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,

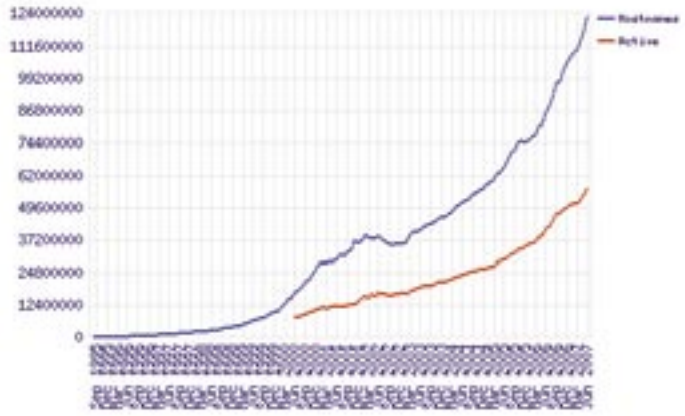


Рис. 1. На сегодняшний день в Интернете насчитывается свыше 120 млн. серверов (hostnames — количество имен серверов в WWW, из них на момент проверки около половины активных — active)

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.

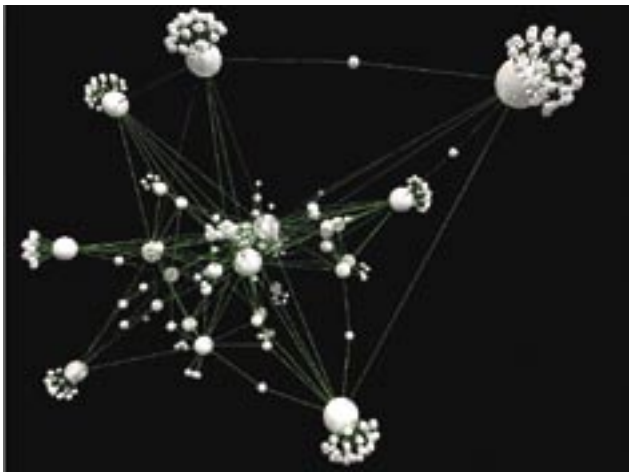


Рис. 2. Модель фрагмента вэб-пространства «по-бирмингемски»; размер узлов пропорционален количеству исходящих связей

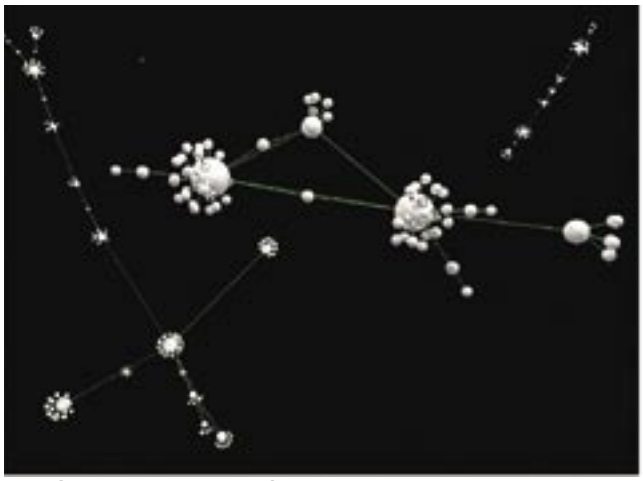


Рис. 3. Некоторые фрагменты Сети формируют компактные группы

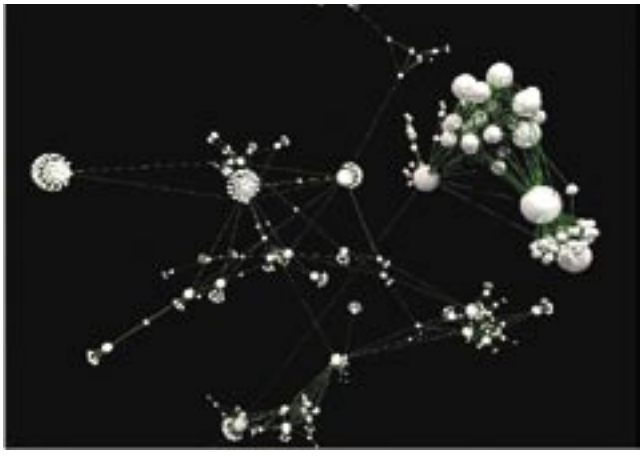


Рис. 4. Отдельные группы, соединенные между собой «дальними» связями, образуют «малые миры»

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

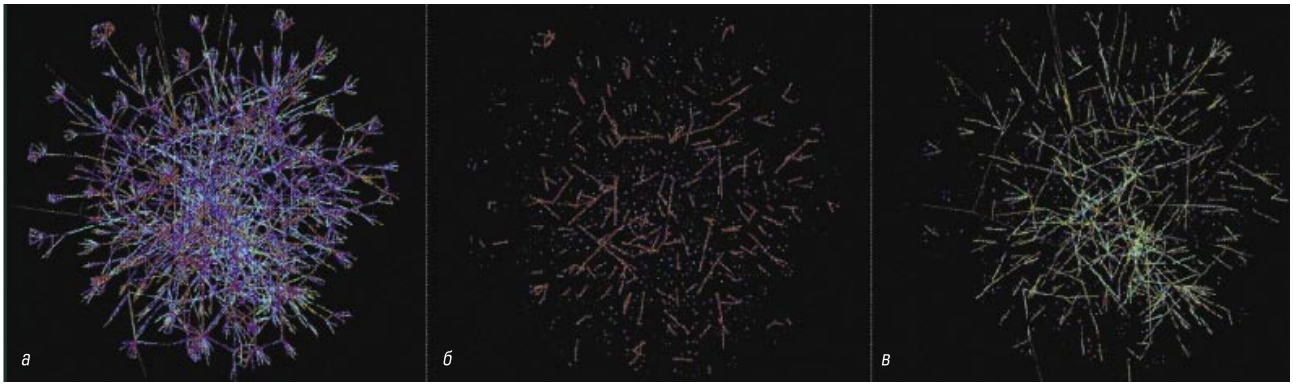
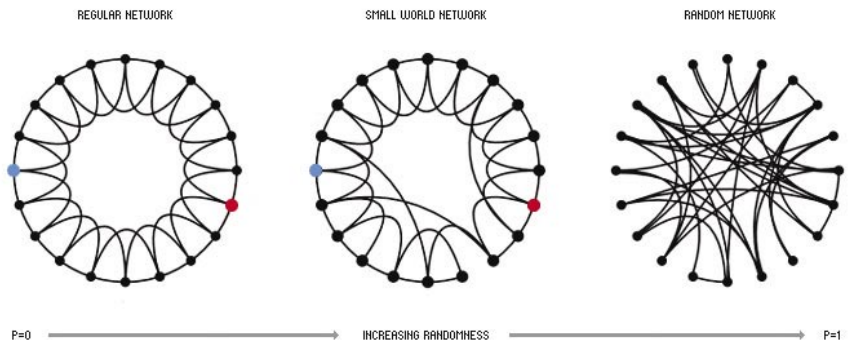
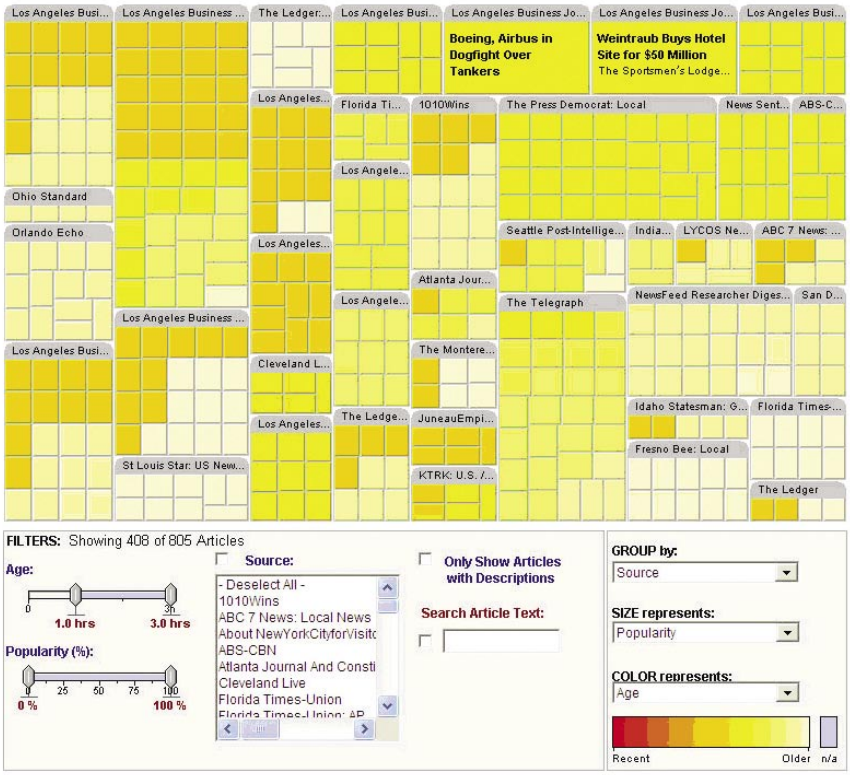
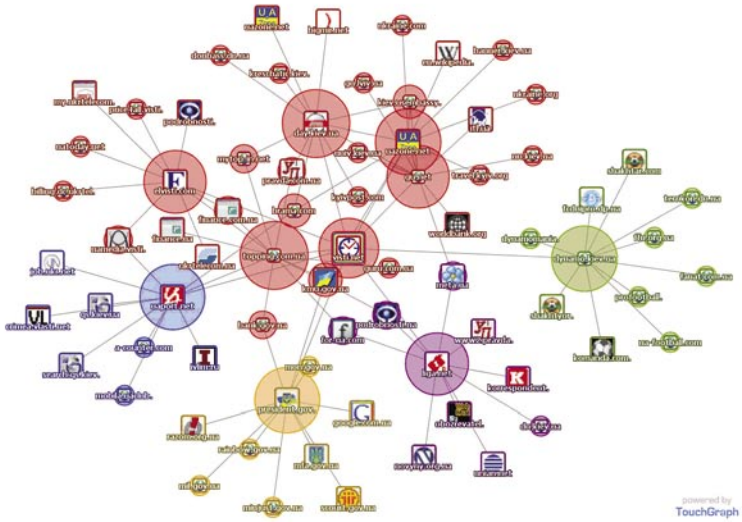


Рис. 7. Именно дальние, «слабые», связи обеспечивают цельность сети:
‰ — полная карта сети социальных коммуникаций;
¾ — социальная сеть, из которой удалены слабые связи, разбивается на множество изолированных участков;
¿ — карта сети, из которой удалены сильные связи: структура сохраняет сквозную проводимость





Item Map for Category 'Regional/North America/United States' (English) Generated on 23/06/07 14:49


powered by
TouchGraph