Попытки объять необъятное, или World Wide Web под прицелом

Дмитрий ЛАНДЭ,
Андрей СНАРСКИЙ

Особенности web-пространства

Сеть Интернет была создана более 30 лет тому назад в рамках проекта ARPANET, став в настоящее время крупнейшей информационной магистралью. Надстроенная поверх узлов Интернет сеть вэб-сайтов, в свою очередь, стала крупнейшим феноменом информационных технологий, мощнейшим за всю историю человечества информационным ресурсом. Он содержит свыше 20 млрд. документов, размещенных более чем на 120 млн. серверах (**рис. 1**, статистика сайта www.netcraft.com).

В свою очередь, WWW стала базой для построения многочисленных ***подсетей, малых миров*** (Small Worlds), многие из которых по объемам информации превышают объемы сети WWW трех-пятилетней давности. По-видимому, причины резкого роста объемов и динамики информации в Сети обусловлены тем, что если в начале ее существования небольшое количество вэб-сайтов публиковало информацию немногих авторов для относительно большого количества посетителей,

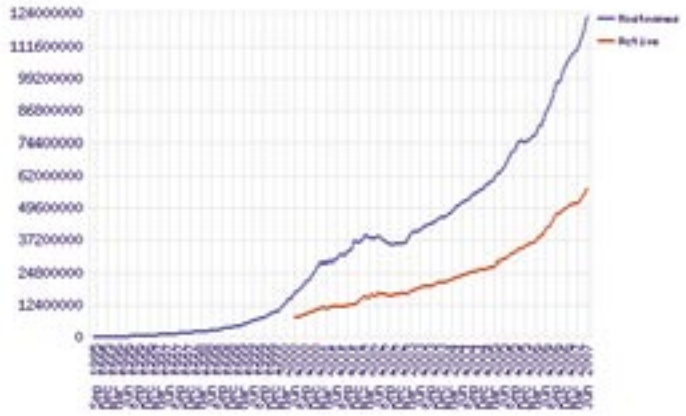

Рис. 1. На сегодняшний день в Интернете насчитывается свыше 120 млн. серверов (hostnames — количество имен серверов в WWW, из них на момент проверки около половины активных — active)

то сегодня Сеть «поддержали массы»: посетители вэб-сайтов сами активно участвуют в создании контента. Т.е. произошел качественный скачок от сети распространения к сети публикации информации.

Моделирование структуры WWW

Стоит отметить, что как сама сеть WWW, так и ее отдельные фрагменты и даже сайты несут значительную социальную нагрузку. Поэтому их можно сравнивать на содержательном уровне с «сетями» человеческих отношений или цитирования в науке.

Web-пространство характеризуется большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок, поэтому его можно с полным правом считать социальной сетью, исследование которой можно проводить, базируясь на существующем подходе анализа таких сетей — SNA (Social Network Analysis). Многие сетевые службы, позволяющие людям устанавливать связи в Сети, автоматически формируют социальные сети. Само понятие «социальная сеть» появилось уже давно, его ввели в употребление английские социологи еще в 50-х годах XX века. В качестве узлов таких сетей стали рассматривать не только представителей социума, но и другие объекты, которым присущи социальные связи.

Поскольку стало понятно, что WWW — тоже социальная сеть, к ней оказалось возможным применить некоторые стандартные процедуры, которые позволяют понять, с одной стороны, логику развития этой сети, а с другой — некоторые феномены, отличающие ее от обычных сетей, например, транспортной.

В анализе любых сетей главная задача заключается в ***выявлении сетевых подструктур или клик.*** Клики — это подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик. Одна из первых моделей сети появилась в результате исследования группы ученых из Бирмингема, которые написали программу, отображающую реальные гиперссылки в виде трехмерной схемы. Клики, блоки, группировки, перемычки стали видны сразу же, как и в любой другой социальной сети. Первая модель, позволившая выявить подструктуры web-пространства, была построена в 1995 го-ду. По адресу http://www.igd.fhg.de/archive/1995_www95/proceedings/posters/35/index.html находится отчет с результатами моделирования (один из примеров приведен на **рис. 2**), которые могут показаться наивными на сегодняшний взгляд. Да и моделированием то, что представлено, назвать трудно. Скорее, это «отпечаток» состояния Сети на момент снятия ее карты.

Из множества рисунков, отображающих структуру Сети на тот момент, видно, что некоторые ее части образуют компактные группы (**рис. 3**), а отдельные группы соединены между собой «дальними» связями (**рис. 4**). Именно такие группы формируют «малые миры», речь о которых будет идти ниже.

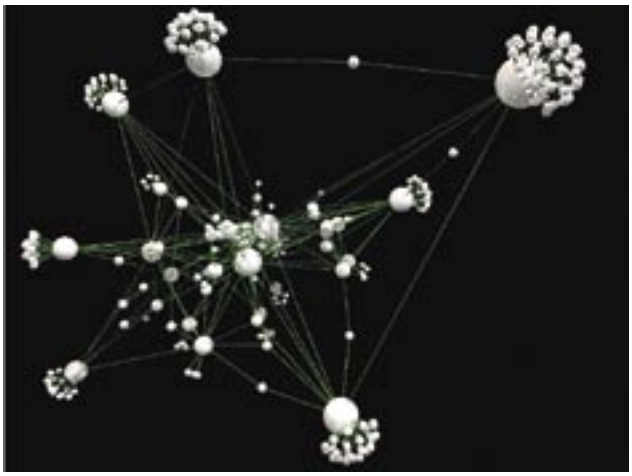

Рис. 2. Модель фрагмента вэб-пространства «по-бирмингемски»; размер узлов пропорционален количеству исходящих связей

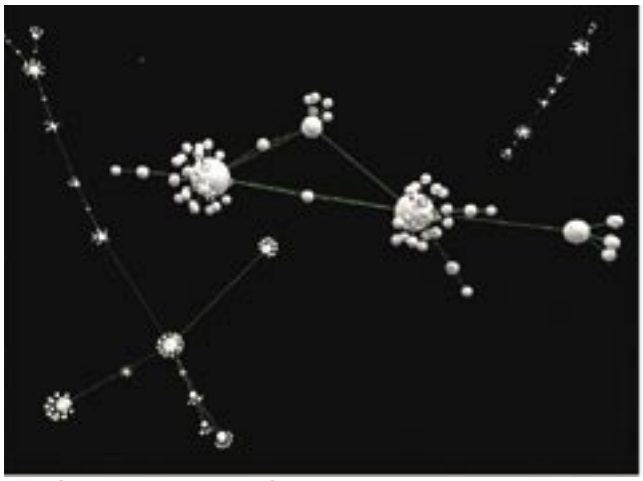

Рис. 3. Некоторые фрагменты Сети формируют компактные группы

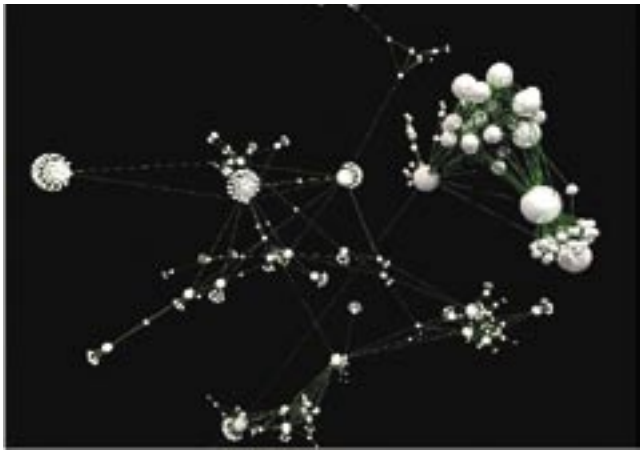

Рис. 4. Отдельные группы, соединенные между собой «дальними» связями, образуют «малые миры»

Ученые удивились компактности отдельных подмножеств сетей, но сформулировать этот эффект смогли только на языке «слабых связей» и «малых миров».

Кластеризация и семантические карты

Но не все так просто. Связи между объектами в любой социальной сети бывают скрытыми, тайными или латентными. Учет таких неочевидных связей и группировка объектов по ним — удел аналитиков, а на помощь им приходят специальные методы. Один из таких методов группировки на основе неочевид-

Рис. 5. Карта понятий, соответствующая запросу «семантический web»

Рис. 6. «Джин» выполняет кластеризацию вэб-пространства

ных связей — кластеризация. Различные сети по-разному поддаются кластеризации. В 1998 году исследователи из Корнелльского университета (США) Д. Уаттс и С. Строгатц даже ввели такую характеристику сетей, как ***коэффициент кластеризации***, который соответствует уровню связности узлов в сети.

Читатель, имеющий дело с поисковыми системами, конечно же, знаком с примерами кластеризации. Многие современные поисковики группируют свои ответы на запросы, предоставляя пользователям так называемые «информационные портреты», или «семантические карты», соответствующие запросам. Первой такой известной системой в свое время стала Vivisimo, больших успехов добилась российская ***мультипоисковая система Nigma*** (http://www.nigma.ru), реализовав специальный AJAX-интерфейс (кстати, Nigma — это один из трех родов пауков семейства Dictunidae).

В результате запросов поисковая система выдает не только традиционный набор результатов в виде гиперссылок, но и карту (**рис. 5**) близких, семантически связанных, понятий.

Если вернуться к структуре вэб-пространства, то процесс кластеризации серверов при указании первого, гипертекстовые связи которого исследуются, наглядно демонстрирует ***систему KartOO*** (**рис. 6**, http://www.kartoo.com/). К сожалению, этот сервис, реализованный на флеш-технологиях, не адаптирован к кириллическим шрифтам.

Эластичность и перколяция

Еще один очень интересный параметр сети — ее ***эластичность***. Это свойство как бы отвечает на вопрос: что же будет с сетью, если из нее удалить некоторые узлы или, наоборот, если некоторые узлы добавить. Как изменится расстояние между другими узлами, нарушится ли связность? Для большинства сетей, если узлы из них будут удаляться, длина путей между остальными узлами будет увеличиваться, и, в конечном счете, связность сетей нарушится.

Нарушение связности сетей — это проблема безопасности, для решения которой мобилизованы огромные научные коллективы.

Исследования, недавно проведенные американскими учеными, показали, что сеть WWW обладает достаточно высокой эластичностью по отношению к удалению (отказу) случайных узлов (сайтов), но высокочувствительна к преднамеренной атаке на сайты с высокими степенями связей с другими сайтами.

При изучении свойств WWW как социальной сети в плане ее безопасности и устойчивости оказался интересен подход, логически связанный с понятием ***перколяции*** (протекания), популярным в современной физике. Оказывается, что многие вопросы, возникающие при анализе структуры Интернета, имеют прямое отношение к теории перколяции. Ведь протекание или просачивание в любой физической среде в некотором смысле эквивалентно целостности гиперсвязей в Интернете. При нарушении таких связей, например, сайт или целый «малый мир» могут стать недоступными для индексирования поисковыми системами, уйти в раздел «скрытого» (invisible, deep) Web.

Перед теорией перколяции стоит множество задач, самая важная из которых имеет следующий вид. «Дана решетка из связей, случайная часть которой проводит сигнал (воздух, ток, информацию...),

а остальная часть его не проводит. Вопрос: чему равна минимальная концентрация проводящих связей, при которой еще существует путь через всю решетку?».

В настоящее время известно много важных обобщений перколяционной задачи, например, рассматриваются случаи, когда «непроводящие» связи проводят, но много хуже проводящих (в Интернете это может быть связано с пропускной способностью каналов к отдельным сайтам или возможностью сайтов адекватно реагировать на множественные запросы); можно говорить о различных значениях проводимостей для разных связей; можно рассматривать однонаправленные «диодные» связи (большинство связей, реализованных гиперссылками в WWW, именно такие) и т.п.

К задачам, решаемым в рамках теории перколяции для анализа стабильности сетей, относятся такие, как определение порогового уровня проводимости (пропускной способности), изменения длины пути и его траектории (извилистости, запараллеленности) при приближении к пороговому уровню проводимости, количества узлов (сайтов), которое необходимо вывести из строя, чтобы нарушить связность Сети.

Применение перколяционного подхода, в частности, представили ученые из Стенфордского университета. Они разработали простой алгоритм случайного поиска для пиринговых сетей по принципу «пчелиного роя». В предложенной системе каждый узел переправлял полученный поисковый запрос дальше по сети, причем по одному случайно выбранному адресу. Алгоритм, разработанный в Калифорнийском университете, делает то же самое, только параллельно. Он использует принцип порога перколяции связей, т.е. порога протекания связей между тесно связанными узлами. На этапе перколяции связей запрос попадает на один из базовых серверов Сети, которые соединены друг с другом мощными каналами связи. Американские ученые обнаружили, что полноценный процесс поиска может проводить «локально», т. е. при опросе только соседних серверов. При таком методе каждый запрос генерирует относительно малый трафик, объем которого растет медленнее, чем вся сеть в целом.

Таким образом, перенесенные из мира физики понятия эластичности и перколяции оказались фундаментальными при исследовании такой быстрорастущей сети, как World Wide Web. С одной стороны, эти понятия дают объяснения некоторых эффектов, возникающих в процессе эволюции Сети, а с другой — предоставляют эффективные алгоритмы поиска и навигации в ней.

«Слабые связи» и «малые миры»

По отношению к некоторым социальным сетям справедлива модель «слабых связей». Если в обществе аналогом «сильных связей» можно считать отношения людей с родственниками или сослуживцами, то аналогом слабых связей являются, например, отношения с дальними знакомыми и коллегами. В некоторых случаях такие связи оказываются более эффективными, чем связи «сильные». Так, в области мобильной связи группой ученых из Великобритании, США и Венгрии был получен концептуальный вывод, что «слабые» социальные связи между индивидуумами оказываются самыми важными для существования социальной сети.

Для исследования были проанализированы звонки 4,6 млн. абонентов мобильной связи, что составляет около 20% населения средней европейской страны. В сети было выявлено 7 млн. социальных связей, то есть взаимных звонков от одного абонента другому и обратно (если обратные звонки были сделаны в течение 18 недель). Частота и длительность разговоров использовалась для того, чтобы определить силу каждой социальной связи.

Именно слабые социальные связи (один-два обратных звонка) связывают воедино большую социальную сеть (**рис. 7а**). Если эти связи убрать, то сеть распадется на отдельные фрагменты (**рис. 7б**). Если же убрать сильные связи, то ничего страшного с сетью не произойдет — она останется единой (**рис. 7в**).

На основании проведенных исследований ученые сделали вывод, что именно слабые связи являются тем феноменом, который связывает большое общество в единое целое. Надо полагать, что данный вывод

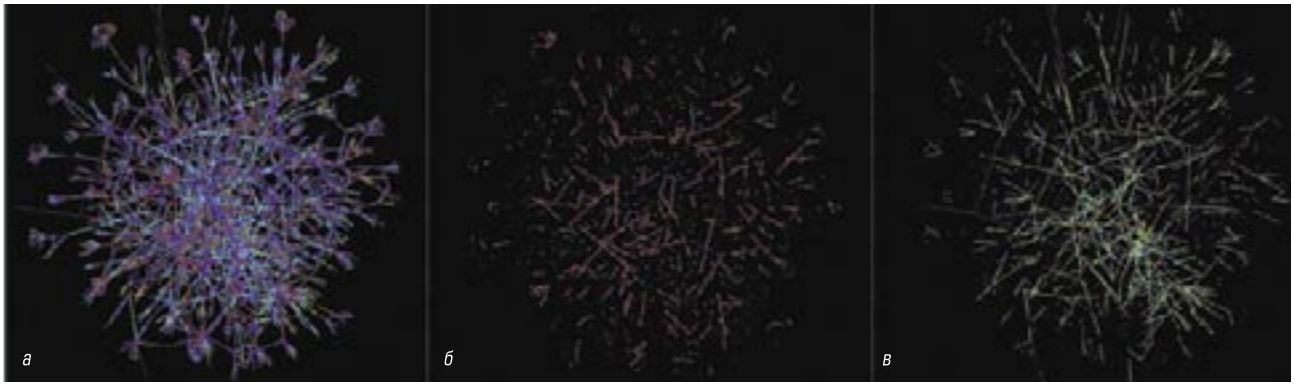

Рис. 7. Именно дальние, «слабые», связи обеспечивают цельность сети:
а — полная карта сети социальных коммуникаций;
б — социальная сеть, из которой удалены слабые связи, разбивается на множество изолированных участков;
в — карта сети, из которой удалены сильные связи: структура сохраняет сквозную проводимость

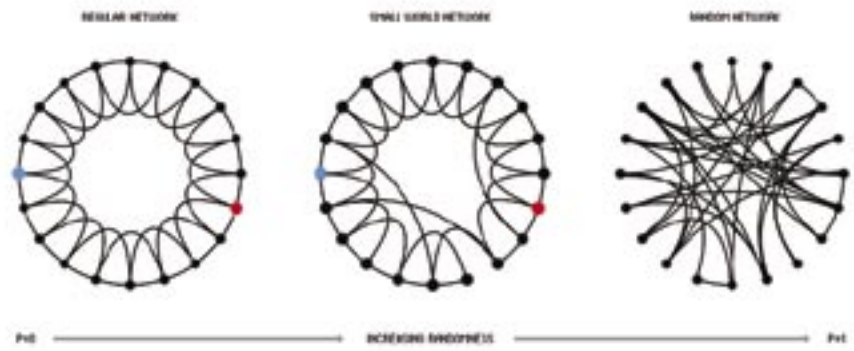

Рис. 8. Модель Уаттса и Строгатца, поясняющая эволюцию сети: регулярная сеть — сеть «малых миров» — случайная сеть в зависимости от количества замен «ближних» связей «дальними»

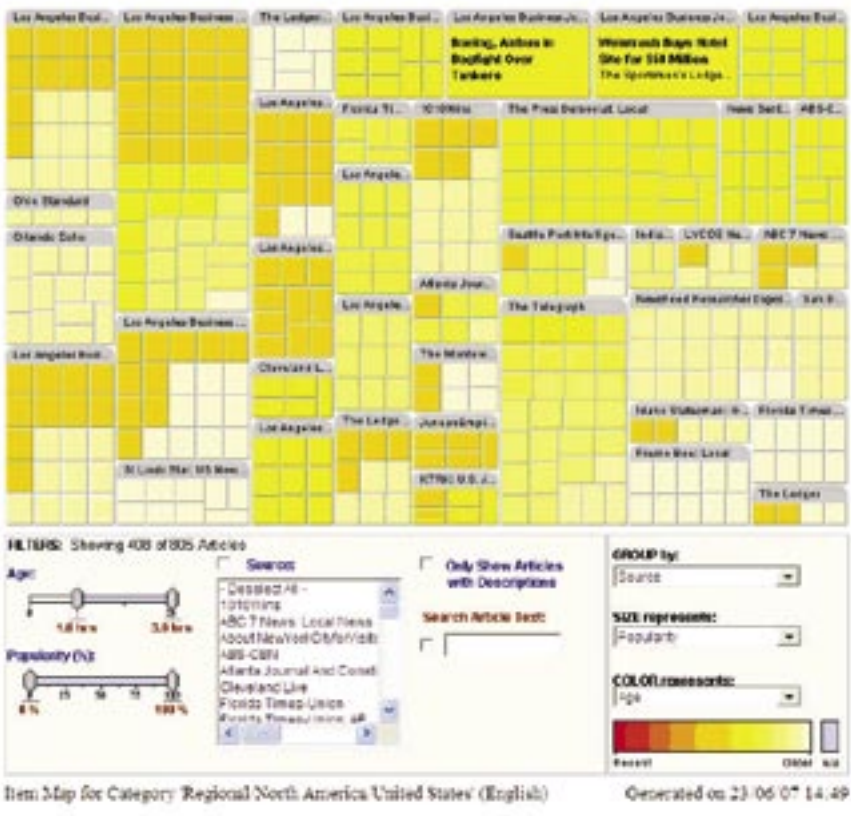

Рис. 9. Карта новостного Web от News Is Free

справедлив и для Web, хотя академических исследований в этой области до сих пор не проводилось.

Именно «слабые связи» объяснили такой феномен WWW, как появление «малых миров», который базируется на наблюдении психолога С. Милграна, что существует цепочка знакомств, в среднем длиной, равной шести, практически между двумя любыми гражданами США.

Эффект «малых миров» наглядно демонстрируется следующей процедурой, представленной Д. Уаттсом и С. Строгатцом. Рассматриваются три состояния сети (**рис. 8**): регулярная сеть, каждый узел которой соединен с четырьмя соседними, та же сеть, у которой некоторые «ближние» (сильные) связи случайным образом заменены «дальними» (слабыми) связями (именно в этом случае возникает феномен «малых миров»), и случайная сеть, когда количество подобных замен превысило некий порог. Как оказалось, именно те сети, узлы которых имеют одновременно некоторое количество локальных и «дальних» связей, демонстрируют и эффект малого мира, и большой уровень кластеризации.

Именно к таким сетям относится WWW, для которой подтвержден феномен малых миров. Об свидетельствует исследование Ши Жоу и Рауля Дж. Мондрагона из Лондонского университета. Проведенный ими анализ топологии Сети показал, что большие узлы имеют больше связей между собой, чем с малыми узлами, тогда как малые узлы имеют больше связей с большими узлами, чем между собой. Ученые назвали этот феномен «клубом богатых» (rich-club phenomenon). Исследование показало, что 27% всех соединений имеют место между всего 5% крупнейших узлов, 60% приходится на соединения остальных 95% узлов с 5% крупнейших и только 13% — соединения между узлами, которые не входят в лидирующие 5%.

Эти исследования дают основания полагать, что зависимость WWW от больших узлов существенно выше, чем предполагалось ранее, т.е. Сеть еще более уязвима в отношении злонамеренных атак. С концепцией «малых миров» связан практический подход, именуемый «сетевой мобилизацией», реализуемый над структурой «малых миров». В частности, скорость распространения информации благодаря эффекту «малых миров» возрастает на порядки, по сравнению с теоретически возможной. Экспертами по безопасности эффект «малых миров» в последнее время все чаще связывается с сетями террористических организаций, надстроенных поверх Интернета.

Агрегация новостей в Интернете

В Интернете существуют сотни малых миров, к которым можно отнести форумы, общения по интересам, блоги, сообщества профессиональных сайтов и т.п. Всех их объединяют «слабые связи» с каталогами, поисковыми системами или просто с отдельными участниками Сети. Рассмотрим более подробно один из «малых миров», а именно мир новостей, публикуемых в WWW. К этому миру можно отнести сайты информационных агентств, онлайновых СМИ, агрегаторов новостей, новостные разделы сайтов государственных учреждений и т.п.

В отличие от существующих моделей web-пространства, при анализе социальной сети, образованной новостным Web, приходится учитывать чрезвычайно высокую

динамику информационных потоков, контекстные (а не только гипертекстовые) ссылки, эффект содержательного дублирования. Стандартный подход к анализу web-пространства в случае «мира новостей» нельзя считать корректным по ряду причин. Среди них можно назвать, например, то, что в новостных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части web-пространства (это зачастую могут быть ссылки на ресурсы, доступные только по паролю, или даже оффлайновые публикации изданий, возможно, и присутствующих в Интернете). При построении модели структуры новостного web-пространства наибольшее внимание должно оказываться именно web-сайтам, на которых публикуются новостные сообщения, а не отдельным web-страницам или самим сообщениям.

Состояние новостного web-пространства в виде ссылок на источники и отдельные сообщения демонстрирует «карта новостей» (News Map), представленная на сайте агрегатора новостей News Is Free (http://newsisfree.com). В этом интерфейсе (**рис. 9**) учитывается два основных параметра отображения — ранг популярности и «свежесть» информации. Карта новостей представляет собой своеобразный пульт управления информацией, инструмент визуализации не столько самих новостей, сколько информационных предпочтений пользователей.

Фильтр, отбирающий новости, «по умолчанию» настроен на отбор только тех сообщений, со времени публикации которых прошло не более двух часов. Таких сообщений обычно немного, и поэтому картина получается достаточно наглядной. В рамках этой модели можно наблюдать «дробление» групп источников при увеличении ранга популярности и «свежести» изданий. Когда этот ранг становится достаточно высоким, дробление уже не позволяет без особых усилий читать названия источников и идентифицировать отдельные документы.

Инструменты визуализации вэб-пространства

Одним из направлений анализа социальных сетей является их визуализация. В качестве примеров визуализации сетей рассмотрим некоторые разработки компании TouchGraph. Так, например, TouchGraph Amazon отображает сеть, порожденную книгами и связями между ними (тематиками, авторами, издательствами).

Одним из самых динамичных новостных ресурсов Интернета сегодня можно считать так называемые живые журналы, или блоги. Компания TouchGraph, в частности, реализовала интерфейс для построения социосетей на основе Livejournal — TouchGraph LiveJournal Browser.

В случае визуализации WWW средствами TouchGraph Google Browser (http://www.touchgraph.com/TGGoogleBrowser.html) ребрами выступают не гипер-ссылки, а отношения подобия. Google Browser, представляющий собой Java-апплет, позволяет визуализировать связи подобия между веб-сайтами, рассчитывающиеся в поисковой системе Google. В этом интерфейсе можно увидеть все сайты, связанные отношением подобия с исходным заданным. При этом пользователь может задавать глубину связей и отображать взаимосвязи различных сайтов **(рис. 10).** TouchGraph Google Browser — весьма полезный инструмент также при поиске сайтов, связанных с исходным общей тематикой.

Рис. 10. Карта сайтов, подобных выбранному

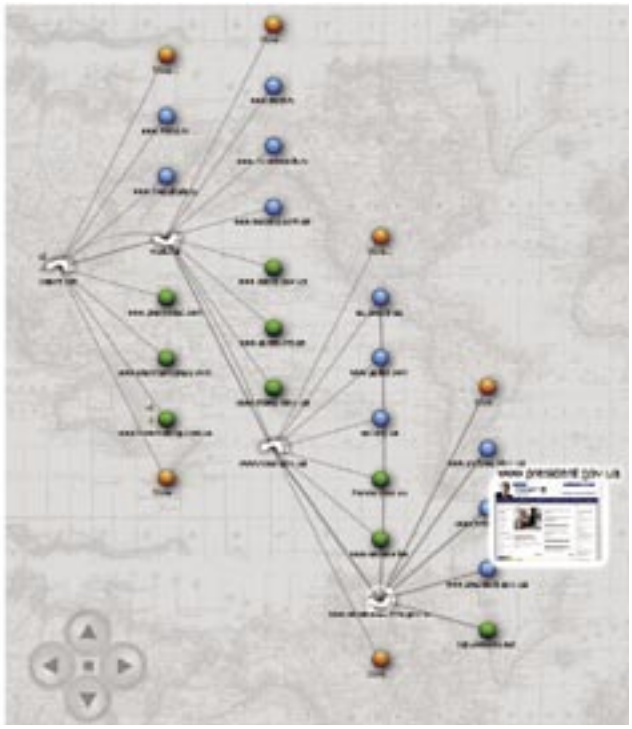

Рис. 11. Сеть как игра — результат выполнения программы Walk2Web

В качестве еще одного инструмента для анализа web-пространства можно привести программу Walk2Web (http://www.walk2web.com/), бесплатный онлайн-сервис, поддерживающий кириллическую кодировку (**рис. 11**).

В этой связи можно упомянуть о некогда очень популярных вэб-кольцах (web-rings), «сообществах» сайтов одной тематики, связанных между собой гиперссылками. Каждый сайт, участник вэб-кольца, имеет ссылку на

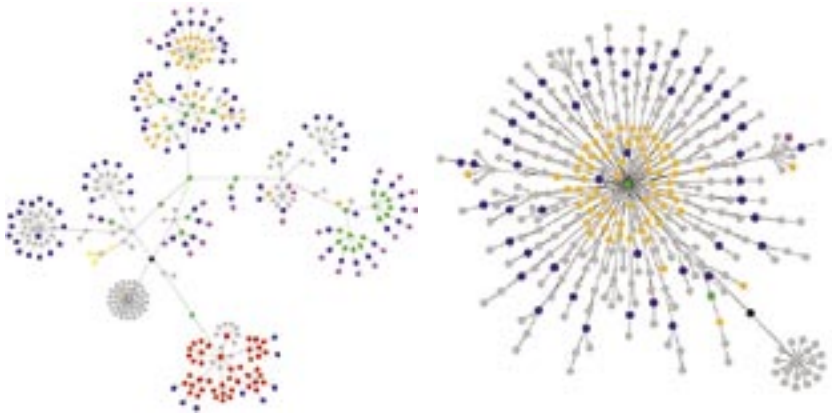

Рис. 12. Примеры графов — виртуальных цветов из «сада» Интернет

следующий сайт в кольце, позволяя посетителям совершать целевой тур по сайтам участников. К примеру, на идеологии web-колец, предположительно не связанных с другими частями Интернета, базировались разработчики так называемых детских вэб-браузеров, домашними страницами которых были сайты типа портала Уолта Диснея. Основная идея заключалась в том, что дети, заходя в начале на такие сайты и перемещаясь в дальнейшем только по гиперссылкам (ввод адресов был умышленно запрещен), не выйдут за пределы сайтов детской тематики. Реальность «малых миров» развеяла эти иллюзии. Однако некоторые проекты, такие как My Kids Browser (http://www.mykidsbrowser.com/), успешно развиваются благодаря преференциям иного рода.

Еще один интересный, но, к сожалению, мало применимый на практике инструмент, посвященный анализу Сети, представлен на сайте Websites as Graphs (http://www.aharef.info/static/htmlgraph/). На этом ресурсе с помощью технологии Java реализуется своеобразная «игра в жизнь». Достаточно ввести адрес исходного сайта, как буквально на глазах пользователя начинает разрастаться граф, разноцветные узлы которого — сайты, а ребра — гиперсвязи. Как и процесс обхода сайтов, картинки также формируются динамически, вплоть до достижения некоторого предельного уровня разрешения. На этом же ресурсе эстеты могут полюбоваться галереей **(рис. 12)** уже полученных графов (http://flickr.com/photos/tags/websitesasgraphs/), а «ботаники» —попытаться их классифицировать ☺.

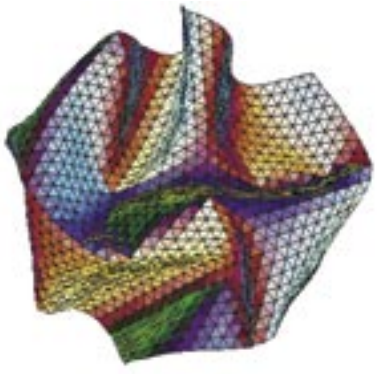

Рис. 13. Модель Бйорнеборнма — «мятый Web»

Исследователей заинтересует то, что на этом ресурсе приведены исходные коды применяемых апплетов.

Путешествие по Интернету в рамках этого сервиса превращается в своеобразную визуальную игру, на каждом шагу по которой пользователю открываются все новые сайты.

Сегодня весьма успешно изучаются масштабируемые, статические, иерархические, самосинтезирующиеся, «малые миры» и другие сети, исследуются их фундаментальные свойства, такие как устойчивость к деформациям и перколяции. В частности, недавно физики из университета Гранады доказали, что наибольшую информационную проводимость имеет особый класс сетей, названных «запутанными» (entangled networks). Они характеризуются максимальной однородностью, минимальным расстоянием между любыми двумя узлами и очень узким спектром основных статистических параметров. Ученые-теоретики полагают, что запутанные сети могут найти широкое применение в области информационных технологий, в частности, в новых поколениях WWW, позволяя существенно снизить объемы сетевого трафика.

Приходится констатировать, что ученые все еще далеки от полного понимания структуры сложных сетей и процессов, происходящих в них. Пожалуй, лучшим отражением существующего положения с изучением топологии web-пространства является модель «мятого Web». Ее предложил датчанин Леннарт Бйорнеборн. «Мятый Web» в этой модели ассоциируется с мятой бумагой (**рис. 13**). При этом путь между выбранными точками на мятой бумаге зачастую короче, поскольку противоположные части листа бумаги соединены. В соответствии с этой моделью каждая новая гиперссылка изменяет все существующие связи, создавая новые деформации «мятой» сети. Т.е. каждая новая гиперсвязь — «крючок», растягивающий или деформирующий форму существующей сети WWW. Вместе с тем, исследования в области изучения топологии WWW приобретают все большее значение — как для науки, так и для практики. Экспоненциальный рост объемов информации в Сети сегодня ставит перед пользователями вопрос о рациональности ее использования. Информационный шум, спам, дублирование снижают эффективность работы в Интернете без привлечения интеллектуальных программных средств, которые должны решить проблемы навигации и поиска, выхода на действительно необходимую и достоверную информацию. Очевидно, для этого необходимы глубокие знания о структуре WWW. И кто знает, сможет ли человек постичь структуру Сети, им созданной?

Дмитрий ЛАНДЭ,
д.т.н., заместитель директора Информационного центра «ЭЛВИСТИ»

Андрей СНАРСКИЙ,
д.ф.-м.н., профессор НТУУ «КПИ»