Handbook of Massive Data Sets. Massive Computing, Springer, 2002, vol. 4, https://doi.org/10.1007/978-1-4615-0005-6_2.
6. Laliwala Z., Shaikh A. *Web Crawling and Data Mining with Apache Nutch*, Packt Publishing, 2013.
7. Nasraoui O. *ACM SIGKDD Explorations Newsletter*, 2008, DOI: https://doi.org/10.1145/1540276.1540281.
8. Chakrabarti S. *Mining the Web: Discovering knowledge from hypertext data*, Elsevier, 2003.
9. Castillo C. *ACM SIGIR Forum*, 2005, DOI: https://doi.org/10.1145/1067268.1067287.
10. Boeing G., Waddell P. *Journal of Planning Education and Research*, 2017, no. 4(37), DOI:10.2139/ssrn.2781297.
11. *Practical Web Scraping for Data Science*, Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4842-3582-9_6.
12. Bloch J. *Companion to the 21st ACM SIGPLAN symposium on Object-oriented programming systems, languages, and applications*, 2006, pp. 506–507.
13. Robillard M.P. et al. *IEEE Transactions on Software Engineering*, 2012, no. 5(39), pp. 613–637.
14. Ofoeda J., Boateng R., Effah J. *International Journal of Enterprise Information Systems (IJEIS)*, 2019, no. 3(15), pp. 76–95.
15. Qi L. et al. *IEEE transactions on big data*, 2020, no. 3(8), pp. 685–698.
16. https://eais.rkn.gov.ru/. (in Russ.)
17. *HTML::LinkExtor - Extract links from an HTML document*, http://search.cpan.org/dist/HTML-Parser/lib/HTML/LinkExtor.pm.
18. http://habrahabr.ru/post/185816/. (in Russ.)
19. http://seopult.ru/subscribe.html?id=76. (in Russ.)
20. http://habrahabr.ru/post/23456/. (in Russ.)
21. http://habrahabr.ru/post/130258/. (in Russ.)
22. http://socio.escience.ifmo.ru/content/files/file/network+centered.pdf. (in Russ.)
23. http://download.yandex.ru/company/techno/YandexTech_1.pdf. (in Russ.)
24. http://habrahabr.ru/post/123671/. (in Russ.)
25. *HtmlUnit – JavaScript Tutorial*, https://htmlunit.sourceforge.io/javascript-howto.html.
26. https://timeweb.com/ru/community/articles/poddomeny-chto-eto-takoe-i-zachem-oni-nuzhny. (in Russ.)
27. *RFC1035: Domain Names – Implementation and Specification. Network Working Group*, November 1987, http://www.faqs.org/rfcs/rfc1035.htm>.
28. https://habr.com/ru/company/click/blog/478758/. (in Russ.)
29. *A Standard for Robot Exclusion*, http://www.robotstxt.org/orig.html.
30. Kuleshov S., Zaytseva A., Aksenov A. *Natural Language Search and Associative-Ontology Matching Algorithms Based on Graph Representation of Texts* in Intelligent Systems Applications in Software Engineering. Advances in Intelligent Systems and Computing, Springer, Cham, 2019, vol. 1046, DOI 10.1007/978-3-030-30329-7_26.
31. Mikhailov S.N., Kuleshov S.V. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta* (Proceedings of the Southwest State University), 2013, no. 6-2(51), pp. 40–43. (in Russ.)
32. Zaytseva A.A., Kuleshov S.V., Mikhailov S.N. *SPIIRAS Proceedings*, 2014, no. 37, pp. 144–155. (in Russ.)
33. Moskalenko A.A., Laponina O.R., Sukhomlin V.A. *Modern Information Technology and IT-education*, 2019, no. 2(15), pp. 413–420. (in Russ.)
34. Ignatiev A.G., Lindre Yu.A. *Aktual'nyye trendy regulirovaniya Interneta: ot otkrytogo prostranstva bezgranichnoy svobody k regional'noy i stranovoy fragmentatsii* (Current Trends in Internet Regulation: from an open Space of Unlimited Freedom to Regional and Country Fragmentation), Moscow, 2023, 30 p., EDN EHZLLW. (in Russ.)
35. Kulikova A.V. *Indeks bezopasnosti*, 2015, no. 1(21), pp. 115–120, EDN XBFPKZ. (in Russ.)

Data on authors

Sergey V. Kuleshov	—	Dr. Sci., Professor RAS; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Research Automation Laboratory; Senior Researcher; E-mail: kuleshov@iias.spb.su
Alexandra A. Zaytseva	—	PhD; St. Petersburg Federal Research Center of the RAS, Senior Researcher; E-mail: cher@iias.spb.su

Received 28.08.2023; approved after reviewing 07.09.2023; accepted for publication 27.10.2023.