ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 004.738:002

В. И. Якимов, В. М. Ефременкова, В. Г. Севастьянов

Поиск в Интернете электронных ресурсов по узкотематическим направлениям. Методика оценки информативности и достоверности научных сайтов

Анализ динамики развития Интернета выявил увеличение числа научных сайтов. В рамках статьи описаны основные поисковые машины и каталоги. Показана важность предварительного поиска информации в глобальной сети. Проведен поиск сайтов по направлению "фуллерены" в поисковой машине Google, каталоге DMOZ и базе данных SCOPUS. Определены основные критерии оценки и степень научной важности сайтов отдельных ученых и научных групп.

1. ВВЕДЕНИЕ

Основным толчком в создании и развитии всемирной паутины в современном ее виде послужила новая "сетевая инициатива", выдвинутая Альбертом Гором в 1991 г., — декларация о национальной и глобальной информационной инфраструктуре. Созданная в 1989 г. в Европейской лаборатории физики частиц (CERN) в Женеве сеть передачи данных World Wide Web [1], с помощью которой ученые могли обмениваться результатами при выполнении проекта, легла в основу информационной системы для распространения информации в различных областях знания. World Wide Web и графический Web-браузер Mosaic X, разработанный в 1993 г. в NCSA (National Center for Supercomputing Application), открыли доступ к Интернету широкому кругу пользователей. Публикация знаменитого меморандума Б. Клинтона — А. Гора "Технологии для экономического роста США: новые направления, которые предстоит создать" (1993), как и знаменитый доклад Мартина Бангеманна "Рекомендации ЕС и глобальное информационное сообщество" (1994) фактически подтолкнули мир к принятию Интернета. В 1994 г. на первой Всемирной конференции по развитию телекоммуникаций, состоявшейся в Буэнос-Айресе (Аргентина), вице-президент США Альберт Гор определил основную задачу глобального информационного общества: "... создать глобальное сообщество, в котором население соседних стран рассматривает друг друга не как потенциальных врагов, а как потенциальных партнеров, как членов одной семьи в огромной, все в большей степени взаимосвязанной человеческой семье" [2]. Именно 1994 г. считается годом создания Интернета.

Число пользователей с 1994 г. многократно возросло. До 2002 г. оно ежегодно в среднем увеличивалось на 2%. Но уже в 2003 г. по результатам исследования компании **Ipsos-Reid** [3], число пользователей Интернета в мире выросло на 7% и, по данным компании **VeriSign Inc.**— администратора доменов COM и NET, составило 580 млн. [3].

Наиболее активно пользуются Интернетом в Канаде: 71% взрослого населения страны выходит в сеть не реже раза в месяц. В пятерку мировых лидеров по этому показателю также вошли Южная Корея (70%), США (68%), Япония (65%) и Германия (60%). В России, согласно отчету Ipsos-Reid, Интернетом пользуется 10% городского населения. Необходимо учесть, что число активных пользователей в России, по данным Фонда "Общественного мнения" (ФОМ), составляет 34,9% от общего числа пользователей [3]. Одновременно с ростом числа пользователей, росло и число Интернет ресурсов. По данным VeriSign Inc. к концу 2003 г. в мире всего было зарегистрировано 60 млн. доменных имен, что на 16% больше, чем год назад. Как и прежде самые популярные доменные области — это COM и NET, на их долю приходится 52% всех доменных имен. В абсолютных цифрах это более 30,4 млн. доменов, причем 1,7 млн. имен в этих двух зонах было зарегистрировано в последнем квартале 2003 г. [3].

Для ученых, наряду с традиционными источниками информации, представляют интерес сайты научной тематики. (Сайт — единая информационная структура, состоящая из связанных между собой документов — страниц [4]). Так, по данным рейтинга Rambler's Top100, 10,8% всех сайтов составляют сайты научной тематики, к которым мы отнесли сайты, посвященные образованию, технологии, фармацевтике, электронике и медицине. Для сравнения, сайты, посвященные Интернет-торговле, составляют всего 8,53%. В связи со столь внушительным объемом научной информации в глобальной сети встал вопрос об оценке ее достоверности [5–7].

Цель работы — определение критериев и создание методики оценки информативности и достоверности научных сайтов по узкотематическим направлениям.

2. СПОСОБЫ ПОИСКА ИНФОРМАЦИИ В ИНТЕРНЕТЕ

Существуют три основных способа поиска информации в Интернете.

1. Поиск сайтов и страниц через поисковые машины. Самая крупная и известная поисковая машина (или Поисковик) — Google имеет адрес в сети www.google.com. Она включает базу данных по 8 058 044 651 страниц (информация на 6 февраля 2005 г. 14:42 по московскому времени). Google вносит в свою базу данных сайты на разных языках, расположенные на различных географических доменах. В этом поисковике, существует множество языковых кластеров, например, английский www.google.com, немецкий www.google.de, итальянский www.google.it, японский www.google.jp, русский www.google.ru (www.google.com.ru) и др., поиск по каждому из которых ведется на языке той страны, чей национальный домен он занимает. Как и на других поисковиках в Google есть возможность вести "расширенный поиск". Следует отметить, что в поисковых машинах поиск осуществляется по описаниям страниц сайта, составленным роботом в автоматическом режиме.

2. Поиск через каталоги, имеющие разветвленную — "древовидную" структуру. В связи с большим объемом информации о сайтах, в каталогах подключен поисковый робот. В выдаче робота содержится ссылка на сайт, его название, описание и ссылка на раздел каталога. Таким образом, пользователь имеет возможность просмотреть не только сайты, описания которых соответствуют поисковому запросу, но и сайты, находящиеся в одном разделе с представленными в выдаче.

3. Поиск библиографической информации по узкотематическим направлениям, отраженным в БД, с одновременной выдачей по страницам сайтов, относящихся к рассматриваемой тематике. Поиск сайтов осуществляется специализированной поисковой машиной. В настоящее время такой поиск может быть проведен в БД SCOPUS компании Elsevier, где наряду с традиционным поиском библиографической информации осуществляется поиск сайтов с помощью специализированной поисковой машины, находящейся на платформе SCIRUS. Например, по приоритетному направлению "фуллерены" с 1991 г. по апрель 2005 г. отражено 18 658 публикаций из традиционных источников информации: журналов, трудов конференций, книг, отчетов; 2539 — патентов и 36 320 страниц сайтов.

3. ОПИСАНИЕ КАТАЛОГОВ

Наиболее известным каталогом в Интернете считается DMOZ, доступ к нему можно получить как по адресу www.dmoz.org, так и через поисковую машину Google, так как база данных "открытого каталога" DMOZ подключена к Google Directory. Отличием является способ поиска. DMOZ осуществляет поиск исключительно по названию и описанию сайта в каталоге, а Google Directory ищет сайты, подключая базу поисковой машины, т. е. учитывает в поиске полное содержание всех страниц сайта. Как поисковая машина Google, так и Google Directory имеют многоязычный пользовательский интерфейс. Регистрация сайтов в Open Directory Project (ODP) www.dmoz.org производится по стандартной схеме регистрации в каталогах. Владелец сайта предлагает модераторам OPD описание и раздел каталога. Однако окончательное решение принимает модератор.

Конкуренцию DMOZ и Coogle Directory составляет каталог Yahoo!. Так же как и в DMOZ в нем существует форма для подачи заявки на регистрацию сайтов и поиск ведется по названиям и каталожным описаниям. Часто выдачи каталогов Yahoo! и DMOZ, содержат одни и те же сайты. Но при этом ссылка на сайт может находиться в различных разделах каталогов. При поиске сайтов близкой тематики необходимо просматривать разные разделы каталогов.

Отдельно стоит каталог www.altavista.com компании Yahoo. Сайты в него вносятся только модераторами, при этом владельцы сайта не имеют возможности предлагать свои сайты к рассмотрению. В каталог внесены только самые крупные сайты по различным областям знания, содержащие полезную для наибольшего, по мнению модераторов, числа людей информацию. В связи с этим поиск сайтов по узкоспециализированным направлениям через этот каталог не принесет никаких результатов.

Каталог коммерческих сайтов — www.overture.com принадлежит поисковой системе Yahoo!. Размещение ссылок в этом каталоге — платное. Сайты, размещенные в коммерческом каталоге, попадают при поиске в выдачу Yahoo! — раздел "SPONSOR RESULTS". Каталог Altavista выдает первые две ссылки из www.overture.com, а далее — из своей БД. В случае же неудачного поиска по запросу в каталоге Altavista происходит переключение на сайт Overture.

4. ПОИСК САЙТОВ НАУЧНОЙ ТЕМАТИКИ

Поиск по сайтам открывает пред учеными возможность получать сведения об интересах научных групп и отдельных ученых и при этом наиболее оперативно иметь контактную информацию и, в ряде случаев, бесплатный доступ к полному тексту документов. В последнее время авторы все чаще размещают статьи в электронных изданиях и на сайтах. Поэтому при поиске по научно-техническим дисциплинам уже нельзя ограничиваться библиографической и/или полнотекстовой литературой без ущерба для "полноты" поиска.

В качестве примера сайтов узкой специализации рассмотрим сайты по приоритетному направлению "фуллерены".

Как и в обычном поиске по БД, в Интернет-поиске немаловажную роль играет поисковый запрос. При составлении Интернет-запроса нужно принять во внимание одно из важнейших свойств информации — *некоммутативность* (неперестановочность: суммарное количество полученной информации зависит от последовательности поступления (получения) информационных сообщений ($A+B \neq B+A$, где A и B — разные информационные сообщения)). В ряде случаев это свойство можно не учитывать, используя программные возможности "расширенного поиска" в поисковых машинах и каталогах так же, как и в БД. Так, например, выдача

по запросу "fullerene nanotubes" будет отличаться от выдачи по запросу "nanotubes fullerene".

При поиске по запросу (если не учитывать дополнительные, постоянно меняющиеся с развитием поисковой машины критерии, влияющие на позицию — ранг ссылки в выдаче.) следующим образом распределяются сайты, содержащие:

1) запрос целиком;

2) все слова запроса в том же порядке, в котором они находятся в запросе, например, — fullerene** nanotubes, где ** — произвольная фраза и чем она меньше, тем выше позиция в выдаче;

3) все слова запроса в произвольном порядке;

4) не все слова запроса, в выдаче сайт находится тем выше, чем больше слов запроса он содержит.

4.1. Анализ выдачи информации из поисковой машины Google

Для того чтобы оценить тип, информативность и достоверность научных сайтов узкой тематической направленности, проанализируем первые 20 ссылок из выдачи поисковой машины Google по запросу "fullerene".

1. http://sbchem.sunysb.edu/msl/fullerene.html
2. www.fullerene.com/
3. www.godunov.com/Bucky/Patents.html
4. www.chemistry.wustl.edu/~edudev/Fullerene/fullerene.html
5. www.chemistry.wustl.edu/~edudev/Fullerene/
6. www.sussex.ac.uk/Users/kroto/
7. www.sussex.ac.uk/Users/kroto/FullereneCentre/
8. www.mindspring.com/~kimall/ Fuller/
9. www.dekker.com/servlet/product/productid/FST
10. http://www.ifw-dresden.de/iff/14/forschg/fulleren/wassindfullerene/
11. www.fullereneinternational.com/
12. www.univie.ac.at/spectroscopy/
13. www.susx.ac.uk/Users/kroto/fullgallery.html
14. http://dc2.uni-bielefeld.de/dc2/fullerene/
15. http://en.wikipedia.org/wiki/Fullerene
16. www.mcfullerene.com/
17. www.geocities.com/upwardthrust/carbon/fullerene.html
18. www.nanoword.net/library/def/Fullerene.htm
19. www.worldofmolecules.com/materials/fullerene.htm
20. www.fullerene-jp.org/

Среди этих двадцати ссылок можно выделить ссылки на следующие сайты:

1) научных групп, содержащие информацию о работе ученого или научной группы;

2) семинаров и конференций;

3) институтов с общей информацией о проводимых в институте работах и кратким описанием той или иной научной тематики;

4) энциклопедические;

5) коммерческие — Интернет-магазины и сайты фирм, торгующих необходимыми оборудованием или материалами;

6) Интернет-издания — сайты Online — журналов или газет;

7) не имеющие отношения к исследуемой научной области, — "шум" для рассматриваемого научного направления.

Распределение сайтов из выдачи Google по запросу "fullerene" представлено на рис. 1.

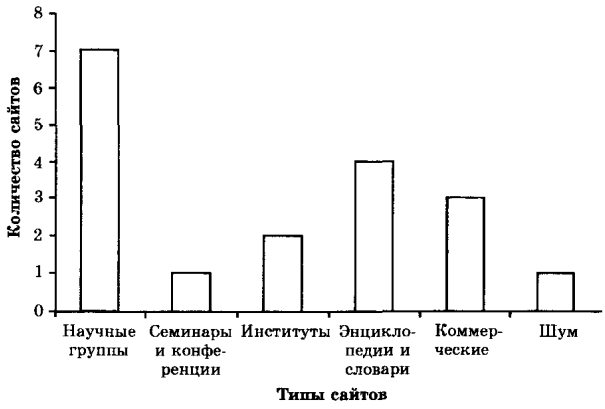

Рис. 1. Распределение по типам сайтов из выдачи Google по запросу "fullerene"

4.2. Анализ выдачи информации из каталога DMOZ

Более точную, но менее полную информацию, как было сказано выше, можно получить в каталогах. Рассмотрим выдачу каталога DMOZ по запросу "fullerene". В выдаче присутствует всего шестнадцать ссылок на сайты (в Google более 500 тыс. страниц).

1. http://www.geocities.com/kuku05/
2. http://www.mcfullerene.com/
3. http://smalley.rice.edu/
4. http://www.rsphysse.anu.edu.au/nanotube/awnf2001/index.htm
5. http://buckminster.physics.sunysb.edu/
6. http://www.uvm.edu/~dcloughe/
7. http://sciencenews.org/20000325/fobl.asp
8. http://www.sesres.com/
9. http://www.ciam.unibo.it/electrochem/
10. http://www.chem.ucdavis.edu/groups/balch/
11. http://gaus90.chem.yale.edu/henmr.html
12. http://www.diederich.chem.ethz.ch/
13. http://www.cchem.berkeley.edu/Ekpvgrp/research.html
14. http://www.mtr-ltd.com/ Science: Chemistry
15. http://www.mtr-ltd.com/ Regional: North America
16. http://www.nottingham.ac.uk/~ppzstm

Распределение по типам сайтов из выдачи открытого каталога DMOZ представлено на рис. 2.

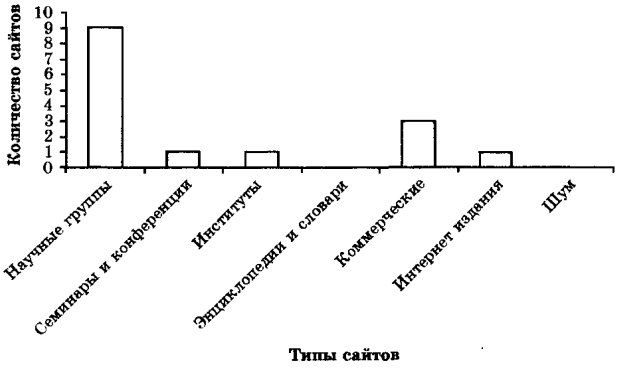

Рис. 2. Распределение сайтов из выдачи DMOZ по запросу "fullerene"

4.3. Анализ выдачи из БД SCOPUS

База данных SCOPUS подключена к поисковой машине SCIRUS издательства Elsevier. Для анализа мы взяли первые двадцать ссылок, представленные в выдаче этого поисковика по запросу "fullerene":

1. http://mozart.chem.nyu.edu/
2. http://www.rsphysse.anu.edu.au/ampl/research/rad/fullerene_genesis.html
3. http://alc41.riken.go.jp/lab/research/fullerene/index.html
4. http://www.uclm.es/dep/flanga/fullerene_chemistry.htm
5. http://www.pa.msu.edu/
6. http://www.ipc.uni-linz.ac.at
7. http://smalley.rice.edu
8. http://www.photon.t.u-tokyo.ac.jp/~maruyama/ fullmd/yokohama/yokohama.html
9. http://cscmr.snu.ac.kr/index_e.html
10. http://www.infochem.ethz.ch-шум
11. http://www.chemindustry.com/searches/B/buckminster_fullerene.html
12. http://www.fastlane.nsf.gov/servlet/showaward?award=0316078
13. http://www.psychcentral.com/wiki/Fullerene
14. http://www.columbia.edu/
15. http://pubs.acs.org/cgi-bin/jcen?jacast/asap/html/ja026069j.html
16. http://ftp.mathe2.uni-bayreuth.de/frib/html2/fullerene/full1_10.html
17. http://search.psychcentral.com/wiki/Fullerene
18. http://www.chemistry.wustl.edu/~edudev/Fullerene/
19. http://www.nasatech.com/Briefs/Feb01/NPO20148.html
20. http://www-lab.imr.edu/~fulltube/info.html

Из распределения по типам сайтов видно, что больше всего в выдаче представлены сайты отдельных ученых и институтов, причем сайтов институтов больше, чем сайтов отдельных научных групп (рис. 3).

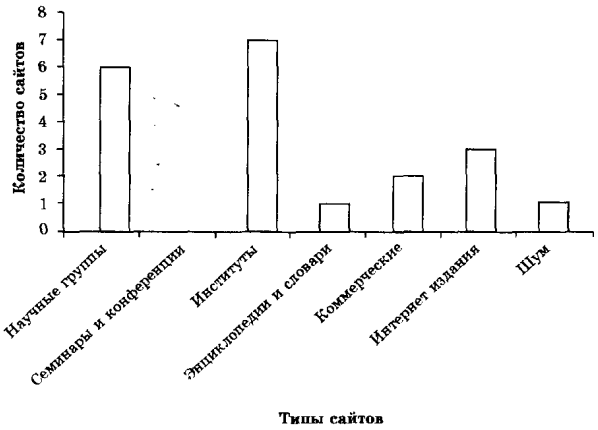

Рис. 3. Распределение по типам сайтов поисковой машины SCOPUS

4.4. Сравнение выдач различных поисковых ресурсов

Как видно из рис. 1–3, в сумме всех трех выдач большая часть сайтов — это сайты отдельных ученых и научных групп. Некоторые из этих сайтов находятся на серверах университетов, в которых работают ученые, что может являться одним из показателей достоверности информации.

Сопоставление выдач из различных поисковых ресурсов показало, что, в отличие от выдачи поисковой машины, в выдачу каталога очень мало вероятно попадание "шума" по рассматриваемой тематике. Кроме того, здесь нет сайтов энциклопедий и словарей, так как их каталожные описания не содержат термин "fullerene", но представлен раздел — "Интернет-издания" в который мы включили сайт, содержащий множество ссылок на сайты, посвященные фуллеренам (рис. 4).

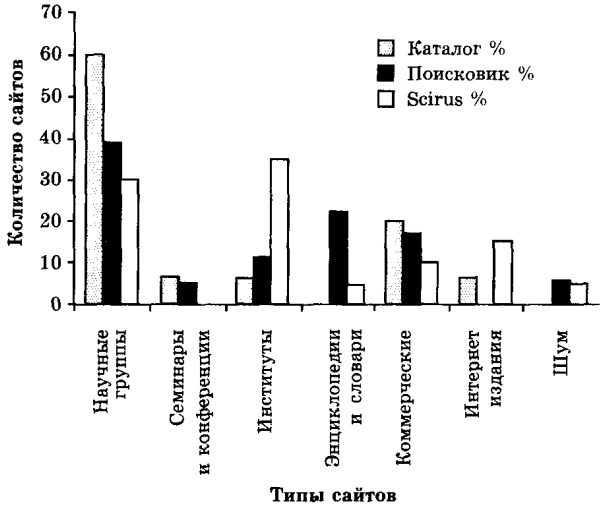

Рис. 4. Соотношение типов сайтов в выдачах каталога DMOZ, поисковой машины Google и Scirus

5. ОПРЕДЕЛЕНИЕ ИНФОРМАТИВНОСТИ И ДОСТОВЕРНОСТИ САЙТОВ

Рассмотрим критерии оценки, выбранные для самой значительной части выдач — "сайтов научных групп":

1. *Список статей.* На сайтах часто можно встретить информацию о статьях, опубликованных научной группой.

2. *Полные тексты статей.* Сайты многих крупных научных групп содержат полный текст опубликованных ими статей, что позволяет серьезно облегчить поиск информации.

3. *Обновляемость.* Этот фактор никак не влияет на достоверность информации, представленной в статьях, расположенных на сайте. Однако по обновляемости можно судить о работе научной группы и об ее интересе к представленной на сайте области исследований.

4. *Ссылка на сайты из политематических БД* (в списках цитируемых публикаций), что является одним из важнейших критериев оценки их информативности и достоверности.

5. *Обратная связь.* Наличие координат для связи с разработчиками и владельцами сайта также может быть критерием для оценки его информативности и достоверности.

Для сайтов ученых и научных групп представлена "оценочная таблица" из выдачи каталога DMOZ, поисковой машины Google и поисковой машины Scirus.

В случае соответствия сайта тому или иному критерию, полю присваивается значение — "1", в противном случае — "0". По сумме этих значений оценивается **степень информативности** сайта: чем она выше, тем более полезным для ученого может быть сайт и тем выше его **достоверность**. Из таблицы видно, что сайты, занесенные в каталог модераторами и представленные выдачей DMOZ, имеют большую степень информативности, т. е. содержат более полную и достоверную информацию.

Электронные ресурсы по «фуллеренам», представленные на сайтах ученых и научных групп

Электронные ресурсы	Полный текст статей	Список статей	Обновляемость	Ссылка из SCI	Обратная связь	Степень информативности
1. DMOZ						
http://www.geocities.com/kuku05/	0	1	1	–	1	3
http://smalley.rice.edu/	1	1	1	–	1	4
http://buckminster.physics.sunysb.edu/	1	1	0	–	1	3
http://www.uvm.edu/~dcloughe/	1	1	1	–	1	4
http://www.ciam.unibo.it/electrochem/	0	1	1	–	1	3
http://www.chem.ucdavis.edu/groups/balch/	0	1	0	–	1	2
http://gaus90.chem.yale.edu/henmr.html	0	0	0	–	1	1
http://www.diederich.chem.ethz.ch/	0	1	1	–	1	3
http://www.cchem.berkeley.edu/%7Ekpvgrp/research.html	0	1	1	–	1	3
2. Google						
http://sbchem.sunysb.edu/msl/fullerene.html	0	1	0	–	0	1
www.godunov.com/Bucky/Patents.html	0	1	0	–	1	2
www.sussex.ac.uk/Users/kroto/FullereneCentre	0	1	0	–	1	2
www.mindspring.com/~kimall/Fuller/	0	0	1		1	2
http://www.ifw-dresden.de/iff/14/forschg/fulleren/wassindfullerene/	1	1	0	–	1	3
www.univie.ac.at/spectroscopy/	0	1	1	–	1	3
http://dc2.uni-bielefeld.de/dc2/fullerene/	0	1	1	–	0	2
3. Scirus						
http://mozart.chem.nyu.edu/	0	1	1		1	3
http://alc41.riken.go.jp/lab/research/fulerene/index.html	0	0	1		1	2
http://www.uclm.es/dep/flanga/fullerene chemistry.htm	0	1	0		0	1
http://smalley.rice.edu	1	1	1		1	4
http://www.photon.t.u-tokyo.ac.jp/	0	0	0		1	1
http://cscmr.snu.ac.kr/index_e.html	1	1	1		1	4

ВЫВОДЫ

1. Анализ динамики развития Интернета выявил увеличение числа научных сайтов.

2. Описаны принципы работы основных поисковых машин и каталогов. Показана важность предварительного поиска информации в глобальной сети.

3. Рассмотрен ряд возможностей поиска информации в Интернете по одному из приоритетных научных направлений — "фуллерены".

4. Предложен метод оценки информативности и достоверности сайтов, который показал хорошие результаты при анализе сайтов отдельных ученых и научных групп.

СПИСОК ЛИТЕРАТУРЫ

1. Пек С., Аррантс С. Web-сервер WebSite: пер. с англ.— К.: Изд. Группа BHV, — 1997.— 344 с.

2. Clinton W. J., Gore A. Technology for America's Economic Growth, a New Direction to Build: Executive Office of the President.— Washington D. C., 1993.— 39 p.

3. Доклад "Статистика развития российского сегмента Интернета". Материал предоставлен RU-Center, при подготовке статьи использованы данные ICANN, Минсвязи РФ, РосНИИРОС, RU-CENTER, ФОМ, Nielsen/NetRating.— info.nic.ru

4. Леонтьев В. П. Новейшая энциклопедия Интернет 2003.— М.: "Олама-Пресс", 2003.— 781 с.