МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННЫХ СИСТЕМ

MATHEMATICAL AND SOFTWARE SUPPORT OF INFORMATION SYSTEMS

УДК 004.912: 004.822
DOI: 10.17586/0021-3454-2022-65-11-826-832

ФОРМИРОВАНИЕ ЯДРА ДОКУМЕНТОВ В СИСТЕМАХ ИНТЕРНЕТ-МОНИТОРИНГА В УСЛОВИЯХ РЕСУРСНЫХ ОГРАНИЧЕНИЙ

С. В. Кулешов, А. А. Зайцева*, А. Ю. Аксенов

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

Аннотация. Рассматриваются особенности разработки систем интернет-мониторинга открытого типа с неограниченным количеством источников в условиях ограниченного объема систем хранения собранных данных. Цель работы — решение задачи формирования множества документов минимально необходимого размера (ядра документов), отвечающего требованиям репрезентативности и вариативности тем при мониторниге сети Интернет. Для формализации и решения поставленной задачи разработана теоретико-множественная модель ядра документов. Предложенный подход отличается использованием вытесняющего алгоритма, поддерживающего в базе данных наличие только актуальных документов в пределах доступного объема системы хранения данных. Приведены результаты эксперимента с использованием реальных данных, подтверждающие применимость разработанной модели. Предложенный подход может быть использован в ряде практических задач, в частности для поиска в сети Интернет сведений (документов, страниц), по которым отсутствует априорная информация, необходимая для поиска по ключевым словам.

Ключевые слова: *ядро документов, мониторинг, краулер, поиск документов, интернет-ресурсы*

Благодарности: работа выполнена в рамках реализации Государственного задания на 2022 г., № FFZF-2022-0005.

Ссылка для цитирования: *Кулешов С. В., Зайцева А. А., Аксенов А. Ю.* Формирование ядра документов в системах интернет-мониторинга в условиях ресурсных ограничений // Изв. вузов. Приборостроение. 2022. Т. 65, № 11. С. 826—832. DOI: 10.17586/0021-3454-2022-65-11-826-832.

FORMATION OF THE CORE OF DOCUMENTS IN INTERNET MONITORING SYSTEMS UNDER RESOURCE CONSTRAINTS

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

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

Abstract. The features of development of open-type Internet monitoring systems with an unlimited number of sources in conditions of a limited amount of data storage systems are considered. The purpose of the work is to solve the problem of forming a set of documents of the minimum required size (the core of documents) that meets the requirements of representativeness and variability of topics when monitoring the Internet. To formalize and solve the problem, a set-theoretic model of the document core is developed. The proposed approach is distinguished by the use of a preemptive algorithm that supports the availability of only relevant documents in the database within the available volume of the data storage system. The results of an experiment using real data confirming the applicability of the devel-

oped model are presented. The proposed approach can be used in a number of practical tasks, in particular for searching the Internet for information (documents, pages) for which there is no a priori information needed for keyword search.

Keywords: core of documents, monitoring, crawler, document search, Internet resources

Acknowledgment: the work was carried out as part of the implementation of the State Task for 2022, N FFZF-2022-0005.

For citation: Kuleshov S. V., Zaytseva A. A., Aksenov A. Yu. Formation of the core of documents in Internet monitoring systems under resource constraints. *Journal of Instrument Engineering.* 2022. Vol. 65, N 11. P. 826—832 (in Russian). DOI: 10.17586/0021-3454-2022-65-11-826-832.

Технологии сбора данных в сети Интернет в настоящее время являются технологиями общего назначения, решающими широкий спектр практических задач. Если в начале развития интернет-сетей краулер (программный робот, используемый для обнаружения новых страниц в Интернете, их загрузки и анализа) являлся определяющим признаком поисковой системы, то на нынешнем уровне развития Всемирной паутины эта технология используется даже в небольших проектах, требующих получения актуальных данных из внешних источников.

Большинство практических задач решаются с использованием „закрытого“ (closed domain) мониторинга, при котором сбор данных ведется по конечному списку ресурсов либо конкретному набору документов (мониторинг социальных сетей [1], анализ бизнес-активности конкурентов [2], агрегирование новостей [3], анализ общественного мнения, аналитика изменения цен [4] и тому подобные задачи). Проблема решения задач при ограниченных ресурсах памяти рассматривалась, например, в [5—7] и до сих пор является актуальной, несмотря на все возрастающую емкость систем хранения данных.

Системы сбора данных открытого типа не ограничены конкретным списком ресурсов и могут работать, собирая все доступные документы путем анализа ссылок на уже известные документы, постоянно расширяя список обработанных документов [8—10]. В англоязычной литературе подобные системы сбора данных иногда называются data harvesters [11]. Такие системы сталкиваются с проблемой ограниченных ресурсов, которые могут проявляться как в нехватке пропускной способности сети и вычислительной мощности серверов для обработки постоянно растущей очереди входящих документов, так и в принципиальной ограниченности систем хранения данных (СХД) для размещения постоянно увеличивающегося объема информации [12].

В условиях ограниченных ресурсов возможны различные подходы к решению указанных проблем. Традиционно используется фильтрация контента по различным критериям [13—15], позволяющим на этапе предобработки исключать часть документов из списка, предназначенного для сохранения в СХД. Данный подход позволяет исключать документы, бесполезные с точки зрения решаемой задачи, из очереди размещения в СХД. В зависимости от задачи это могут быть рекламные страницы, форумы, социальные сети, сайты магазинов, ресурсы на иностранном языке, фото- и видео- хостинги и др.

В ситуации когда фильтрация контента не позволяет преодолеть проблему постепенного переполнения СХД (в случаях когда доступных для обработки документов потенциально больше, чем возможностей для их хранения) приходится решать задачу формирования минимально необходимого множества документов (назовем данное множество ядром), производя постепенное замещение документов, признанных неподходящими по заранее заданному критерию, на новые в процессе работы системы мониторинга.

Такими критериями могут быть новизна, степень релевантности документа заданной теме, полнота представления темы имеющимися документами, а также эвристические

критерии, позволяющие оценить субъективные факторы, такие, например, как полезность документа для человека или качество самого документа.

Данный подход применим в ряде практических задач, в частности для поиска в сети Интернет сведений (документов, страниц), по которым отсутствует априорная информация, необходимая для поиска по ключевым словам. Это может быть информация о новых научных направлениях, новых технологиях, новинках на рынках продукции, интересных фактах из области общественной жизни, политики, экономики. Соответственно, требуется разработка вытесняющих алгоритмов обновления ядра, учитывающих не только временную метку документа, но и показатель его полезности, рассчитываемый по результатам внутреннего ранжирования.

На основании вышеизложенного в настоящей статье решается задача формализации построения ядра документов как задача построения отображения $A \to K$, где A — множество всех документов, доступных в Интернете в открытом доступе, имеющих URL (уникальный адрес документа или ресурса в сети Интернет), K — множество документов, входящих в ядро и сохраняемых в СХД.

Введем дополнительные обозначения:

L — множество „legacy“ документов, т.е. документов, содержащихся в ядре, но отсутствующих в интернет-доступе по тем же URL;

R — множество документов, размещение которых в ядре нецелесообразно: документы рекламного характера; документы, содержащие запрещенный законодательством контент; документы специального вида, имеющие целью поисковую оптимизацию (SEO-оптимизацию);

U — множество всех имеющихся в момент времени t_{current} документов в сети Интернет;

D — множество загруженных системой документов d на момент времени t_{current}, $K \subseteq D \subseteq U$.

Множество U служит для теоретической верхней оценки полноты ядра. Полнота ядра может быть оценена как $|K|/|G|$, где $\forall d_i \in G$, если $\forall d \in U, d \notin R$.

В связи с тем, что ядро, как и сам Интернет, является динамически меняющейся структурой, все вышеперечисленные множества определены в момент времени t_{current}:

$$K \cup L = D_{t-1}, \quad D \cap A = K.$$

В свою очередь, множество K делится на подмножества K_1 — актуализированных и K_2 — не актуализированных на данной итерации алгоритма документов:

$$K = K_1 \cap K_2.$$

На рис. 1 отображено графическое представление теоретико-множественной модели принципа формирования ядра документов и работы вытесняющего алгоритма.

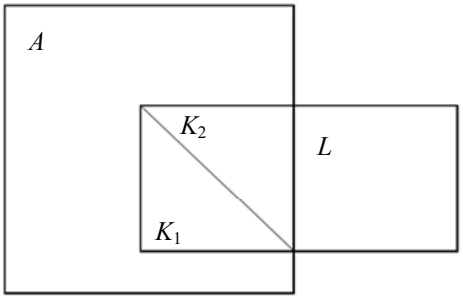

Рис. 1

Множество документов ядра (множество K) должно формироваться при следующих условиях:

— минимизируется дублирование документов:

$$\forall d_i \in K, d_j \in K, \lim_{t\to\infty} \left| d_i \equiv d_j \right| \to 0;$$

— минимизируется количество „legacy“ документов $\lim_{t\to\infty} \left|L\right| \to 0$;

— условие ограниченных ресурсов системы хранения данных выражается в виде

$$\lim_{t\to\infty} \left|K\right| \to k_{\max},$$

т.е. количество документов, входящих в ядро, не должно превышать возможности системы, ограниченной техническими ресурсами ($k_{\max}$ — максимальное количество документов для СХД).

Для доменов удовлетворительного качества все документы должны присутствовать в максимальном объеме, для доменов плохого качества (с низким суммарным показателем качества документов) наличие документов в ядре должно быть минимальным.

Предложенный подход основан на вытесняющем алгоритме, который, используя очередь документов, подлежащих обработке (множество K_1), последовательно проверяет каждый документ на актуальность (если документ уже есть в множестве D) либо пытается разместить его в ядре документов K (если документ имеется в множестве A, но ранее не загружался и не размещался в СХД, т.е. не принадлежит множеству D).

При работе алгоритма используется комплексная эвристическая функция $e(d)$ оценки качества документа [13]:

$$e(d) = \begin{cases} \text{true, если документ } d \text{ удовлетворяет требованиям качества,} \\ \text{false, если иначе,} \end{cases}$$

которая позволяет классифицировать документы перед помещением их в ядро.

За одну итерацию каждый известный домен обрабатывается один раз, при этом показатель качества q домена M пересчитывается после каждой итерации:

$$q(M) = \min\left(100, 1 + \frac{n_{\text{good}}}{\left|M\right|}\right),$$

где $n_{\text{good}} = \left|G\right|$ — количество документов, признанных качественными, т.е. $\forall d \in G : e(d) = \text{true}$; $\left|M\right|$ — общее количество документов, загруженных для данного домена (мощность множества M).

Значение порога вытеснения документов Q определяется после каждой итерации на основе вычисления объема документов, значение показателя качества домена $q(M)$ которых больше заданного.

Итоговое решающее правило на каждой итерации алгоритма для документа d имеет следующий вид:

— новый документ принадлежит ядру (помещается на хранение в СХД) $d \in K$, если $d \notin L \vee d \notin R \vee e(d) = \text{true}$;

— документ исключается из ядра, если $d \in L \vee d \in M_i, q\left(M_i\right) < Q$.

Условие минимизации дублирования документов в ядре достигается за счет дополнительных алгоритмов поиска схожих документов во время каждой итерации алгоритма замещения.

Проведенный с использованием реальных данных эксперимент показал применимость предложенного алгоритма для поддержания актуальной коллекции (ядра) документов в процессе непрерывного мониторинга вновь появляющихся документов без превышения ограничения на количество хранимых экземпляров документов, накладываемого системой хранения данных. На рис. 2 приведен график фактического распределения количества замещаемых документов (N) при различных значениях Q, полученных в результате экспериментальной проверки.

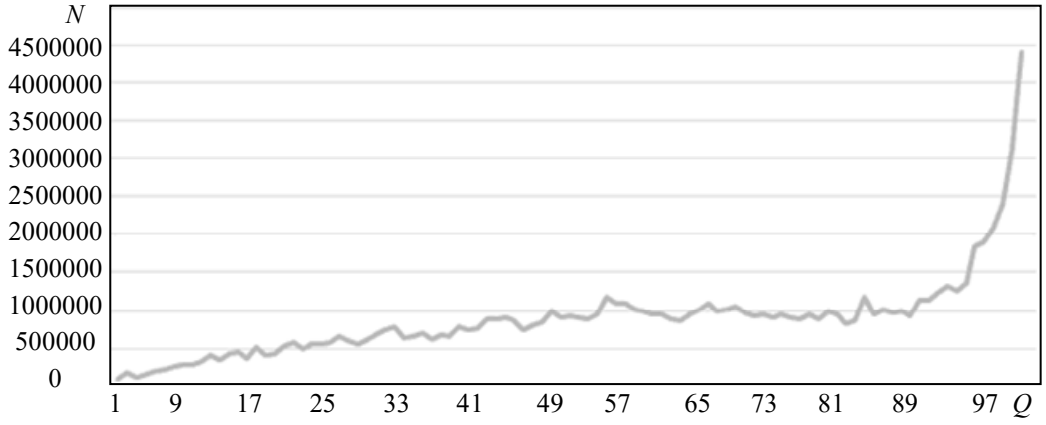

Рис. 2

Выводы. В статье решена задача формализации построения ядра документов в системах интернет-мониторинга в условиях ограниченного объема СХД. Предложенный подход отличается использованием вытесняющего алгоритма, поддерживающего в базе данных наличие только актуальных документов в пределах доступного объема СХД.

Использование данного подхода позволяет поддерживать актуальное состояние ядра документов в СХД ограниченного объема, не превышая заданное количество $k_{\max}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Zachlod C., Samuel O., Ochsner A., Werthmüller S.* Analytics of social media data – state of characteristics and application // Journal of Business Research. 2022. Vol. 144, P. 1064—1076. DOI: 10.1016/j.jbusres.2022.02.016.

2. *Fink C., Toivonen T., Correia R. A., Di Minin E.* Mapping the online songbird trade in Indonesia // Applied Geography. 2021. P. 134. DOI:10.1016/j.apgeog.2021.102505.

3. *Han H., Wang C., Zhao Y., Shu M., Wang W., Min Y.* SSLE: A framework for evaluating the “Filter bubble” effect on the news aggregator and recommenders // World Wide Web. 2022. N 25(3). P. 1169—1195. DOI: 10.1007/s11280-022-01031-4.

4. *Krewinkel A., Sünkler S., Lewandowski D.* et al. Concept for automated computer-aided identification and evaluation of potentially non-compliant food products traded via electronic commerce // Food Control. 2016. N 61, P. 204—212. DOI:10.1016/j.foodcont.2015.09.039.

5. *Беляевский К. О.* Формирование октодерева по облаку точек при ограничении объема оперативной памяти // Научно-технический вестник СПбПУ. Информатика. Телекоммуникации. Управление. 2019. Т. 12, № 4. С. 97—110.

6. *Puzak T.R.* Analysis of Cache Replacement-Algorithms: Doctor’s Thesis. 1985.

7. *Wilson P. R.* et al. Dynamic storage allocation: A survey and critical review // Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 1995. Vol. 986. P. 1—116.

8. *Laliwala Z., Shaikh A.* Web Crawling and Data Mining with Apache Nutch. Packt Publ., 2013.

9. *Nasraoui O.* Web data mining: exploring hyperlinks, contents, and usage data // ACM SIGKDD Explorations Newsletter. 2008.

10. *Van den Broucke S., Baesens B.* From Web Scraping to Web Crawling. Practical Web Scraping for Data Science. Berkeley, CA: Apress, 2018. P. 155—172.

11. *Alkalbani A. M., Hussain W., Kim J. Y.* A Centralised Cloud Services Repository (CCSR) Framework for Optimal Cloud Service Advertisement Discovery from Heterogenous Web Portals // IEEE Access. 2019. Vol. 7. P. 128213—128223. DOI: 10.1109/ACCESS.2019.2939543.

12. *Wu Z., Cai Z., Tang, X., Xu Y., Deng T.* A forward and backward private oblivious RAM for storage outsourcing on edge-cloud computing // Journal of Parallel and Distributed Computing. 2022. Vol. 166. P. 1—14. DOI: 10.1016/j.jpdc.2022.04.008.

13. *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.

14. *Кулешов С. В., Зайцева А. А., Левашкин С. П.* Технологии и принципы сбора и обработки неструктурированных распределенных данных с учетом современных особенностей предоставления медиа-контента // Информатизация и связь. 2020. № 4. С. 62—66.

15. *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, CoMeSySo 2019; Advances in Intelligent Systems and Computing. 2019. Vol. 1046. P. 7—26. DOI 10.1007/978-3-030-30329-7_26.

Сведения об авторах

Сергей Викторович Кулешов — д-р техн. наук, профессор; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; гл. научный сотрудник; E-mail: kuleshov@iias.spb.su

Александра Алексеевна Зайцева — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; ст. научный сотрудник; E-mail: cher@iias.spb.su

Алексей Юрьевич Аксенов — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; ст. научный сотрудник; E-mail: a_aksenov@iias.spb.su

Поступила в редакцию 18.07.2022; одобрена после рецензирования 27.07.2022; принята к публикации 30.09.2022.

REFERENCES

1. Zachlod C., Samuel O., Ochsner A., & Werthmüller S. *Journal of Business Research*, 2022, vol. 144, pp. 1064–1076, DOI: 10.1016/j.jbusres.2022.02.016.
2. Fink C., Toivonen T., Correia R. A., & Di Minin E. *Applied Geography*, 2021, pp. 134, DOI: 10.1016/j.apgeog.2021.102505.
3. Han H., Wang C., Zhao Y., Shu M., Wang W., & Min Y. *World Wide Web*, 2022, no. 3(25), pp. 1169–1195, DOI: 10.1007/s11280-022-01031-4.
4. Krewinkel A., Sünkler S., Lewandowski D. et al. *Food Control*, 2016, vol. 61, pp. 204–212, DOI: 10.1016/j.foodcont.2015.09.039.
5. Beliaevskii K.O. *Peter the Great St. Petersburg Polytechnic University. Computing, Telecommunications and Control*, 2019, no. 4(12), pp. 97–110. (in Russ.)
6. Puzak T.R. *Analysis of Cache Replacement-Algorithms,* Doctor's thesis, 1985.
7. Wilson P.R. et al. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1995, vol. 986, pp. 1–116.
8. Laliwala Z., Shaikh A. *Web Crawling and Data Mining with Apache Nutch*., Packt Publishing, 2013.
9. Nasraoui O. *Computer Science*, 2008, DOI:10.1145/1540276.1540281.
10. Van den Broucke S., Baesens B. *From Web Scraping to Web Crawling. Practical Web Scraping for Data Science*, Apress – Berkeley, CA, 2018, pp. 155–172.
11. Alkalbani A.M., Hussain W. & Kim J.Y. *IEEE Access*, 2019, vol. 7, pp. 128213–128223, DOI: 10.1109/ACCESS.2019.2939543.
12. Wu Z., Cai Z., Tang, X., Xu Y., & Deng T. *Journal of Parallel and Distributed Computing*, 2022, vol. 166, pp. 1–14, DOI:10.1016/j.jpdc.2022.04.008.
13. Zaitseva A.A., Kuleshov S.V., Mikhailov S.N. *Trudy SPIIRAN* (SPIIRAS Proceedings), 2014, no. 37, pp. 144—155. (in Russ.)
14. Kuleshov S.V., Zaytseva A.A., Levashkin S.P. *Informatization and communication*, 2020, no. 5, pp. 22–28. (in Russ.)
15. Kuleshov S., Zaytseva A., Aksenov A. *Systems Applications in Software Engineering. CoMeSySo 2019. Advances in Intelligent Systems and Computing*, 2019, vol. 1046, pp. 7–26, DOI 10.1007/978-3-030-30329-7_26.

Data on authors

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

Alexandra A. Zaytseva — PhD; 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: cher@iias.spb.su

Alexey Yu. Aksenov — PhD; 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: a_aksenov@iias.spb.su

Received 18.07.2022; approved after reviewing 27.07.2022; accepted for publication 30.09.2022.