УДК 001.8:004.738.5
DOI 10.20913/2618-7515-2020-1-86-97

ИНФРАСТРУКТУРА ДЛЯ ШИРОКОМАСШТАБНОГО СБОРА ВЕБОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ

INFRASTRUCTURE FOR LARGE SCALE HARVESTING OF WEBOMETRIC INDICATORS

© ***Косяков Денис Викторович***
заместитель директора по развитию, Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук (ГПНТБ СО РАН), научный сотрудник, Институт нефтегазовой геологии и геофизики им. А. А. Трофимука Сибирского отделения Российской академии наук (ИНГГ СО РАН), Новосибирск, Россия, kosyakov@spsl.nsc.ru

Kosyakov Denis Viktorovich
Deputy Director for Development, State Public Scientific Technological Library of the Siberian Branch of the Russian Academy of Sciences (SPSTL SB RAS), Trofimuk Institute of Petroleum Geology and Geophysics of the Siberian Branch of the Russian Academy of Sciences (IPGG SB RAS), Novosibirsk, Russia, kosyakov@spsl.nsc.ru

Основной целью научных рейтингов является не только сравнение, но и стимулирование развития университетов и научных институтов, улучшение результативности их образовательных и исследовательских процессов. Однако существующие вебометрические рейтинги плохо для этого подходят из-за отсутствия возможности их анализа и достоверизации. Решением проблемы является изменение подхода к вебометрическим измерениям, в основе которого лежат принципы регулярного сбора и открытости исходных данных. Рассматриваются индикаторы, использующиеся в задачах академической вебометрики, их достоверность и устойчивость, обосновывается необходимость регулярного сбора значений этих индикаторов для повышения качества данных и анализа их динамики. На основе опыта реализации проекта по исследованию российского академического веб-пространства (http://www.webometrix.ru) анализируются проблемы, возникающие при сборе значений вебометрических индикаторов сайтов научно-исследовательских организаций и учреждений высшего образования в мировом масштабе с частотой не менее 1 раза в месяц. Описывается реализованная автором распределенная система сбора значений вебометрических индикаторов и производится оценка необходимой степени распараллеливания процесса. Разработанный подход является универсальным для задач сбора больших объемов информации методом извлечения данных со страниц веб-сайтов, а сбор вебометрических данных является также актуальным для задач анализа других тематических сегментов веб-пространства, например сайтов библиотек.

Ключевые слова: *вебометрика, научные организации, вузы, вебсайты, индикаторы*

The main purpose of scientific rankings is not only comparison, but also stimulating the development of universities and research institutes, improving the effectiveness of their educational and research processes. However, existing webometric ratings are not suitable for this task because of the lack of the possibility of their analysis and verification. The solution is to change the approach to webometric measurements, to perform them based on the principles of regular collection and openness of source data. The article discusses the indicators used in the tasks of academic webometrics, their reliability and stability, substantiates the need for regular collection of the values of these indicators to improve the quality of data and analyze their dynamics. Based on the experience of implementing a project to research the Russian academic web space (http://www.webometrix.ru), problems are analyzed that arise when collecting the values of webometric indicators of sites of research organizations and institutions of higher education on a global scale with a frequency of at least 1 time per month. The author describes a distributed system for collecting values of webometric indicators and evaluates the necessary degree of parallelization of the process. The developed approach is universal for the tasks of collecting large amounts of information by the method of extracting data from website pages, and the collection of webometric data is also relevant for the tasks of analyzing other thematic segments of web space, such as library websites.

Keywords: *webometrics, research institutions, higher educational institutions, websites, indicators*

Введение

Возможно ли изучить и оценить объемы научной информации, опубликованной в World Wide Web? Работы нескольких групп исследователей, начатые в середине 1990-х гг., посвященные анализу различных вебометрических индикаторов [1–4], привели к созданию в 2008 г. постоянно обновляющегося мирового рейтинга университетов, исследовательских центров и репозиториев (www.webometrics.info) [5]. В последующие годы этот рейтинг с некоторыми вариациями был многократно повторен на национальных, региональных и тематических уровнях группами исследователей из нескольких стран [6–12]. Часть работ носила разовый характер, некоторые продолжались (и продолжаются) на регулярной основе в течение длительного промежутка времени.

В значительной степени интерес к количественному измерению академического веб-пространства был обусловлен активным ростом количества результатов научных исследований, публикующихся онлайн. Авторы рейтинга webometrics.info отмечают, что «влияние электронных публикаций значительно превышает влияние традиционных журналов и книг. Вебсайты являются наиболее эффективным и дешевым способом усиления всех трех основных академических задач: образования, исследований и передачи знаний» [5]. Важность сетевых коммуникаций в этой области с тех пор только увеличивается, в значительной степени в связи с массовым принятием концепции открытой науки.

Опубликованные он-лайн сведения об исследователях, научных проектах, результатах, учебные материалы не только облегчают информационный обмен в академической среде, но и служат важным ресурсом для новых поколений интеллектуальных информационных систем, основанных на технологиях автоматического анализа текстов на естественных языках, извлечения данных и знаний из неструктурированной информации. Технологии искусственного интеллекта уже активно используются ведущими коммерческими поисковыми системами, все большее распространение получают элементы семантического веба, что делает информацию в веб-пространстве все более доступной для автоматизированной обработки. По данным исследования 2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground, проведенного компанией Imperva (https://www.incapsula.com/blog/bot-traffic-report-2015.html), около 50% «посетителей» веб-сайтов являются разного рода роботами. Близкие данные приводятся в исследовании компании Statista (https://www.statista.com/statistics/670782/bot-traffic-share).

В связи с этим не теряет актуальности задача количественного измерения академического веба с целью анализа динамики его развития и оценки объемов и эффективности публикации академических материалов.

Индикаторы

Развитие коммерческих поисковых систем предоставило исследователям в области вебометрики необходимый инструментарий. С течением времени сложился определенный консенсус в части основных вебометрических индикаторов. В общем смысле индикаторы могут быть разделены на две группы – определяющие видимость информации и ее качество. В качестве индикаторов видимости рассматривается размер сайта (или группы сайтов, расположенных в одном домене), измеряемый как количество результатов поиска с ограничением по доменному имени. Качество предлагается оценивать показателями «цитирования», то есть количеством внешних ссылок на материалы, размещенные на сайте. В роли уточняющих видимость показателей, в некоторой степени учитывающих и качество информации, рассматривается измеряемое с помощью коммерческих поисковых систем количество полнотекстовых документов в распространенных форматах DOC(X), PPT(X), PDF, PS и количество документов, проиндексированных специализированной академической поисковой системой Google Scholar, в качестве таких документов могут выступать как полные тексты публикаций, так и подробные библиографические метаданные.

Количество внешних ссылок на сайты изначально измерялось с помощью тех же самых поисковых систем, но позже, в связи со снижением значения этого фактора в ранжировании результатов поиска, коммерческие поисковики перестали обновлять эту информацию. Однако имеется несколько коммерческих сервисов, занимающихся сбором и анализом внешних ссылок с использованием собственных роботов, в их числе сервисы Majestic SEO (https://majestic.com) и Ahrefs (https://ahrefs.com). Объем базы данных (БД) Ahrefs уже в 2015 году превысил миллиард проиндексированных страниц (https://ahrefs.com/blog/one-billion-pages-crawled-by-content-explorer), что, однако, все еще значительно меньше, чем размеры индексов коммерческих поисковых систем. С помощью этих сервисов можно получить оценку не только общего количества внешних ссылок, но и количества отдельных доменов, с которых приходят внешние ссылки, IP-адресов и IP-подсетей. Все вместе позволяет в некоторой степени оценить среднее качество ссылок.

В отечественной практике также часто рассматривался тематический индекс цитирования, рассчитываемый по специальной формуле компанией «Яндекс», однако в 2018 г. этот индекс перестал рассчитываться и отображаться.

Наличие независимых источников оценки пользовательского трафика сайтов, старейшим из которых является Alexa Internet, начавший работу в 1996 г., позволяет добавить к числу вебометрических индикаторов данные, характеризующие посещаемость сайтов. К сожалению, Alexa оценивает трафик только для доменов второго уровня. Эту проблему решает появившийся в 2007 г. и активно развивающийся сервис SimilarWeb (https://similarweb.com). С помощью сервиса SimilarWeb можно получить оценки количества пользовательских сессий в месяц, среднего количества просмотренных страниц за один визит (глубина просмотра), среднего времени, проведенного на сайте, долей поискового трафика, трафика из социальных сетей и трафика по ссылкам на других сайтах.

Объект измерения

Объектом измерения для большей части проводившихся исследований является домен (совокупность сайтов, размещенных в пространстве отдельного DNS домена). В ряде случаев часть измерений может быть проведена относительно конкретного сайта или даже его части. Однако в полном объеме значения всех указанных индикаторов могут быть измерены только относительно отдельного домена.

Исследователи неоднозначно относятся к тому, что веб-ресурсы организации могут быть размещены в разных доменах. В оригинальном рейтинге авторы придерживаются убеждения, что хорошей практикой является размещение всех ресурсов в одном доменном пространстве, и такую практику следует поощрять, присутствует также точка зрения, что необходимо учитывать всю совокупность ресурсов организации. Однако это не влияет на выбор объекта измерения, им все равно остается отдельный DNS домен.

Недостатки имеющихся проектов

Из упомянутых выше вебометрических проектов только рейтинг webometrics.info имел широкий охват и аккумулировал данные по значительной части мировых научно-исследовательских организаций, научных онлайн-архивов и высших учебных заведений. Сегодня в активном состоянии остался только рейтинг вузов. Все остальные проекты имели национальный или еще более узкий охват и существовали относительно недолго, а некоторые проекты ограничились публикацией результатов только одного цикла сбора и анализа данных. Рейтинги, созданные Институтом вычислительных технологий Сибирского отделения Российской академии наук (СО РАН) и Дальневосточным геологическим институтом Дальневосточного отделения РАН, обновлялись в течение нескольких лет, однако и они прекратили свое существование. Ни один из известных проектов не публиковал исходные собранные данные без коррекций и/или агрегации, что делает практически невозможной независимую проверку результатов. Отсутствие в публичном доступе первичных данных и низкая частота сбора данных не позволяют также анализировать изменение значений индикаторов отдельных доменов в динамике и общие тенденции развития академического веб-пространства. И наконец, в опубликованных рейтингах не используются показатели, связанные с посещаемостью сайтов, что значительно обедняет возможности анализа.

В связи с вышеизложенным представляется актуальной задача создания широкого вебометрического проекта, выполняющего сбор значений индикаторов по регулярному, желательно ежемесячному графику. С начала 2015 г. на инициативной основе выполняется проект по сбору вебометрических показателей российских академических институтов, позже расширенный на вузы и другие научные организации. В 2017 г. в экспериментальном порядке охват был расширен на все страны постсоветского пространства, но эти результаты недоступны онлайн. Также в закрытом режиме выполняется сбор вебометрических данных сайтов библиотек России и крупнейших мировых библиотек. Сайт проекта расположен по адресу http://www.webometrix.ru. Опыт, полученный при реализации проекта, позволяет нам обозначить и разобрать проблемы, возникающие при решении задачи регулярного сбора вебометрических данных доменов большого количества исследовательских институтов, учреждений высшего образования и других типов организаций.

Сбор базы данных доменов организаций

Прежде всего должна быть сформирована БД доменов организаций, что может оказаться нетривиальной задачей. Пример Российской Федерации показывает, что только относительно недавно Рособрнадзор сформировал и опубликовал в общем доступе БД высших учебных заведений, которую можно использовать в качестве основы. Даже наличие такого реестра требует дополнительных усилий по поиску официальных веб-сайтов вузов, что дополнительно затрудняется наличием у некоторых вузов развитой филиальной сети, зачастую с собственными официальными сайтами, размещенными в доменах, отличных от основной организации. Единый же реестр научно-исследовательских организаций отсутствует. Для других стран проблема лишь усугубляется.

Представляется разумным для формирования реестра организаций опираться прежде всего на библиометрические БД, такие как eLibrary,

Scopus, Web of Science. Можно предположить, что нас прежде всего интересуют организации, сотрудники которых публикуются в научных журналах и других изданиях. Соответственно, по аффилиациям авторов публикаций может быть составлена и обновляться БД организаций, имеющих отношение к научным исследованиям.

Адреса веб-сайтов организаций и соответствующие им домены могут быть уточнены с использованием поисковых систем общего назначения, другие данные организации могут быть найдены на веб-сайте и уточнены с использованием различных информационно-справочных систем. Так, для российских организаций мы интенсивно использовали сервис «Контур Фокус» для уточнения текущего статуса организации.

Контроль актуальности базы данных доменов

К сожалению, сформированная БД организаций нуждается в постоянном уточнении. Достаточно часто происходит ликвидация отдельных организаций, их слияние, переименование, сопровождающиеся ликвидацией сайта или переносом сайта на другой домен, что также может происходить и по другим причинам.

Беглый анализ БД российских организаций, представленных в рейтинге webometrics.info, проведенный в сентябре 2016 г., показал, что из 11 998 доменов высших учебных заведений 78 на тот момент были ликвидированы полностью, 48 не отвечали на запросы, для 78 происходило перенаправление на другое доменное имя и как минимум для двух доменов закончился срок регистрации. Последнее число, скорее всего, больше, так как обычной практикой киберсквоттеров, захватывающих популярные доменные имена при их освобождении, является сохранение элементов заголовка страницы и ее описания, в связи с этим может быть затруднено автоматизированное определение таких доменов. Таким образом, около 2% доменов (а возможно, и больше) в БД этого рейтинга нуждались в уточнении. Соответствующие цифры для исследовательских центров еще хуже – из 7 209 доменов 101 был удален, 97 не отвечали, 183 были перенаправлены, у не менее двух закончился срок регистрации, всего более 5% нуждались в актуализации. В этих цифрах не учтены отсутствующие в рейтинге организации, хотя только в РФ в рейтинге webometrics.info не были учтены более 80% академических институтов.

Актуальность БД может быть обеспечена автоматической проверкой доступности сайта организации, проверкой отсутствия переадресации, сверкой редко изменяющихся значений заголовка страницы, описания, другой имеющейся метаинформации, для последующего визуального контроля может быть создан снимок окна браузера. В зависимости от имеющихся ресурсов такая проверка может выполняться с разной частотой. Также индикатором изменения ситуации может служить резкое падение размера домена. Домены, попавшие под подозрение по приведенным причинам, должны быть отсмотрены в ручном режиме с принятием соответствующего решения. Пополнение БД может выполняться на основе мониторинга библиометрических БД и доступных реестров.

Устойчивость показателей

J. Bar-Ilan с соавторами исследовали устойчивость оценки количества результатов (хитов) при ответе на запрос в различных поисковых системах [13]. В качестве основных проблем ими отмечены некорректная оценка количества хитов и нестабильность этого значения с течением времени. В работе C. I. Font-Julian и др. отмечается противоречивость результатов, выдаваемых поисковыми системами – количество результатов может варьироваться в зависимости от страницы, оценка количества хитов на первой и последней страницах результатов поиска может заметно различаться [14]. Детальный анализ устойчивости показаний поисковых систем и взаимосвязи оценки количества результатов с внутренними технологическими изменениями был приведен в статье A. Van den Bosch и др. [15].

В полученных нами данных также можно увидеть вариативность значений, которую невозможно объяснить изменениями самого объекта измерения. В качестве примера корректных показателей можно рассмотреть график изменения значений индикаторов размеров страниц в домене ipgg.sbras.ru, принадлежащем Институту нефтегазовой геологии и геофизики (ИНГГ) СО РАН и контролируемому в указанный период автором данной статьи (рис. 1).

С января 2015 г. в домене не происходило существенных изменений за исключением регулярного пополнения материалов на официальном сайте и добавления ресурсов, связанных с проведенными конференциями. Этот тренд виден на показаниях по поисковой системе Google. Но, несмотря на устойчивый и плавный рост фактического количества страниц в домене, в данных измерений разных поисковых систем наблюдаются разного вида и амплитуды искажения.

Достоверность показателей

Под сомнение также ставилась достоверность результатов измерения в связи с получением различных результатов измерений в зависимости от использованной поисковой системы,

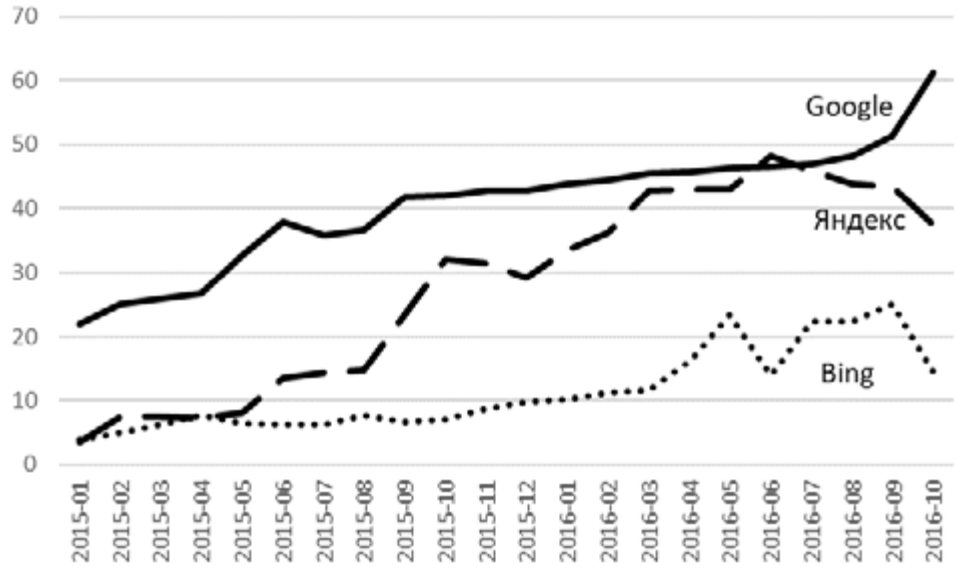

Рис. 1. Динамика размера домена ipgg.sbras.ru (ИНГГ СО РАН), по данным различных поисковых систем (тыс. стр.)

предположительно связанных с алгоритмами оценки количества результатов [16]. Некоторыми авторами предлагались различные подходы к решению проблемы достоверности с помощью использования дополнительных техник типа разделения запросов [17] или использования веб-краулера для уточнения результатов [18]. Однако эксперименты, проведенные над находящимися под управлением авторов сайтами, показывают, что оценки количества результатов, получаемые такими методами, часто оказываются заметно ниже, чем количество проиндексированных страниц, указанное в интерфейсе веб-мастера соответствующей поисковой системы, и реальное количество страниц на сайте. Оценка, получаемая с помощью веб-сервиса или веб-интерфейса поисковой системы, обычно несколько превышает реальное количество имеющихся и проиндексированных страниц.

Результаты анализа массива, накопленного за весь период последовательного ежемесячного сбора данных, показывают ряд повторяющихся паттернов «искажений». Эти искажения могут быть связаны как с особенностями работы модулей индексирования и расчета количества результатов при поиске в самих поисковых системах, так и с изменениями объекта измерения – удалением или появлением сайта или раздела, изменением формата страниц, улучшением метаданных, что может заметно отразиться на качестве индексации сайта поисковыми

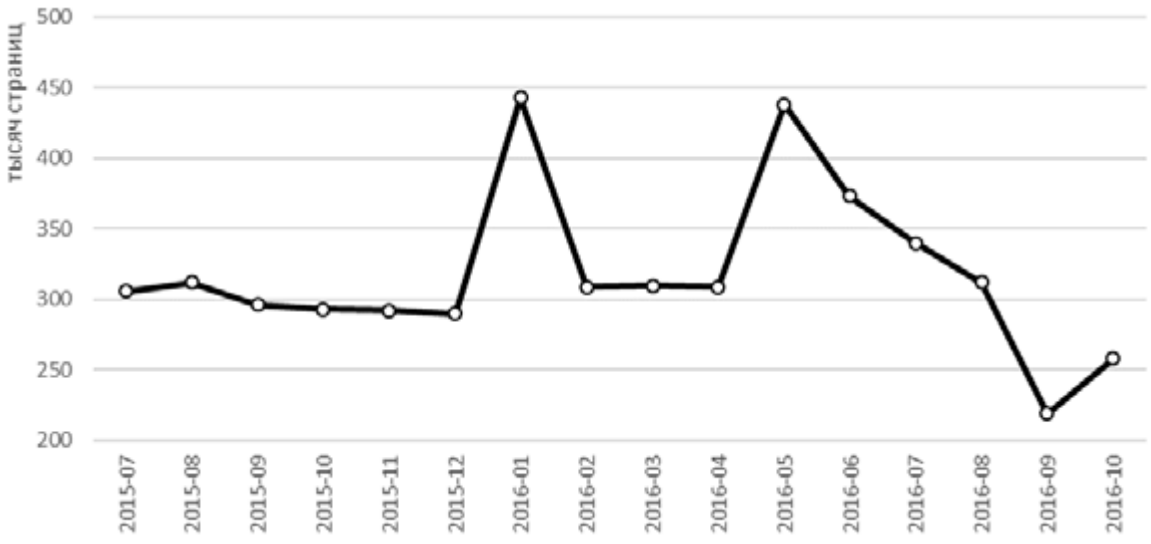

Рис. 2. Динамика размера домена unn.ru (Университет Лобачевского), по данным веб-интерфейса Google (тыс. стр.)

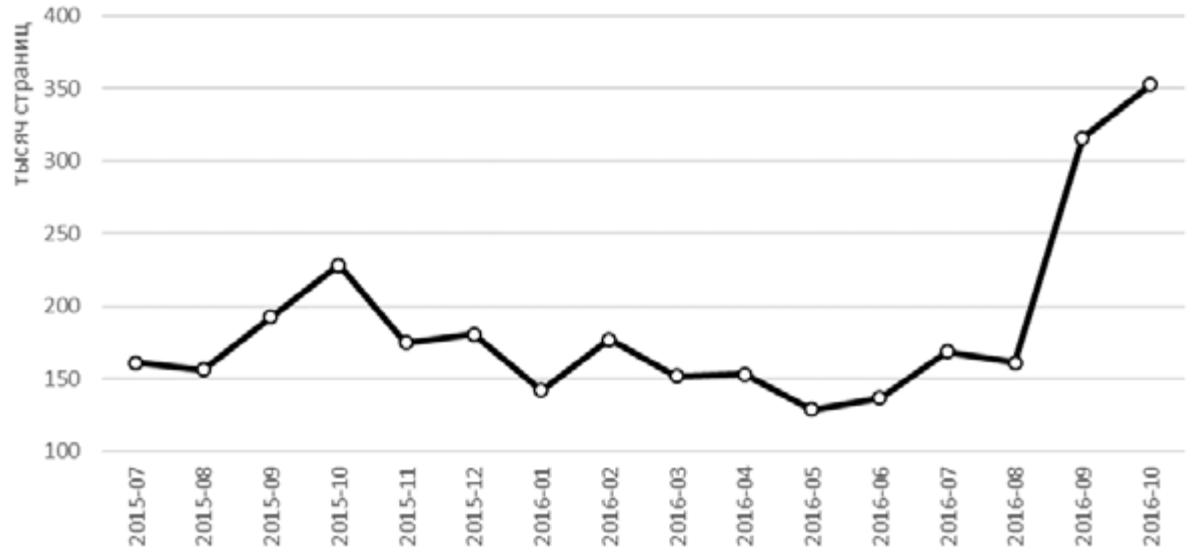

Рис. 3. Динамика размера домена nstu.ru (Новосибирский государственный технический университет), по данным веб-сервиса поиска Google (тыс. стр.)

системами и, соответственно, количестве проиндексированных страниц.

Рассмотрим некоторые характерные паттерны искажений. Достаточно часто попадаются одиночные «выбросы» (рис. 2). В конце декабря 2015 г. – начале января 2016 г. в Национальном исследовательском Нижегородском государственном университете им. Н. И. Лобачевского (Университет Лобачевского) запустили новую версию сайта. Резкий выброс вверх с 290 до 443 тыс. страниц в январе 2016 г. может быть обусловлен тем, что в индексе Google присутствовали и новая, и старая версия сайта. Начиная с апреля 2016 г. мы видим другую динамику: резкое повышение с последующим снижением, причем снижение может быть достаточно плавным. Такая динамика часто бывает связана с публикацией нового ресурса или заменой старого, когда какое-то время в индексе поисковой системы присутствуют две версии ресурса, но старая версия постепенно вычищается из индекса вследствие недоступности ресурсов по сохраненным в поисковом индексе путям. «Всплеск» может быть откорректирован для улучшения качества данных как имеющий технический, а не содержательный характер. В дальнейшем мы также видим ряд всплесков, но общая динамика может быть достаточно стабильной в течение довольно протяженных периодов.

И выбросы, и всплески могут быть также обусловлены ошибками индексации сайта поисковой системой, например, избыточной индексацией большого количества страниц, доступных по адресам с использованием параметров URL. Зачастую такая избыточная индексация корректируется внутренними алгоритмами поисковых систем.

Другим часто повторяющимся паттерном является «гребенка», вероятно связанная с особенностями работы алгоритмов оценки количества ответов поисковых систем или с территориальным распределением дата-центров, обрабатывающих запросы. Характерный пример гребенки показан на рисунке 3. Мы можем наблюдать колебания значений вокруг плавно снижающегося среднего значения с октября 2010 г. по июнь 2016 г. Для выявления общей динамики этот график может быть сглажен одним из распространенных алгоритмов, например, скользящим средним.

Наконец, последний характерный паттерн представлен на рис. 4 и представляет собой «горб». Интерпретация таких паттернов может быть неоднозначной, в некоторых случаях такой горб или, наоборот, провал хорошо распознаются, так как остальная кривая значений носит достаточно монотонный характер, но в большей части случаев ситуация не так однозначна. Нам представляется, что применение алгоритмов сглаживания к таким участкам позволяет лучше проявить общие тренды развития веб-домена.

Широкий спектр алгоритмов сглаживания данных хорошо справляется с большей частью искажений. На рисунке 5 к данным поисковых систем применен алгоритм скользящего среднего.

Необходимо отметить, что в большей или меньшей степени неустойчивость значений во времени характерна практически для всех измеряемых индикаторов, хоть и по разным причинам (рис. 6). Очевидно, что измерения, проводимые единовременно или достаточно редко (раз в год или полгода), могут давать сильно искаженные данные. При частом же сборе данных у нас появляется возможность лучше оценить их качество, выполнить

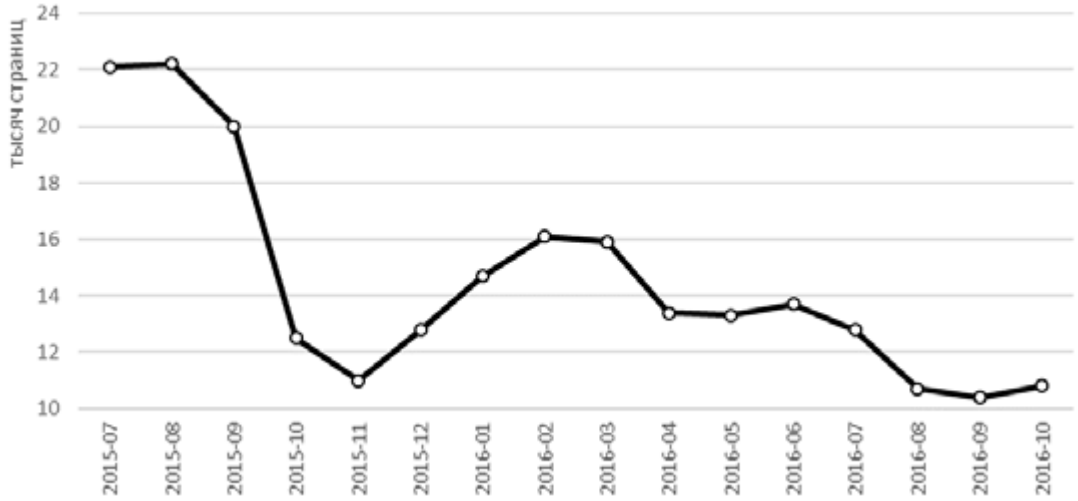

Рис. 4. Динамика размера домена dstu.ru (Дагестанский государственный технический университет), по данным веб-сервиса поиска Google (тыс. стр.)

коррекцию с использованием фильтрации и/или сглаживания, что не приведет к значительному изменению данных, но при этом эффективно очистит их от искажений и позволит выделять заметные тренды на длительном промежутке измерений.

Кроме того, наличие данных, измеряемых регулярно, позволяет добавить дополнительные производные индикаторы, связанные со скоростью и характером изменений значений индикаторов во времени.

Имеются также сложности, связанные с несоответствием ожиданиям самих объектов измерения, например, наличием на веб-сайтах домена научного института контента, не имеющего отношения к его области исследований и т. д. Часть этих вопросов была нами разобрана на примере сайтов академических институтов в работе [19]. К сожалению, решение этих проблем невозможно без полного анализа контента сайтов, что представляется крайне трудоемкой задачей.

Масштаб сбора данных

Задача сбора вебометрических данных сайтов научно-исследовательских организаций и высших учебных заведений в ежемесячном режиме в мировом масштабе может оказаться

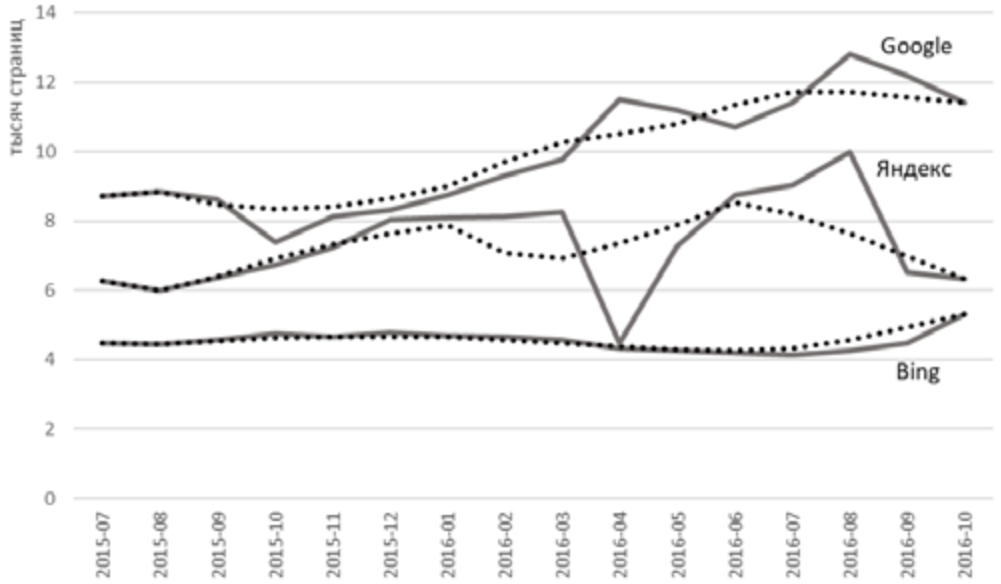

Рис. 5. Динамика размера домена spbau.ru (Санкт-Петербургский национальный исследовательский академический университет им. Ж. И. Алфёрова РАН) по данным трех поисковых систем: измеренные (сплошная линия) и сглаженные (пунктирная линия) значения

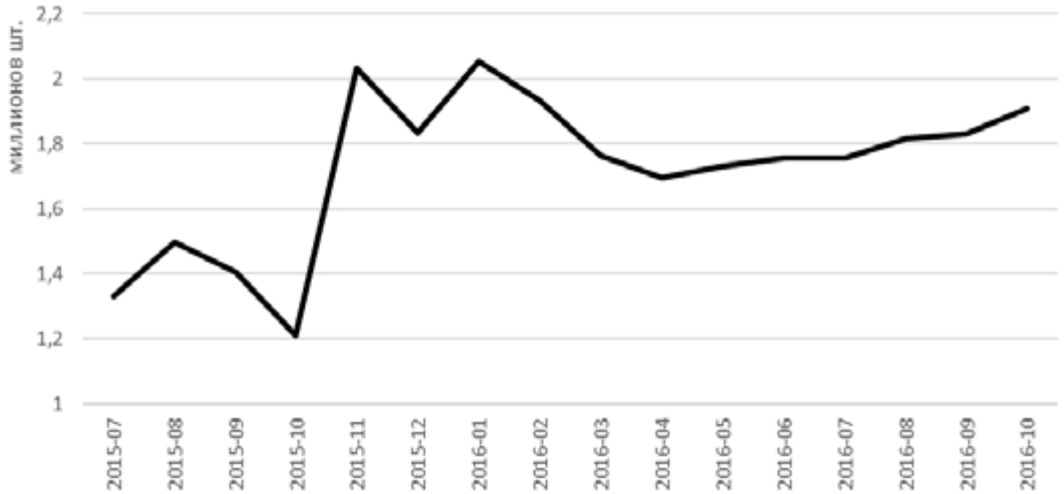

Рис. 6. Динамика количества внешних ссылок на домен ifmo.ru (Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики – Университет ИТМО), по данным сервиса Ahrefs (млн ссылок)

весьма затратной. Попробуем выполнить оценки потенциального объема собираемых данных и затрат времени на один цикл сбора. Рейтинг webometrics.info в период наибольшего охвата содержал данные по 11 998 высшим учебным заведениям (июль 2016), 2 275 репозиториям (июль 2016), 11 999 клиникам (январь 2015), 1 268 школам бизнеса (прерван в июле 2013 г.) и 7 209 исследовательским центрам (январь 2016). Таким образом, общий объем собираемых данных превышал 30 тыс. доменов. С учетом неполного списка исследовательских центров можно считать, что полный объем скорее будет близок к 40 тыс. доменов или даже превышать это значение в случае сбора данных по дополнительным доменам организаций. С учетом измерения значений от 4 до 20 индикаторов для каждого домена необходимо ежемесячно собирать до 1 миллиона значений. Распространенные методы получения значений вебометрических индикаторов практически не позволяют эффективно решить такую задачу, так как процесс измерения связан с использованием веб-сервисов, которые могут ограничивать количество обращений в единицу времени с одного IP-адреса, и загрузкой и разбором веб-страниц, что само по себе может быть связано с задержками обработки http-запросов, кроме того, запрашиваемые веб-серверы могут быть оборудованы одним из вариантов защиты от атак и массового использования роботов. Стандартной практикой при вебометрических измерениях является искусственное ограничение количества запросов к сервису в единицу времени с помощью временных задержек, что ограничивает скорость сбора данных.

Архитектура системы сбора показателей

Необходимая для ежемесячного сбора до 1 миллиона значений вебометрических индикаторов производительность может быть достигнута построением распределенной системы, позволяющей повысить производительность кратно количеству независимых распределенных по разным IP-сетям агентов. Разработанная система состоит из БД, размещенной на отдельном сервере под управлением СУБД MongoDB, веб-сервере / сервере приложений, построенном на базе Node.js и обеспечивающем диспетчеризацию заданий и обработку собранных данных, а также агентов, выполняющих сбор данных по получаемым от сервера приложений заданиям.

Агенты написаны на языке сценариев PowerShell и используют для получения данных как встроенные в PowerShell средства выполнения HTTP запросов, так и управление браузером Internet Explorer по технологии OLE Automation. Взаимодействие между клиентами и сервером выполняется посредством вызовов RESTful API. Весь процесс контролируется через веб-интерфейс администратором, выполняющим общие задачи подготовки и завершения ежемесячного цикла сбора данных, и операторами, контролирующими работу агентов.

Весь цикл сбора состоит из следующих этапов (рис. 7):

- Администратор системы через веб-интерфейс инициирует подготовку базы заданий, формирующейся из списка исследуемых доменов и набора измеряемых индикаторов.
- Оператор через веб-интерфейс получает генерируемый автоматически уникальный ключ приложения.
- Оператор на рабочей станции выполняет запуск агента сбора данных с ключом приложения, полученным на предыдущем этапе.
- Агент с ключом приложения и списком индикаторов, которые он может собирать, обращается к серверу за сессионным ключом.

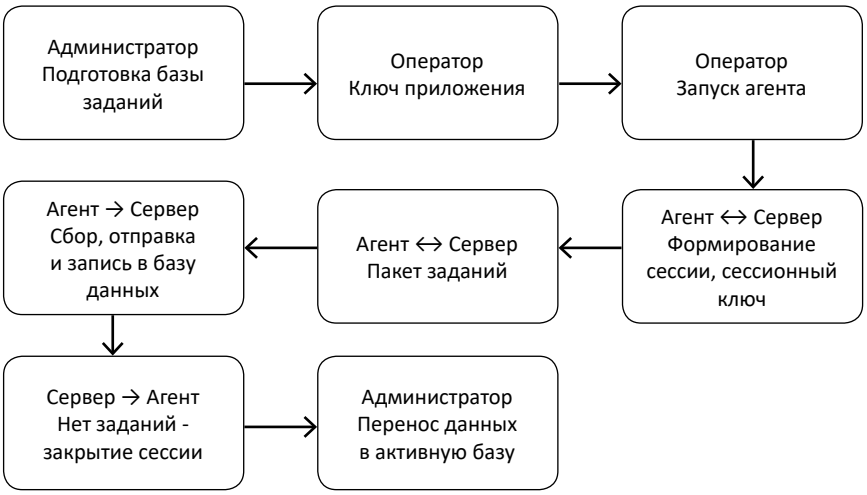

Рис. 7. Схема процесса ежемесячного сбора данных

- Сервер формирует сессию, генерирует сессионный ключ и возвращает его агенту.
- Агент с сессионным ключом обращается за пакетом заданий определенного размера.
- Сервер формирует пакет заданий, блокирует соответствующие записи в базе заданий и отправляет пакет агенту.
- Агент обрабатывает пакет заданий, отправляя полученные данные серверу.
- Сервер получает данные и записывает их в базу заданий.
- При исчерпании очереди заданий сессия закрывается.
- По окончании сбора данных администратор через веб-интерфейс инициирует перенос данных в общую БД, и они становятся доступны на сайте проекта.

Администратор и операторы могут наблюдать в веб-интерфейсе ход выполнения заданий, видеть очередь, принудительно закрывать сессии агентов, а также визуально контролировать данные, переданные агентами, и, при необходимости, возвращать отдельные задания в очередь ожидания. В связи с особенностями сбора отдельных индикаторов были реализованы разные агенты для разных источников данных (табл. 1). Количество операторов и активных агентов не ограничено, что позволяет масштабировать систему в соответствии с потребностями.

БД обеспечивает хранение в двух основных коллекциях информации об исследуемых организациях и принадлежащих им доменах и результатов сбора данных, включая дату и время получения каждого

Таблица 1

Источники данных, собираемые вебометрические индикаторы и методы сбора

Источник данных	Индикаторы	Метод сбора значений
Google	Размер сайта, количество документов	Веб-интерфейс под визуальным контролем и во взаимодействии с оператором
Bing	Размер сайта, количество документов	Программный интерфейс (API)
Яндекс	Размер сайта, количество документов, тИЦ – значение и ранг	Программный интерфейс (API)
Ahrefs, Majesstic SEO	Количество внешних ссылок, ссылающихся доменов, ссылающихся IP, ссылающихся IP подсетей	Веб-интерфейс во взаимодействии с оператором, обработка выгруженных в формате csv данных
Google Scholar	Количество публикаций (файлов или карточек с полными метаданными) на сайтах	Веб-интерфейс под визуальным контролем и во взаимодействии с оператором
SimilarWeb	Количество визитов в месяц, среднее время на сайте, средняя глубина просмотра, количество отказов, доли поиска, социальных сетей и внешних ссылок во входящем трафике	Веб-интерфейс под визуальным контролем оператора

Таблица 2

Временные затраты при сборе данных для 3853 доменов

Агент № п/п.	Источники	Количество индикаторов	Общее время (часов:минут)	Среднее время на один домен (мин:сек)
1	Google, Bing, Яндекс	6	144:14	2:15
2	Google Scholar	1	297:13	4:38
3	Ahrefs	4	0:14	менее 1 с.
4	Majestic SEO	4	0:17	менее 1 с.
5	SimilarWeb	7	147:14	2:18
6	Яндекс тИЦ	2	10	0:10

измеряемого значения. Служебные коллекции также содержат: данные пользователей (администратора и операторов) системы, необходимые для их аутентификации и разграничения прав доступа; список сессий (сеанс работы отдельного агента) с указанием контролирующего его оператора, времени начала и окончания сбора данных, набора измеряемых индикаторов; задания на обработку, из которых формируются пакеты по запросу агентов и производится блокирование «занятых» заданий.

Оценка производительности

Результаты оценки производительности на примере реального сбора данных с учетом задержек различного характера показывают следующие временные затраты для каждого из агентов при сборе данных для 3853 доменов (табл. 2).

Таким образом, на сбор данных по одному домену уходит ориентировочно 9 мин 20 сек. При оценке общего количества доменов для сбора данных по всем странам мира в 40 тыс. общее время сбора составит 6222 часа, или чуть меньше 260 суток. Чтобы уложиться в месячный цикл, параллельно должно работать не менее 10 агентов. Необходимо отметить, что один оператор может параллельно контролировать работу нескольких агентов.

Актуализация списка организаций

Подсистема редактирования данных организаций (рис. 8) позволяет редактировать основные данные на двух языках (национальном и английском), а также проверять состояние домена переходом на веб-сайт организации для визуального контроля (кнопка справа

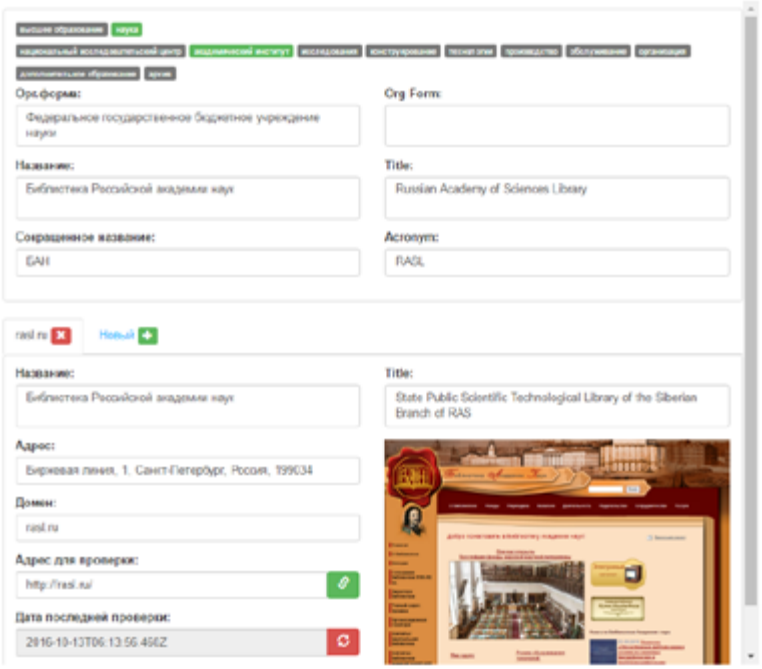

Рис. 8. Элемент интерфейса редактирования карточки организации с возможностью добавления и удаления доменов

от поля «Адрес для проверки») и автоматически, обработкой обращения к веб-сайту, выполняемого на стороне сервера с проверкой возможной автоматической переадресации, сверкой заголовка страницы с сохраненной в базе, созданием копии окна браузера (кнопка справа от поля «Дата последней проверки»).

Заключение

Сочетание постоянной изменчивости веб-пространства, являющейся, вероятно, его фундаментальным свойством, и специфики применяемых измерительных инструментов приводит к недостаточно достоверным данным вебометрических измерений. Однако повышение частоты измерений позволяет значительно улучшить качество данных и добавляет возможность анализа динамики изменения значений отдельных индикаторов. Достаточно частый сбор данных для большого числа доменов исследовательских организаций и вузов при широкомасштабных вебометрических исследованиях, охватывающих большое количество стран, возможен только в распределенной среде.

Предложенная схема сбора данных, реализованная в виде программной системы с клиент-серверной архитектурой, может обеспечить ежемесячный сбор вебометрических данных в масштабах, сравнимых с проектом webometrics.info, с приемлемыми трудозатратами. Улучшенные за счет сглаживания и фильтрации временных рядов данные могут быть основой не только для рейтинга организаций, но и для исследований динамики и направлений изменения научного и академического веб-пространства и его отдельных сегментов [20].

Список источников

1. *Björneborn L., Ingwersen P.* Toward a basic framework for webometrics // Journal of the American Society for Information Science and Technology. 2004. V. 55, N14. P. 1216–1227. DOI: 10.1002/asi.20077.

2. *Faba-Pérez C. и др.* Comparative analysis of webometric measurements in thematic environments // Journal of the American Society for Information Science and Technology. 2005. V. 56, N8. P. 779–785. DOI: 10.1002/asi.20161.

3. *Payne N., Thelwall M.* A longitudinal study of academic webs: Growth and stabilisation // Scientometrics. Springer Netherlands, 2007. V. 71, N3. P. 523–539. DOI: 10.1007/s11192-007-1695-y.

4. *Aguillo I. F., Granadino B., Ortega J. L., Prieto J.* A Scientific research activity and communication measured with cybermetrics indicators // Journal of the American Society for Information Science and Technology. 2006. V 57, N10. P. 1296–1302. DOI:10.1080/03797720802254031.

5. *Aguillo I. F., Ortega J. L., Fernández M.* Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments // Higher Education in Europe. 2008. V. 33, N2–3. P. 233–244. DOI: 10.1080/03797720802254031.

6. *Vargas-Quesada B. и др.* Web structure and influence of the Arab universities of the MENA zone (Middle East and North Africa): Visualization and analysis. 2013. V. 65, N6. P. 623–643. DOI:10.1108/AP-10-2012-0082.

7. *Tafaroji R. и др.* Webometric analysis of Iranian medical universities according to visibility, size and rich files // Webology. University of Tehran, 2014. V. 11, N1.

8. *Ma F., Qiu J.* Study on web performance of Chinese universities // Wuhan Daxue Xuebao (Xinxi Kexue Ban)/ Geomatics and Information Science of Wuhan University. 2010. V. 35, N SPECIAL ISSUE2. P. 146–151.

9. *Шокин Ю. И., Клименко О. А., Рычкова Е. В., Шабальников И. В.* Рейтинг сайтов научных организаций СО РАН // Вычислительные технологии. 2008. Т. 13, № 3. С. 128–135.

10. *Антопольский А., Поляк, Ю., Усанов В.* Развитие вебометрического индекса научно-образовательных учреждений России // Информационные ресурсы России. 2013. Т. 4. С. 16–24.

11. *Ханчук А. И., Наумова В. В.* Информационное пространство Дальневосточного отделения РАН // Вестник Дальневосточного отделения Российской академии наук. 2009. № 4. С. 122–129.

12. *Jati H., Dominic D. D.* A New Approach of Indonesian University Webometrics Ranking Using Entropy and PROMETHEE II // Procedia Computer Science. Elsevier B. V., 2017. V. 124. P. 444–451. DOI: 10.1016/j.procs.2017.12.176

13. *Bar-Ilan J., Levene M., Mat-Hassan M.* Dynamics of search engine rankings - A case study // CEUR Workshop Proceedings. 2004. V. 703. P. 3–13.

14. *Font-Julian C. I., Ontalba-Ruipérez J. A., Orduña-Malea E.* Hit count estimate variability for website-specific queries in search engines: The case for rare disease association websites // Aslib Journal of Information Management. Emerald Group Publishing Ltd., 2018. V. 70, N2. P. 192–213. DOI:10.1108/AJIM-10-2017-0226.

15. *Van den Bosch A., Bogers T., de Kunder M.* Estimating search engine index size variability: a 9-year longitudinal study // Scientometrics. Springer Netherlands, 2016. N April. DOI:10.1007/s11192-016-1863-z .

16. *Uyar A.* Investigation of the accuracy of search engine hit counts // Journal of Information Science. 2009. V. 35, N4. P. 469–480. DOI:10.1177/0165551509103598.

17. *Thelwall M.* Extracting accurate and complete results from search engines: Case study windows live // Journal of the American Society for Information Science and Technology. 2008. V. 59, N1. P. 38–50. DOI: 10.1002/asi.20704.

18. *Печников А. А.* О вебометрическом индикаторе «размер сайта» // Обозрение прикладной и промышленной математики. 2013. Т. 20, № 4. С. 568.

19. *Косяков Д. В., Гуськов А. Е., Быховцев Е. С.* Академические институты России в зеркале вебометрики

// Вестник РАН. 2016. Т. 86, № 11. С. 1015–1025. DOI: 10.7868/S086958731610011X.

20. *Kosyakov D., Guskov A., Bykhovtsev E.* Webometric analysis of Russian scientific and education web // CEUR Workshop Proceedings. 2017. Т. 1839.

References

1. Björneborn L., Ingwersen P. Toward a basic framework for webometrics // Journal of the American Society for Information Science and Technology. 2004. V. 55, N14. P. 1216–1227. DOI: 10.1002/asi.20077.

2. Faba-Pérez P. и др. Comparative analysis of webometric measurements in thematic environments // Journal of the American Society for Information Science and Technology. 2005. V. 56, N8. P. 779–785. DOI:10.1002/asi.20161.

3. Payne N., Thelwall M. A longitudinal study of academic webs: Growth and stabilisation // Scientometrics. Springer Netherlands, 2007. V. 71, N3. P. 523–539. DOI: 10.1007/s11192-007-1695-y.

4. Aguillo I. F., Granadino B., Ortega J. L., Prieto J. A. Scientific research activity and communication measured with cybermetrics indicators // Journal of the American Society for Information Science and Technology. 2006. V. 57, N10. P. 1296–1302. DOI: 10.1080/03797720802254031.

5. Aguillo I. F., Ortega J. L., Fernández M. Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments // Higher Education in Europe. 2008. V. 33, N2–3. P. 233–244. DOI: 10.1080/03797720802254031.

6. Vargas-Quesada B. и др. Web structure and influence of the Arab universities of the MENA zone (Middle East and North Africa): Visualization and analysis. 2013. V. 65, N6. P. 623–643. DOI: 10.1108/AP-10-2012-0082.

7. Tafaroji R. и др. Webometric analysis of Iranian medical universities according to visibility, size and rich files // Webology. University of Tehran, 2014. V. 11, N1.

8. Ma F., Qiu J. Study on web performance of Chinese universities // Wuhan Daxue Xuebao (Xinxi Kexue Ban)/ Geomatics and Information Science of Wuhan University. 2010. V. 35, N SPECIAL ISSUE2. P. 146–151.

9. Shokin Yu.I., Klimenko O. A., Rychkova E. V., Shabalnikov I. V. Рейтинг сайтов научных организаций СО РАН // Вычислительные технологии. 2008. V. 13, N3. P. 128–135.

10. Antopolsky A., Polyak Yu., Usanov V. Development of a vebometric index of scientific and educational institutions of Russia // Information resources of Russia. 2013. V. 4. P. 16–24.

11. Khanchuk A. I., Naumova V. V. Information Space of the Far East Branch of the Russian Academy of Sciences. // Herald of the FEB RAS. 2009. N4. P. 122–129.

12. Jati H., Dominic D. D. A New Approach of Indonesian University Webometrics Ranking Using Entropy and PROMETHEE II // Procedia Computer Science. Elsevier B. V., 2017. V. 124. P. 444–451. DOI: 10.1016/j.procs.2017.12.176.

13. Bar-Ilan J., Levene M., Mat-Hassan M. Dynamics of search engine rankings - A case study // CEUR Workshop Proceedings. 2004. V. 703. P. 3–13.

14. Font-Julian P. I., Ontalba-Ruipérez J. A., Orduña-Malea E. Hit count estimate variability for website-specific queries in search engines: The case for rare disease association websites // Aslib Journal of Information Management. Emerald Group Publishing Ltd., 2018. V. 70, N2. P. 192–213. DOI: 10.1108/AJIM-10-2017-0226.

15. Van den Bosch A., Bogers T., de Kunder M. Estimating search engine index size variability: a 9-year longitudinal study // Scientometrics. Springer Netherlands, 2016. N April. DOI: 10.1007/s11192-016-1863-z.

16. Uyar A. Investigation of the accuracy of search engine hit counts // Journal of Information Science. 2009. V. 35, N4. P. 469–480. DOI: 10.1177/0165551509103598.

17. Thelwall M. Extracting accurate and complete results from search engines: Case study windows live // Journal of the American Society for Information Science and Technology. 2008. V. 59, N1. P. 38–50. DOI:10.1002/asi.20704.

18. Pechnikov A. A. About the webometric indicator «size of a website» // Applied and Industrial Mathematics Review. 2013. V. 20, N4. P. 568.

19. Kosyakov D. V., Guskov A. E., Bykhovtsev E. S. Russia's academic institutes as mirrored by webometrics // Herald of the Russian Academy of Sciences. 2016. V. 86, N6. P. 490–499. DOI:10.1134/S1019331616050063.

20. Kosyakov D., Guskov A., Bykhovtsev E. Webometric analysis of Russian scientific and education web // CEUR Workshop Proceedings. 2017. V. 1839.

Статья поступила в редакцию 24.01.2020
Получена после доработки 03.02.2020
Принята для публикации 09.02.2020

Received 24.01.2020
Revised 03.02.2020
Accepted 09.02.2020