УДК 004.7

А.И.ХАНЧУК, В.В.НАУМОВА

Информационное пространство Дальневосточного отделения РАН

Описываются результаты анализа информационного пространства Дальневосточного отделения РАН методами вебометрики. На примере сайтов ДВО РАН установлены свойства и характеристики научного сайта, определяющие его высокий рейтинг в Интернете.

Ключевые слова: информатика, вебометрика, научный сайт, Дальневосточное отделение РАН.

Information Space of the Far East Branch of the Russian Academy of Sciences. A.I.KHANCHUK, V.V.NAUMOVA (Far East Geological Institute, FEB RAS, Vladivostok).

This article represents results of analysis of the Far East Branch of RAS Information space by methods of Web Metrics. The main point of this analysis for us is to understand what characteristics and peculiarities a science web-site should possess to be highly rated in epy Internet. To reveal such special features science web-sites of FEB RAS have been taken as an example and analyzed.

Key words: informatics, Web Metrics, a science web-site, Far East Branch of RAS.

Важнейшей задачей Дальневосточного отделения РАН на сегодняшнем этапе является разработка и развитие высокорейтинговых научных интернет-ресурсов.

Работа выполнена методами вебометрики – раздела информатики, в котором исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к World Wide Web [5, 6]. Вебометрика включает 4 основных направления: www-индикаторы сайтов (индексы цитирования, размеры, «научность», «видимость», ранжирование и др.), поиск и сбор данных в www (информационный поиск, каталоги, поисковые машины, роботы, краулеры, «черви», «пауки» и др.), социальные феномены в www (социальные сети, сообщества сайтов, форумы, самоорганизация и др.), анализ гиперссылок (связи между сайтами, мотивация ссылок, структура фрагментов www и др.) [2, 3].

Нами проведены исследования в Дальневосточном научном сегменте глобальной сети. Доступ к информации Сети осуществлялся главным образом через поисковые машины – ключевые инструменты измерения видимости и воздействия на аудиторию Интернета. Измерения с помощью поисковых машин используются многими исследователями WWW. В настоящее время используют следующие поисковые машины: Google, Yahoo, Yahoo Search, Live Search, Exalead и Google Scholar. В России в этот список добавляется Yandex, Rambler и т.д.

Для проведения измерений поисковые машины имеют соответствующие сервисы. Значения индикаторов существенно изменяются в зависимости от даты измерений.

Современную web-метрическую методологию ранжирования сайтов научных организаций мира представляют результаты исследований киберметрической лаборатории Центра научной информации и документации Национального исследовательского совета Испании [5, 6]. Результаты ранжирования доступны на сайте http://research.webometrics.info.

ХАНЧУК Александр Иванович – академик, первый заместитель председателя ДВО РАН, НАУМОВА Вера Викторовна – доктор геолого-минералогических наук, заведующая лабораторией (Дальневосточный геологический институт ДВО РАН, Владивосток). E-mail: naumova@fegi.ru

Что измеряют: размер сайта (S – size) – общее количество страниц, обнаруживаемых на сайте поисковыми машинами; видимость сайта (V – visibility) – общее количество обнаруживаемых уникальных гипертекстовых ссылок с других веб-ресурсов; количество полнотекстовых файлов (R – rich files) – суммарное количество файлов с расширениями PDF, DOC, PS и PPT; научность сайта (Sc – «scholar») – количество обнаруживаемых Google Scholar ссылок на сайт.

Алгоритм ранжирования научных сайтов, предложенный этой лабораторией:

V – измеряются Yahoo Search, Live Search и Exalead, затем нормируются по каждой поисковой машине и суммируются для каждого сайта, далее ранжируются; QV – место в рейтинге;

S – измеряются Google, Yahoo, Live Search и Exalead, нормируются по каждой поисковой машине, отбрасываются max и min и суммируются, затем ранжируются; QS – место в рейтинге;

R – рассчитываются как S, QR – место в рейтинге;

Sc – рассчитываются с помощью Google Scholar; QSc – место в рейтинге.

Ranking Web of World Research Centers рассчитывается следующим образом:

$$WR = 5 \times QV + 2 \times QS + 1{,}5 \times QR + 1{,}5 \times QSc. \quad (1)$$

В этом рейтинге в число 55 российских сайтов научных организаций из Дальневосточного отделения РАН входит только информационный сервер Дальневосточного геологического института, занимая в мировом рейтинге 1486-е место, в российском – 39-е (табл. 1).

Таблица 1

Российский фрагмент рейтинга сайтов научных организаций мира http://research.webometrics.info/rank_by_country.asp?country=ru

Мировой / российский	Исследовательский центр	Интернет-адрес
80/1	Сибирское отделение РАН	http://www.nsc.ru
120/2	Российская академия наук	http://www.ras.ru
246/3	Государственный институт информационных технологий и телекоммуникаций	http://www.informika.ru
274/4	Объединенный институт ядерных исследований	http://www.jinr.ru
335/5	Институт автоматики и электрометрии СО РАН	http://www.www-sbras.nsc.ru
394/6	Институт космических исследований РАН	http://www.iki.rssi.ru
417/7	Уральское отделение РАН	http://www.uran.ru
467/8	Вычислительный центр им. А.А.Дородницына РАН	http://www.ccas.ru
546/9	Математический институт им. В.А.Стеклова РАН	http://www.mi.ras.ru
564/10	Институт цитологии и генетики СО РАН	http://www.bionet.nsc.ru
1486/39	Дальневосточный геологический институт ДВО РАН	http://www.fegi.ru

Институт вычислительных технологий Сибирского отделения РАН провел ранжирование сайтов СО РАН по сходной методологии, но на основе предложенной ими формулы [4]:

$$W_{Shok} = \log10(V) + \log10(S) + 2 \cdot \log10(R) + 1{,}5 \cdot Sc1, \quad (2)$$

где $V = [V_{Яндекс} + V_{Google} + V_{Yahoo}]/3$, $S = [S_{Яндекс} + S_{Google} + S_{Yahoo}]/3$, $R = [R_{Яндекс} + R_{Google} + R_{Yahoo}]/3$, $Sc1 = [\log10(Sc_{Яндекс}) + \log10(Sc_{Google})]/2$.

Результаты ранжирования доступны на сайте http://www.ict.nsc.ru/ranking.

Вебометрические исследования web-сайтов университетов России проведены на таких индикаторах, как число индексируемых страниц и ссылок на сайт. Предложен критерий ранжирования сайтов по их узнаваемости в Интернете, на его основе ранжированы официальные университетские сайты [2].

Наличие высоких рейтинговых оценок для сайта очень важно с точки зрения его более высокой доступности аудитории пользователей Интернета, поскольку рейтинговые оценки используют практически все поисковые системы. Наличие рейтинговых систем для сайтов, разрабатываемых и применяемых крупными поисковиками (Google, Yandex и др.), при

выдаче результатов поиска пользователям Интернета позволяет быстрее находить наиболее качественную и отвечающую запросу информацию. Результаты запросов сортируются поисковыми машинами и предоставляются в порядке уменьшения их рейтингов. Таким образом, наличие высоких рейтингов ставит сайт в лучшее положение по отношению к другим сайтам.

Анализ связности структур научного интернет-пространства России, оптимизационные модели для размещения ссылок в сообществе научных сайтов представлены в работе [3].

В этом исследовании для нас важным является вопрос о том, какие свойства и характеристики научного сайта определяют его высокие рейтинги в Интернете. Выявлению этих параметров на примере сайтов Дальневосточного отделения РАН и их анализу посвящена эта работа.

Информационное пространство Дальневосточного отделения РАН состоит из разрозненных web-сайтов институтов и организаций, не объединенных в web-структуры. Сайты Отделения практически не имеют между собой перекрестных ссылок.

В анализе информационного пространства ДВО РАН задействованы 83 сайта, которые были найдены с использованием трех (не полных) Каталогов ресурсов ДВО РАН, размещенных на официальном сайте Президиума ДВО РАН http://www.febras.ru и на двух коммуникаторах: Базовой сети ДВО РАН (ИАПУ) http://www.dvo.ru и Информационном сервере Дальневосточного геологического института ДВО РАН http://www.fegi.ru.

Мы называем научным www-коммуникатором научный сайт, который может и не являться официальным сайтом научного учреждения и/или организации РАН, имеет «входящие ссылки с» и/или «исходящие ссылки на» множество официальных и других научных сайтов [2, 3]. К научным www-коммуникаторам относят научные порталы и сайты библиотек, конференций, научных журналов, научных обществ, фондов и т.д.

Анализировались 83 сайта (на доменах 2-го уровня – 23, 3-го и в директориях – 60), найденные с использованием трех (неполных) каталогов ресурсов ДВО РАН, размещенных на официальном сайте Президиума ДВО РАН и на двух коммуникаторах: базовой сети ДВО РАН ИАПУ и информационном сервере ДВГИ ДВО РАН.

Web-сайты институтов и организаций ДВО РАН разрознены, не объединены в web-структуры, практически не имеют перекрестных ссылок, информационное пространство технологически неоднородно. IP-хостинг сайтов осуществляют ИАПУ (Научно-образовательная сеть Владивостока), ИПМТ (Интернет-центр ДВО РАН), ИГиП (Корпоративная сеть ДВО РАН), а также множество внешних интернет-провайдеров (рис. 1).

Наиболее высокая web-активность отмечена в Приморском научном центре (53 сайта). Результаты web-активности научных центров Отделения представлены на рис. 2.

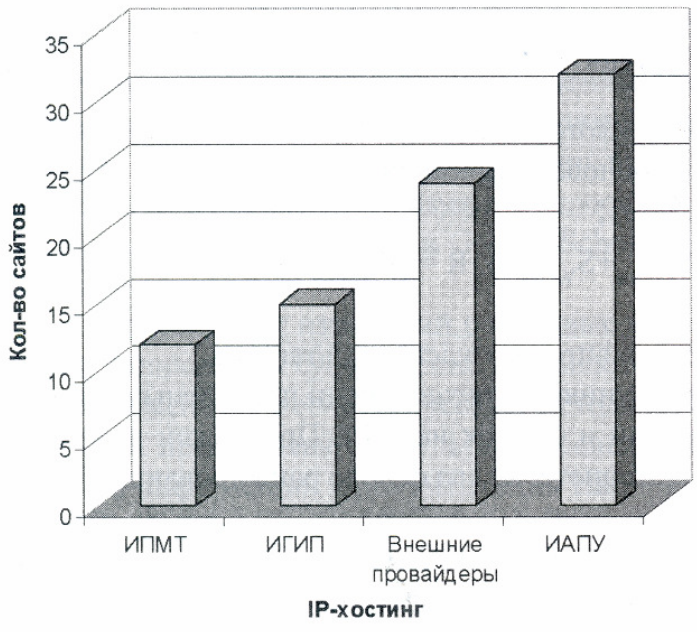

Рис. 1. IP-хостинг сайтов ДВО РАН

Наибольшее количество сайтов в Отделении относится к наукам о Земле (30), биологическим (18), физико-математическим (13), остальные представлены единично. График показывает количество сайтов по различным научным направлениям в ДВО РАН (рис. 3).

Важной характеристикой любого web-сайта является его временная устойчивость, под которой мы понимаем его присутствие в Интернете в течение продолжительного времени при неизменности имени и адреса сайта. По этому признаку проанализировано

65 сайтов, у 18 возраст определить не удалось (рис. 4). В ДВО РАН 6 сайтов старше 10 лет: официальный сайт Президиума, два сайта ДВГИ – информационный сервер (ИС) и региональный портал «Приморский край России», три ТОИ – официальный (http://www.poi.dvo.ru), ИС «Океанография и состояние морской среды Дальневосточного региона России» (http://www.pacificinfo.ru), Архив электронных научных публикаций InfoNet ТОИ (http://infonet.dvo.ru). Первые 3 обладают наивысшей среди сайтов ДВО РАН временной устойчивостью, поскольку никогда не меняли ни имени, ни адреса.

Присутствие научных сайтов в интернет-каталогах – важная характеристика доступности. График анализа присутствия сайтов ДВО РАН в крупных российских каталогах (Yandex, Рамблер, Апорт, Dmoz, Mail.ru) представлен на рис. 5.

Только 4 сайта ДВО РАН присутствуют во всех 5 каталогах: http://www.febras.ru, http://www.fegi.ru, http://www.iacp.dvo.ru/lib/, http://lipid.narod.ru. 40 сайтов ДВО РАН (50%) отсутствуют во всех пяти.

Одним из индикаторов присутствия сайтов в Интернете является количество страниц сайта, проиндексированных поисковыми машинами. Это означает, что поисковый робот посетил сайт, проанализировал его и занес информацию в базу данных (БД) системы. Если страница занесена в индекс поисковика, она может быть показана в результатах поиска. Если страница в индексе отсутствует, то поисковая система ничего не знает о ней, следовательно, никак не может использовать информацию с нее. Однако при проведении замеров обнаруживается, что разные российские и зарубежные поисковые машины индексируют различное количество страниц на одном и том же сайте. Механизмы индексации страниц

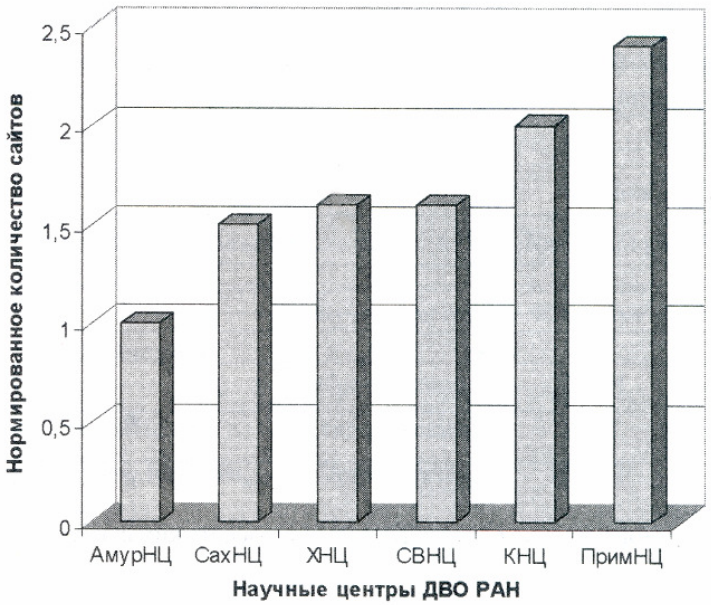

Рис. 2. Количество сайтов в научных центрах ДВО РАН

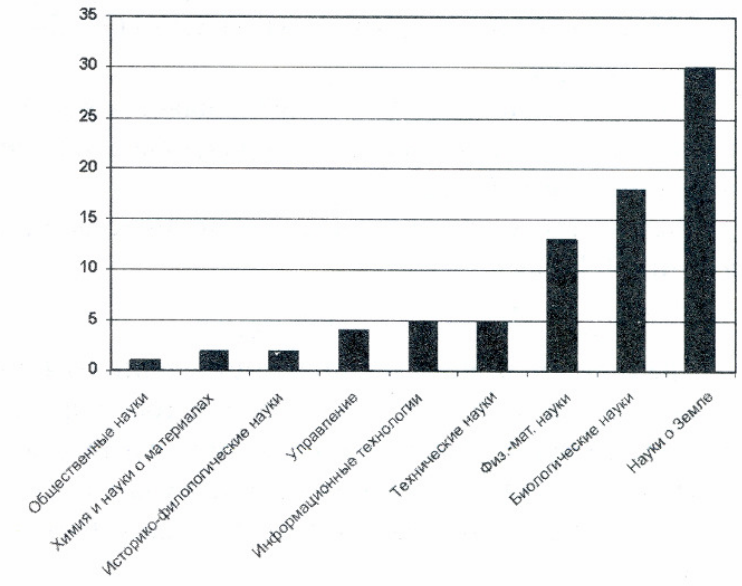

Рис. 3. Количество сайтов по различным научным направлениям в ДВО РАН

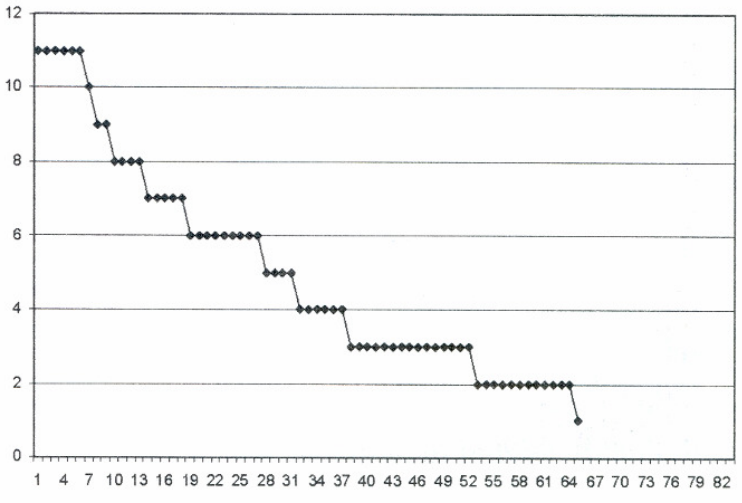

Рис. 4. Продолжительность существования сайтов ДВО РАН. По оси абсцисс – номера сайтов

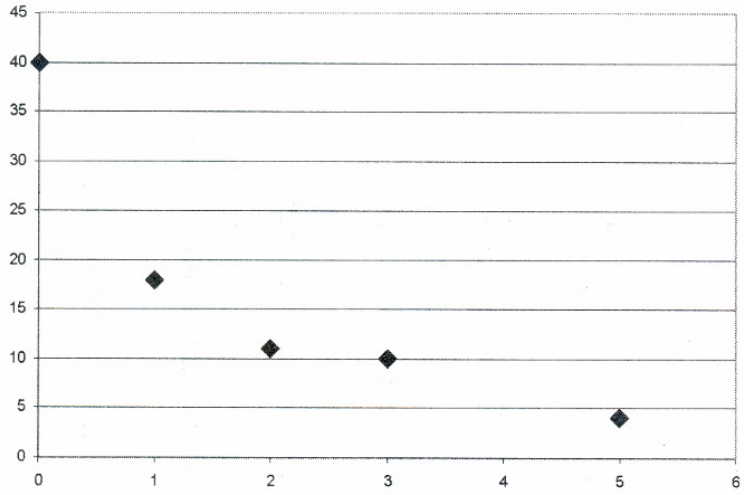

Рис. 5. Присутствие сайтов ДВО РАН в российских каталогах

являются секретной особенностью поисковых машин и, по-видимому, существенно влияют на контекстный поиск по поисковым фразам. Можно предположить, что чем больше страниц сайта проиндексировано, тем вероятнее возможность попадания на сайт через эту поисковую машину.

Таким образом, индикатор присутствия сайта может вычисляться как некий интегральный показатель по замерам нескольких наиболее распространенных поисковых машин.

Нами определено среднее количество страниц для всех сайтов ДВО РАН, проиндексированных поисковиками Google, Yahoo, MSN, Yandex, Рамблер, Апорт:

$$S = [S_{Яндекс} + S_{Google} + S_{Yahoo} + S_{MSN} + S_{Апорт} + S_{Rambler}]/6. \quad (3)$$

Наибольшее среднее количество страниц у сайтов Камчатского научного центра (ИВиС) (26638) и базовой сети ДВО РАН (ИАПУ) (25824). Для остальных сайтов это число значительно меньше.

Важной характеристикой сайта является количество ссылок с других сайтов. На рис. 6 представлена диаграмма, характеризующая количество ссылок на сайты Отделения в поисковой системе Google.

Вычислялись средние значения количества ссылок в поисковых системах Google, Yahoo, MSN, Yandex:

$$V = [V_{Яндекс} + V_{Google} + V_{Yahoo} + V_{MSN}]/4. \quad (4)$$

Традиционная мера цитируемости, принятая в академической среде, равна количеству ссылок на документ. Хотя она дает некоторое представление о качестве документа, создатели Google пошли дальше и предложили считать ссылки с разных страниц не равными друг другу. Ими была предложена мера популярности web-страницы, не зависящая от запроса пользователя, которая называется PageRank. Вычисляется она итеративно для всех страниц Интернета:

$$PR(A) = (1-d) + d(PR(t1)/C(t1) + \ldots + PR(tn)/C(tn)), \quad (5)$$

где t1–tn – страницы, ссылающиеся на A, C(tn) – количество исходящих ссылок на соответствующей странице; d – коэффициент, обычно равный 0,85. Данная формула аппроксимирует модель случайного блуждания пользователей в Интернете. Можно придерживаться примерно такой градации: PageRank от 4 до 5 – наиболее типичный для большинства сайтов средней известности, 6 – очень популярный проект, 7 – очень значимый и известный проект, 8–10 имеют исключительно популярные и значимые проекты. Например,

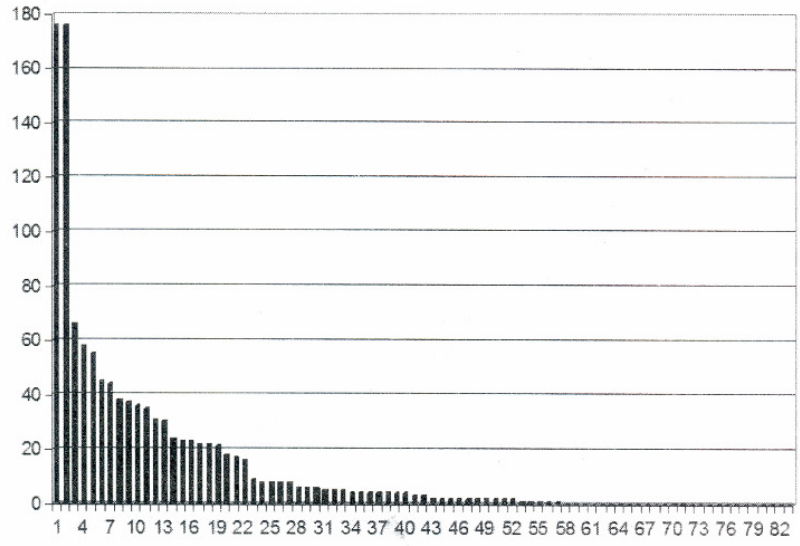

Рис. 6. Количество ссылок на сайты ДВО РАН по информации Google

в данный момент PR у сайта русской Википедии равен 8, у английской Википедии и сайта Microsoft – 9. Значение 10 имеют всего несколько десятков сайтов, в их числе Google. Google использует показатель PageRank найденных по запросу страниц, чтобы определить порядок выдачи этих страниц посетителю в результатах поиска (табл. 2).

Таблица 2

Сайты Дальневосточного отделения РАН, имеющие наибольшее значение (6) PageRank Google

Название сайта и интернет-адрес	Научный центр	Институт
Информационный сервер Дальневосточного геологического института ДВО РАН http://www.fegi.ru	Приморский	ДВГИ
Институт экономических исследований http://www.ecrin.ru	Хабаровский	ИЭИ
Официальный сайт Президиума ДВО РАН http://www.febras.ru	Приморский	ИПМТ
Региональный портал «Приморский край России» [1] http://www.fegi.ru/primorye	– " –	ДВГИ
Управление научно-исследовательского флота http://www.unif.febras.ru	– " –	УНИФ

У 9 сайтов ДВО РАН PR = 5, у 36 – 4, у 7 – 3, у 3 – 2, у 21 сайта PR = 0.

Специализированный научный поисковик Google Scholar предназначен для поиска научной литературы, включая рефераты диссертаций, статьи, книги, другие научные публикации и ссылки на них. Результаты поиска сортируются по релевантности (табл. 3). Google Scholar настроен на автоматическое извлечение цитат из документов, которые в данный момент отсутствуют в свободном доступе. В его БД входят как документы открытого доступа, так и исключительно подписные материалы. Последние включаются в БД по согласованию с издателями, а полные версии статей доступны только подписчикам. Прочие пользователи могут ознакомиться лишь с рефератами таких материалов. Впрочем, пока Google Scholar работает не со всеми научными издательствами, и сам сервис еще тестируется.

Таблица 3

Индексы релевантности сайтов ДВО РАН в Google Scholar

Название, адрес	Научный центр	Институт	Индекс
Базовая сеть ДВО РАН http://www.dvo.ru	Приморский	ИАПУ	4400
Биолого-почвенный институт http://www.biosoil.ru	– " –	БПИ	547
Камчатский научный центр http://www.kscnet.ru	Камчатский	ИВиС	451
Информационный сервер Дальневосточного геологического института ДВО РАН http://www.fegi.ru	Приморский	ДВГИ	224
Официальный сайт Президиума ДВО РАН http://www.febras.ru	– " –	ИПМТ	220

Сайты, не представленные в табл. 3, имеют единичные или нулевые индексы.

Индекс цитирования – принятая в научном мире мера «значимости» трудов ученого, она определяется количеством ссылок на этот труд (или фамилию) в других источниках. Однако для точного значения важно не только количество, но и качество ссылок. Так, значимость ссылок будет разной у авторитетного академического издания, популярной брошюры и развлекательного журнала.

Тематический индекс цитирования Yandex определяет «авторитетность» интернет-ресурсов с учетом качественной характеристики ссылок на них с других сайтов («вес» ссылки), рассчитываемой по специально разработанному алгоритму. Большую роль играет тематическая близость ресурса и ссылающихся на него сайтов, а также количество ссылок на ресурс, но индекс определяется не только количеством ссылок, но и их «весовой» суммой.

Индекс цитирования Yandex как средство определения авторитетности ресурсов призван обеспечить релевантность расположения ресурсов в рубриках каталога Yandex. Сайты

в рубриках Каталога Yandex расположены по убыванию их тематического индекса цитирования. Индекс цитирования Yandex не является чисто количественной характеристикой, поэтому Yandex показывает некоторые округленные значения, которые помогают ориентироваться в «значимости» («авторитетности») ресурсов в каждой области (теме). Индекс цитирования Yandex рассчитывается для всех ресурсов, ссылки на которые Yandex нашел в Интернете, при условии, что результирующее значение индекса для них не меньше 10.

В табл. 4 представлены сайты ДВО РАН, имеющие наивысший индекс цитируемости Yandex.

Таблица 4

Сайты Дальневосточного отделения РАН, имеющие наибольший индекс цитируемости Yandex

Название сайта и интернет-адрес	Научный центр	Институт	тИЦ
Информационный сервер Дальневосточного геологического института ДВО РАН http://www.fegi.ru	Приморский	ДВГИ	1900
Официальный сайт Президиума ДВО РАН http://www.febras.ru	– ” –	ИПМТ	800
Региональный портал «Приморский край России» http://www.fegi.ru/primorye	– ” –	ДВГИ	600
Базовая сеть ДВО РАН http://www.dvo.ru	– ” –	ИАПУ	475
Биолого-почвенный институт http://www.biosoil.ru	– ” –	БПИ	325

22 сайта ДВО РАН имеют индекс цитируемости Yandex > 100, у 33 сайтов – 10–100, у 28 – нулевой, поэтому при поиске в Интернете с использованием самой крупной поисковой машины Yandex российского сегмента Интернета у большей части сайтов Дальневосточного отделения РАН очень низкие шансы представить свои материалы широкой аудитории российских пользователей Интернета.

Ранжирование сайтов ДВО РАН проведено нами с учетом трех различных индексов: PageRank Google, тИЦ Yandex и Google Scholar (табл. 5). Всего в рейтинге участвовало 14 сайтов ДВО РАН, поскольку остальные сайты не имеют ни одной рейтинговой оценки в этих поисковых системах. Оценки были нормированы в интервале от 0 до 1, затем получено среднее значение.

Таблица 5

Ранжирование сайтов ДВО РАН

Место	Институт	Адрес	Значение признака
1	Дальневосточный геологический институт	http://www.fegi.ru	0,68
2	Базовая сеть ДВО РАН (ИАПУ)	http://www.dvo.ru	0,66
3	Официальный сайт Президиума	http://www.febras.ru	0,48
4	Институт экономических исследований	http://www.ecrin.ru	0,33
5	Камчатский филиал Тихоокеанского института географии	http://www.terrakamchatka.org	0,27
6	Биолого-почвенный институт	http://www.biosol.ru	0,26
7	Амурская научная сеть (ИГиП) (на реконструкции)	http://www.ascnet.ru	0,25
8	Камчатский научный центр (ИВиС)	http://www.kscnet.ru	0,23
9–12	Институт морской геологии и геофизики	http://www.imgg.ru	0,17
9–12	Институт машиноведения и металлургии	http://www.imim.ru	0,17
9–12	Институт комплексного анализа региональных проблем	http://www.ikarp.ru	0,17
9–12	Институт космофизических исследований и распространения радиоволн	http://www.ikir.ru	0,17
13	Уссурийская астрофизическая обсерватория	http://www.uafo.ru	0,16
14	Ботанический сад-институт	http://www.botsad.ru	0,11

Результаты ранжирования оказались несколько неожиданными: высокоразвитые сайты ТОИ (http://www.poi.dvo.ru), ИБМ (http://www.imb.dvo.ru), ИВиС (http://www.kscnet.ru/ivis) не вошли в рейтинг. Видимо, отсутствие в их адресах доменов второго уровня и недостаточная работа по увеличению популярности сайтов в Интернете снижают их рейтинги у поисковых систем Google и Yandex.

Анализ информационного пространства ДВО РАН позволяет сформулировать предложения для его дальнейшего развития.

1. Создание централизованного сегмента ДВО РАН в Интернете. Официальный сайт Президиума ДВО РАН не играет роли централизованного сегмента научных ресурсов Отделения в Интернете, поскольку выполняет другие функции. На наш взгляд, необходимо создание по крайней мере двух централизованных ресурсов в Дальневосточном отделении РАН: научного информационного портала ДВО РАН и электронной библиотеки ДВО РАН, которые помимо своих информационных функций взяли бы на себя роль www-коммуникаторов Отделения.

2. Проведение работ по увеличению рейтинговых оценок сайтов ДВО РАН в Интернете. В этих работах необходимо учитывать следующее:

1) наличие домена второго уровня для сайта – необходимое условие для популярности сайта в Интернете. Многие исследователи утверждают, что наличие в имени домена ключевого слова резко повышает шансы на рост рейтинговых оценок;

2) желательно добиться временной устойчивости сайта, т.е. длительного времени жизни, неизменности имени и адреса сайта;

3) необходимо добиваться наличия большого количества уникальных научных материалов на сейте. Такая характеристика, как периодичность обновления информации, в данной работе не рассматривалась, но именно она является тем условием, которое привлекает на сайт постоянных посетителей, тем самым способствуя увеличению рейтинговых оценок сайта в поисковых системах. Статичные сайты быстро теряют свою популярность в Интернете;

4) очень важна высокая степень рекламы сайта в Интернете. Важно обозначить свое присутствие не только в глобальных Каталогах и в БД поисковых машин, но и в научных каталогах, на сайтах близких по направлениям научных организаций и на других сайтах, что дает большое количество ссылок, в том числе и с высоко рейтинговых сайтов;

5) желательно провести работы по увеличению количества проиндексированных страниц основными мировыми поисковыми машинами;

6) непрерывность и безотказность работы сервера в Интернете и высокая скорость ответа сервера на запросы пользователей – очевидная характеристика любого интернет-сайта. И этого нужно добиваться.

Измерения, приведенные в статье, получены в марте 2009 г.

ЛИТЕРАТУРА

1. Наумова В.В. Региональный информационный сервер «Приморский край России» // Информ. бюл. ГИС-ассоциации России. 1999. № 46. С. 10-13.

2. Печников А.А. Вебометрические исследования Web-сайтов университетов России // Информ. технологии. 2008. № 11. С. 74-78.

3. Печников А.А., Чуйко Ю.В. Исследование согласованного поведения малых Интернет-сообществ // Телекоммуникации. 2008. № 10. С. 8-12.

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

5. Aguillo I.F., Granadino B., Ortega J.L., Prieto J.A. Scientific research activity and communication measured with cybermetric indicators // J. American Soc. Information Science and Technology. 2006. N 57 (10). P. 1296-1302.

6. Aguillo I.F., Granadino B., Ortega J.L., Prieto J.A. What the Internet says about Science // The Scientist. 2005. 19 (14): 10, Jul. 18. P. 57-61.