

Анализ структур научных Интернет-ресурсов (на примере ресурсов Дальневосточного отделения РАН)

К.А. Платонов, В.В. Наумова

Дальневосточный геологический институт ДВО РАН

platonov@fegi.ru, naumova@fegi.ru

Аннотация

В статье рассматриваются вопросы разработки подходов для анализа структур научных сайтов в Интернет. Предлагается новый подход для описания и анализа структур научных ресурсов. Подход опробован при проведении многомерного анализа www-ресурсов институтов Дальневосточного отделения РАН.

1. Введение

В 1997 году Tomas C. Almind и Peter Ingwersen в статье [10] определили основные понятия нового раздела информатики. Раздел информатики, в котором исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к World Wide Web назван «webometrics».

Первый опыт анализа научного Интернет-пространства с использованием вебометрических методик получен Кибернетической лабораторией (Cybermetrics Lab) Центра научной информации и документации Национального исследовательского совета Испании. Сотрудниками этой лаборатории предложены методы для получения признаков и ранжирования на их основе официальных сайтов научных и образовательных учреждений. В работе [9] ими были определены основные параметры - вебометрические индикаторы S,V,R,Sc:

- количество уникальных гипертекстовых ссылок с других ресурсов (V – visibility);
- общее количество страниц сайта проиндексированных поисковыми системами (S – size);
- количество полнотекстовых файлов проиндексированных поисковыми системами, под которыми понимаются файлы с расширениями pdf, ps, doc, xls, ppt и ttf (R – rich files);
- количество статей, размещенных на сайте, и их цитируемость проиндексированных системой Google Scholar (Sc – “sholar”).

В той же статье приведена формула для вычисления интегрального показателя, названного «вебометрическим рангом» (WR-Webometrics Rank):

$$WR = 4*RankV + 2*RankS + 1*RankR + 1*RankSc \quad (1)$$

Результаты вычислений мирового рейтинга сайтов доступны на специализированном сайте Cybermetrics Lab [11].

Анализ различных множеств научных Интернет-ресурсов РФ проводят сотрудники Института вычислительных технологий СО РАН [7] и Института научной и педагогической информации РАО [2]. За основу расчетов ими приняты методы и подходы Cybermetrics Lab с некоторыми изменениями вида основной формулы, а также выбора поисковых сервисов.

Анализ различных множеств Интернет-ресурсов проведен Печниковым А.А. с соавторами [5]: научных учреждений РАН, университетов РФ, органов государственной власти Республики Карелия. Этой группой исследователей предложено применение теоретико-графовых моделей для выявления структур Интернет-пространства. На основе полученных результатов построена модель «научного» веба. Предложенная модель позволяет выделить некоторые структуры исследуемого множества: административный каркас, тематические связи, выделить внутренние ресурсы сайта.

В исследованиях Интернет-пространства Дальневосточного Отделения РАН (ДВО РАН) [6] и Отделения наук о земле РАН (ОНЗ РАН) [4] Наумовой В.В. и Платоновым К.А., применены современные методы анализа данных, в частности метод главных компонент, а также некоторые методы теории графов (рис.1).

2. Объединение Интернет-ресурсов научной организации

Доступ к научной информации глобальной сети Интернет для пользователей осуществляется, в основном, с помощью поисковых систем. Наличие высоких рейтинговых оценок для сайта очень важно с точки зрения его более высокой доступности широкой аудитории пользователей Интернет, поскольку рейтинговые оценки используют практически все поисковые системы. Результаты запросов сортируются поисковыми машинами и предоставляются пользователям в порядке уменьшения их рейтингов. Таким образом, наличие высоких рейтингов ставит сайт в лучшее положение по отношению к другим сайтам.

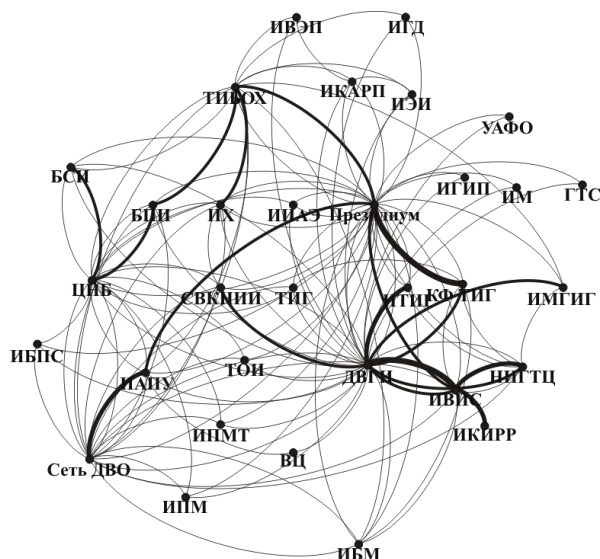


Рис 1. Структура информационного пространства официальных сайтов институтов ДВО РАН на основе теории графов. Условные обозначения в табл. 2

Научные сайты РФ представляют собой следующие группы ресурсов: официальные сайты институтов, университетов и научных организаций, научных лабораторий, групп и отдельных ученых; научные электронные библиотеки, в том числе и сайты научных журналов; сайты конференций; цифровые репозитории, научные информационные системы, ГИС-порталы; научные тематические ресурсы; научно-популярные ресурсы; каталоги научных ресурсов; другие сайты.



Рис 2. Множество Интернет-ресурсов Дальневосточного геологического института ДВО РАН

Если рассматривать сайты Институтов РАН, то они представляют собой сложный информационный комплекс, состоящий из разных групп научных сайтов. Многие институты/учреждения поддерживают несколько различных доменных областей. Таким

образом, их реальное присутствие в Сети определяется множеством адресов. Кроме того, различные подразделения одной организации могут иметь собственные Интернет-ресурсы с адресами, не ассоциируемыми поисковыми машинами с адресом основного сайта.

В качестве примера на рис. 2 представлено множество Интернет-ресурсов Дальневосточного геологического института ДВО РАН.

Таблица 1. Предлагаемое признаковое пространство

№	Признак	Примечание
1	Узлы	Количество ресурсов института
2	Дуги	Количество ссылок между ресурсами института
3	Порядок связанности	Количество узлов, входящих в компоненту сильной связанности
4	Изолированные узлы	Количество изолированных узлов
5	Уровень вложенности	Средний уровень вложенности ресурсов института.
6	Внутренние дуги	Количество исходящих ссылок с ресурсов института внутри множества
7	Внешние дуги	Количество исходящих ссылок с ресурсов института за пределы множества.
8	Входящие дуги	Количество входящих ссылок с ресурсов других институтов множества.
9	Размер	Количество страниц, проиндексированных поисковыми системами по всем ресурсам института.
10	Видимость	Количество уникальных гипертекстовых ссылок с других ресурсов на ресурсы института
11	Кол-во файлов	Количество полнотекстовых файлов, проиндексированных поисковыми системами по всем ресурсам (pdf, ps, doc, xls, ppt и rtf)
12	Научность	Количество статей, размещенных на сайте, и их цитируемость (Google Scholar)
13	Время жизни	Время существования самого старого ресурса института

Таким образом, при анализе научного www-пространства возникает задача объединенного количественного описания всего множества интернет-ресурсов научного института.

В работах Cybermetrics Lab эта задача была поставлена в целях уточнения вебметрических индикаторов научных организаций и университетов, но методов вычисления предложено не было.

Наумовой В.В. [3] предложено использовать формулу объединения известных индикаторов по всем ресурсам одной научной организации.

Для решения этой задачи нами предлагается подход, заключающийся в создании нового признакового пространства на основе объединения известных индикаторов по всем ресурсам научной организации и вычислении некоторого количества новых численных показателей при аппроксимации множества ресурсов организации взвешенным, ориентированным графом (табл. 1).

Тестовая апробация подхода проведена на множестве ресурсов ДВО РАН. Выборка охватывает 67 сайтов, принадлежащих 31 институту. Она содержит 31 официальный сайт, 3 сайта научных центров, 2 сайта филиалов, 5 ГИС-порталов, 2 электронных библиотеки, 1 электронная энциклопедия, 1 научный сервис, 2 научных журнал и 1 сайт газеты, 3 информационных системы, 2 видеопортала, 6 научно тематических ресурсов, 2 сайта лабораторий, сайт образовательной кафедры, 3 административных сайта, сайт музея.

Интерес представляет сравнительный анализ диаграмм многомерной классификации при старом и новом подходе к анализу.

Результаты многомерной классификации официальных сайтов ДВО РАН методом главных компонент в признаковом пространстве V, S, R, Sc представлены на рис.3 (июль 2012 г.), а результаты классификации интегрированных ресурсов институтов ДВО РАН (сентябрь 2013 г.) в предложенном авторами признаковом пространстве - на рис. 4.

Таблица 2. Положение сайтов институтов ДВО РАН в мировом рейтинге научных сайтов [10]

Сокр.	Институт (адрес)	Мировой рейтинг/ рейтинг РАН
ДВГИ	Дальневосточный геологический институт (http://fegi.ru)	1379/38
ИАПУ	Институт автоматизации и процессов управления (http://iacp.dvo.ru)	2291/52
ТОИ	Тихоокеанский океанологический институт (http://poi.dvo.ru)	2347/56
БПИ	Биолого-почвенный институт (http://biosoil.ru)	2856/67-
ИБМ	Институт биологии моря (http://www.imb.dvo.ru)	3889/81
ИКИРР	Институт космофизических исследований и распространения радиоволн (http://www.ikir.ru)	4120/87
ИПМТ	Институт проблем морских технологий(http://imtp.febras.ru)	4309/93
ТИГ	Тихоокеанский институт географии(http://tig.dvo.ru)	4952/97
ИМГГ	Институт морской геологии и геофизики (http://imgg.ru)	5195/99
ИЭИ	Институт экономических исследований (http://ecrin.ru)	5797/105
ИПМ	Институт прикладной математики (http://www.iam.dvo.ru)	6788/111
ЦНБ	Центральная научная библиотека (http://www.cnb.dvo.ru)	7134/114

ГТС	Горнотаежная станция им. В.Л.Комарова (http://gts.febras.ru)	7473/117
ИИАЭ	Институт истории, археологии и этнографии народов ДВ (http://ihaefe.org)	7490/118
ТИБОХ	Тихоокеанский институт биологической химии (http://pihoc.dvo.ru)	7526/119
БСИ	Ботанический сад-институт (http://botsad.ru)	-/-
ВЦ	Вычислительный центр г.Хабаровск (http://ccfebras.ru)	-/-
ИБПС	Институт биологических проблем Севера (http://ibpn.ru)	-/-
ИВЭП	Институт водных и экологических проблем (http://ivep.as.khb.ru)	-/-
ИВИС	Институт вулканологии и сейсмологии (http://www.kscnet.ru/ivs)	-/-
ИГИП	Институт геологии и природопользования (http://ignm.ru)	-/-
ИГД	Институт горного дела (http://igd.khv.ru)	-/-
ИКАРП	Институт комплексного анализа региональных проблем (http://icarp.ru)	-/-
ИМ	Институт материаловедения (http://im.khv.ru)	-/-
ИТИГ	Институт тектоники и геофизики им. Ю. А. Косыгина (http://itig.as.khb.ru)	-/-
ИХ	Институт химии (http://www.ich.dvo.ru)	-/-
Арктика	Международный научно-исследовательский центр "Арктика" (http://arktika.north-east.ru)	-/-
НИГТЦ	Научно-исследовательский геотехнологический центр (http://nigtc.kscnet.ru)	-/-
ПО	Научно-образовательный комплекс "Приморский океанариум" (http://russian-aquarium.ru)	-/-
СВКНИИ	Северо-Восточный комплексный научно-исследовательский институт (http://neisri.ru)	-/-
УАФО	Уссурийская астрофизическая обсерватория (http://uafo.ru)	-/-

И в первом и во втором графике главная компонента X может интерпретироваться как «качество сайта». Но если в первом случае, мы наблюдаем позиции в зависимости от известности сайта и его объема, то во втором - сильное влияние на позицию оказывает связанность Интернет-ресурсов института. Визуально видно, что график рис.4 более корректно отражает присутствие институтов в Интернет, что подтверждается местами в мировом рейтинге научных сайтов институтов - лидеров выборки: ДВГИ, ИАПУ, ТОИ, БПИ, БПИ (табл.2).

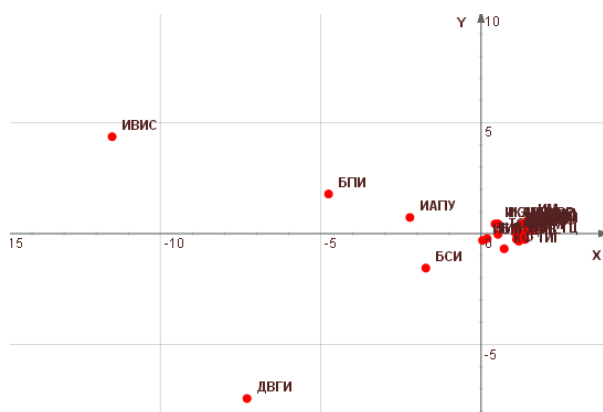


Рис 3. Многомерная классификация ресурсов институтов ДВО РАН методом главных компонент (07.2012 г.)

Компонента $X = (-0,78 * x_1) + (-0,92 * x_2) + (-0,82 * x_3) + (-0,89 * x_4) + (-0,84 * x_5) + (-0,14 * x_6) + (-0,85 * x_7) + (-0,71 * x_8) + (-0,45 * x_9) + (-0,82 * x_{10}) + (-0,80 * x_{11}) + (-0,81 * x_{12}) + (-0,73 * x_{13}) + (-0,77 * x_{14})$

$$\text{Компонента } \mathbf{Y} = (0,49*x_1) + (-0,34*x_2) + (0,54*x_3) + (0,05*x_4) + (0,10*x_5) + (0,15*x_6) + (0,47*x_7) + (-0,56*x_8) + (-0,19*x_9) + (-0,55*x_{10}) + (0,53*x_{11}) + (-0,17*x_{12}) + (-0,59*x_{13}) + (0,50*x_{14})$$

Где : X_1 -S_{google}, X_2 -V_{google}, X_3 -R_{google}, X_4 -SC_{google}, X_5 -S_{yandex}, X_6 -V_{yandex}, X_7 -R_{yandex}, X_8 -Sc_{yandex}, X_9 -S_{yahoo}, X_{10} -V_{yahoo}, X_{11} -R_{yahoo}, X_{12} -S_{bing}, X_{13} -V_{bing}, X_{14} -R_{bing}

Условные обозначения в табл.2.

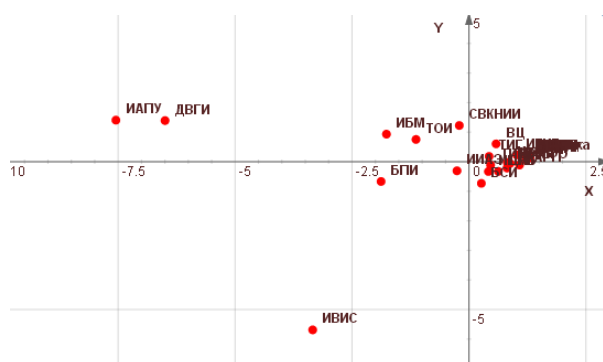


Рис 4. Многомерная классификация ресурсов институтов ДВО РАН методом главных компонент (09.2013 г.)

$$\text{Компонента X} = (-0,82 \cdot x_1) + (-0,95 \cdot x_2) + (-0,78 \cdot x_3) + (-0,79 \cdot x_4) + (-0,84 \cdot x_5) + (-0,77 \cdot x_6) + (-0,45 \cdot x_7) + (-0,67 \cdot x_8)$$

Компонента $Y = (0,35 * x_1) + (0,20 * x_2) + (0,24 * x_3) + (-0,06 * x_4) + (0,38 * x_5) + (-0,37 * x_6) + (-0,84 * x_7) + (-0,40 * x_8)$

Где: X_1 -узлы, X_2 -узлы КСС, X_3 -внут. дуги, X_4 -внеш. дуги, X_5 -Размер, X_6 -Видимость, X_7 -кол-во файлов, X_8 -Научность.

Условные обозначения в табл.2.

3. Выводы

Проведенный анализ структур сайтов Дальневосточного отделения РАН позволяет сделать следующие выводы:

1. Структура www-сайтов ДВО РАН представляет собой открытое множество интернет-ресурсов со свободными связями, с объектами способными

адаптироваться, изменять свою структуру. Таким образом, в изучаемом множестве нами выделены основные признаки самоорганизующейся системы, следуя следующему определению:

Самоорганизующиеся системы — это открытые системы, свободно обменивающиеся с внешней средой энергией, веществом и информацией [1]. Одной из основных особенностей самоорганизующихся систем является способность противостоять энтропийным тенденциям, способность адаптироваться к изменяющимся условиям, преобразуя при необходимости свою структуру.

2. Положение официального сайта института в Интернет зависит в первую очередь от профессиональности web-мастера. Существенная его работа по прописыванию сайтов в тематические каталоги, регистрация сайта в большом количестве поисковых системах, интеграция с существующими родственными тематическими сайтами, добавление на сайт профессиональных ссылок.

3. Положение официального сайта института в Интернет так же зависит от объемов профессионального информационного материала в виде полнотекстовых файлов и систематического наполнения ресурса новым содержимым.

4. Крайне важным является наличие полной англоязычной версии официального сайта института.

5. Наличие у официального сайта института ресурсов-спутников, таких как цифровые репозитории, электронные библиотеки, сайты конференций, ГИС-порталы, информационные системы, различные тематические ресурсы, в том числе и научно-популярные и научно-образовательные ресурсы, с правильной адресацией в Интернет позволяет привлечь внимание пользователей Интернет к сайту научного института.

6. Несомненно, важными характеристиками научных сайтов являются: высокоскоростной доступ к его ресурсам, безотказность работы серверов, адресация в доменах второго уровня, приличный дизайн, современные программные платформы реализации, «долгоживучесть» в Интернет.

За рамками проведенных исследований остался важный для нас вопрос, решением которого мы предполагаем заняться в дальнейшем.

Влияет ли «имя» и/или публикационная активность и цитируемость научного института на его рейтинговые позиции в научном сегменте Интернет?

В рамках задачи изучения научного сегмента РАН в Дальневосточном геологическом институте Дальневосточного отделения РАН (ДВГИ ДВО РАН) реализован научный сервис: <http://webometrics.fegi.ru>, который осуществляет мониторинг научных ресурсов РАН, а так же содержит ряд сервисов анализа: ранжирование, анализ взаимосвязей, анализ тематических структур.

Работа выполняется при финансовой поддержке гранта ДВО РАН № 13-П-В-08-022 «Анализ региональных структур научных сайтов РАН методами

вебометрики (на примере анализа Интернет-пространства Дальневосточного отделения РАН)».

Литература

- [1] Акимов Т.А. Теория организации: Учеб. пособие для вузов. М.: ЮНИТИ-ДАНА, 2003.
- [2] Антопольский А.Б., Поляк Ю.Е., Усанов В.Е. О российском индексе веб-сайтов научно-образовательных учреждений // Информационные ресурсы России. М.: 2012. № 4.
- [3] Наумова В.В. Мониторинг Информационного пространства Дальневосточного отделения РАН // Вестник ДВО РАН. 2012. №2. С. 117-122.
- [4] Наумова В. В., Платонов К. А. Сервис вебометрического анализа структур научных сайтов // Вестник ОНЗ РАН. 2013. Вып. 1. NZ1001, doi:10.2205/2013NZ000000. [Электронный ресурс]. — URL: <http://onznews.wdcb.ru/publications/v05/2013NZ000117/2013NZ000117.pdf>
- [5] Печников А.А. Разработка математических моделей, методов и программных средств для исследования взаимосвязей регламентируемых веб-сайтов. Автореферат диссертации на соискание ученой степени д.т.н., Петрозаводск. 2011. 35 с.
- [6] Платонов К.А. Анализ региональных структур научных сайтов РАН методами вебометрики (на примере анализа Интернет-пространства Дальневосточного отделения РАН) // Научный сервис в сети Интернет: поиск новых решений: Труды Международной суперкомпьютерной конференции (17-22 сентября 2012 г., г. Новороссийск). – М.: Изд-во МГУ, 2012. 646 с (электронная версия: [Электронный ресурс]. — URL: <http://agora.guru.ru/abrau2012/pdf/646.pdf>).
- [7] Ханчук А.И., Наумова В.В. Информационное пространство Дальневосточного отделения РАН // Вестник ДВО РАН. 2009. №4. С. 122-130.
- [8] 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.
- [9] Aguillo I.F., Granadino B., Ortega J.L., Prieto J.A. Scientific research activity and communication measure with cybermetrics indicators // Journal of the American Society for the Information Science and Technology. 2006. Vol. 57, № 10. P. 1296-1302.
- [10] Tomas C. Almind and Peter Ingwersen "Informetric analyses on the World Wide Web: Methodological approaches to «webometrics»" // Journal of Documentation 53 (4): 404–426.
- [11] Ranking Web of World Research Centers [Электронный ресурс] – 2013. – Режим доступа: <http://research.webometrics.info/en/RAS>

Analysis of Science Websites Structure on the Example of the FEB RAS Internet Recourses

K.A. Platonov, V.V. Naumova

The authors develop new approaches for describing and analyzing the science websites structures. The approach was tested by multivariate analysis of the FEB RAS institutes' web-sites.