ЭЛЕМЕНТЫ ФРАКТАЛЬНОГО АНАЛИЗА ИНФОРМАЦИОННЫХ ПОТОКОВ

Дмитрий Ландэ

Мрак первозданный. Тишина. Вдруг луч,
Пробившийся над рваным краем туч,
Ваяет из небытия слепого
Вершины, склоны, пропасти, хребты,
И твердость скал творя из пустоты,
И невесомость неба голубого.

Герман Гессе. «Игра в бисер»

1) Понятие «фрактал»

Термин *фрактал (*от латинского слова fractus – дробный), был предложен Бенуа Мандельбротом в 1975 году для обозначения нерегулярных самоподобных математических структур. Популярная сегодня фрактальная геометрия получила свое название лишь в 1977 году благодаря его книге «The Fractal Geometry of Nature». В работах ученого использованы научные результаты многих ученых, работавших в этой же области (прежде всего, Пуанкаре, Кантора, Хаусдорфа). Основное определение фрактала, данное Мандельбротом, звучало так: *"Фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому"*.

В самом простом случае небольшая часть фрактала содержит информацию о всем фрактале. Строгое определение самоподобных множеств было дано Дж. Хатчинсоном в 1981 году. Он назвал множество самоподобным, если оно состоит из нескольких компонент, подобных всему этому множеству, т.е. компонент получаемых афинными преобразованиями - поворотом, сжатием и отражением исходного множества.

Однако самоподобие – это хотя и необходимое, но далеко не достаточное свойство фракталов. Ведь нельзя же, в самом деле, считать фракталом точку, или плоскость, расчерченную клетками. Главная особенность фракталов заключается в том, что их размерность не укладывается в привычные геометрические представления. Фракталам характерна геометрическая «изрезанность». Поэтому используется специальное понятие фрактальной размерности, введенное Феликсом Хаусдорфом (1868-1942) и Абрамом Самойловичем Безиковичем (1891-1970). Применительно к идеальным объектам классической евклидовой геометрии она давала те же численные значения, что и известная задолго до нее так называемая топологическая размерность (иначе говоря, была равна нулю для точки, единице - для гладкой плавной линии, двум - для фигуры и поверхности, трем - для тела и пространства). Но совпадая со старой, топологической, размерностью на идеальных объектах, новая размерность обладала более тонкой чувствительностью ко всякого рода несовершенствам реальных объектов, позволяя различать и индивидуализировать то, что прежде было безлико и неразличимо. Так, отрезок прямой, отрезок синусоиды и самый причудливый меандр неразличимы с точки зрения топологической размерности - все они имеют топологическую размерность, равную единице, тогда как их размерность Хаусдорфа - Безиковича различна и позволяет числом измерять степень извилистости. Размерность фрактальных объектов не является целым числом, характерным для привычных геометрических. Вместе с тем, в большинстве случаев фракталы напоминают объекты, плотно занимающие реальное пространство, но не использующее его полностью.

Пусть есть множество G в пространстве R^n. Разобьем пространство R^n на n-мерные кубы с длиной ребра δ и обозначим число кубов, необходимых для покрытия ими множества G, через $N(\delta)$. Тогда величина размерности Хаусдорфа-Безиковича (называемой также фрактальной размерностью) D должна удовлетворять следующему условию:

$$\lim_{\delta \to 0} N(\delta)\delta^d = \begin{cases} 0, & d > D \\ \infty, & d < D \end{cases}.$$

Данное определение можно упростить, сделав его более удобным для практического применения. Видно, что при $\delta \approx 0$, оно эквивалентно:

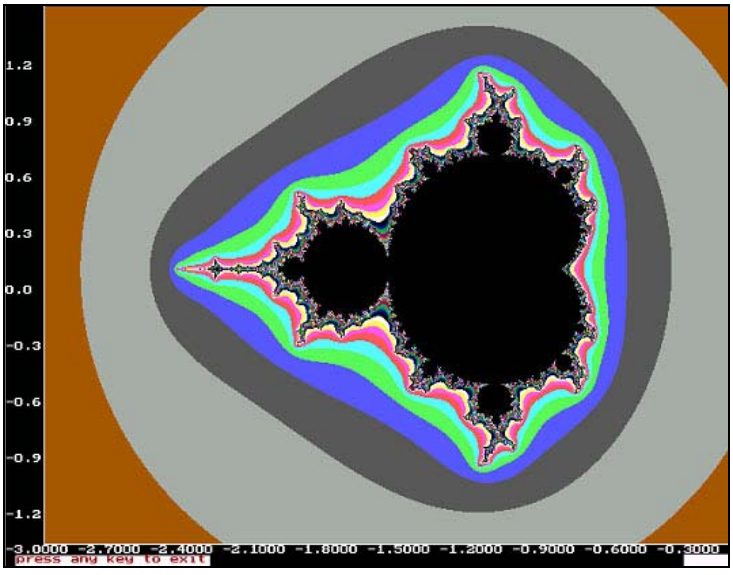
$D \approx - \ln N(\delta) / \ln \delta.$

2) Примеры абстрактных фракталов

Мандельброт предложил не только определение фракталов, но также и алгоритм построения одного из них, получившего название в честь ученого. Алгоритм построения множества Мандельброта основан на итеративном вычислении по формуле:

$$\boldsymbol{Z}[i+1] = \boldsymbol{Z}[i] * \boldsymbol{Z}[i] + \boldsymbol{C},$$

где $\boldsymbol{Z}$ и $\boldsymbol{C}$ - комплексные переменные. Итерации выполняются для каждой стартовой точки $\boldsymbol{C}$ прямоугольной или квадратной области - подмножестве комплексной плоскости. Итерационный процесс продолжается до тех пор, пока $\boldsymbol{Z}[i]$ не выйдет за пределы окружности заданного радиуса, центр которой лежит в точке (0,0), или после достаточно большого числа итераций. В зависимости от количества итераций, в течение которых $\boldsymbol{Z}[i]$ остается внутри окружности, устанавливается цвет точки $\boldsymbol{C}$. Если $\boldsymbol{Z}[i]$ остается внутри окружности в течение достаточно большого количества итераций, то эта точка растра окрашивается в черный цвет. Множеству Мандельброта принадлежат именно те точки, которые в течение бесконечного числа итераций не уходят в бесконечность.

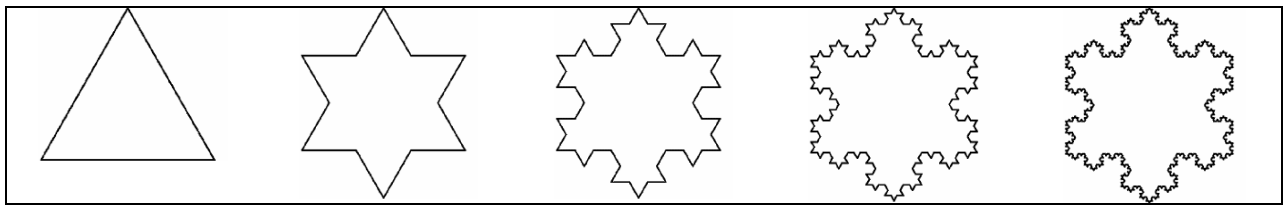


Множество Мандельброта

Так как количество итераций соответствует номеру цвета, то точки, находящиеся ближе к множеству Мандельброта, имеют более яркий цвет.

Увеличенный участок границы множества Мандельброта

Построение другого фрактального множества, снежинки Коха, начинается с правильного треугольника, длина стороны которого равна 1. Сторона треугольника считается базовым звеном для исходного положения. Далее, на любом шаге итерации каждое звено заменяется на образующий элемент – ломанную, состоящую по краям из отрезков длиной 1/3 от длины звена, между которыми размещаются две стороны правильного треугольника со стороной в 1/3 длины звена. Все отрезки - стороны полученной кривой считаются базовыми звеньями для следующей итерации. Кривая, получаемая в результате ***n***-й итерации при любом конечном ***n,*** называется *предфракталом*, и лишь при ***n,*** стремящемся к бесконечности, кривая Коха становится фракталом. Получаемое в результате итерационного процесса фрактальное множество представляет собой линию бесконечной длины, ограничивающую конечную площадь. Действительно, при каждом шаге число сторон результирующего многоугольника увеличивается в 4 раза, а длина каждой стороны уменьшается только в 3 раза, т.е. длина многоугольника на n-й итерации равна $3*(^4/_3)^n$ и стремится к бесконечности с ростом ***n***.



Первые 5 поколений снежинки Коха

Площадь под кривой, если принять площадь образующего треугольника за 1, равна:

$$S = 1 + 1/3\sum_{k=0}^{\infty}(4/9)^k = 1{,}6.$$

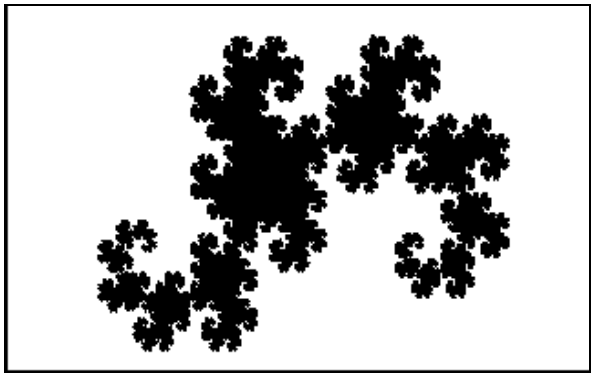
В 80-х годах XX века как простое средство получения фрактальных структур появился метод "Систем Итерационных Функций" (Iterated Functions System - IFS). IFS представляет собой систему функций, отображающих одно многомерное множество на другое. Наиболее простая реализация IFS представляет собой аффинные преобразования плоскости:

$$X' = A*X + B*Y + C$$
$$Y' = D*X + E*Y + F$$

В 80-х годах американские ученые М. Барнсли и А. Слоан предложили идею сжатия и хранения графической информации, основанную на соображениях теории фракталов и динамических систем. На основании этой идеи был создан алгоритм

фрактального сжатия информации, позволяющий сжимать некоторые образцы графической информации в 500-1000 раз. При этом каждое изображение кодируется несколькими простыми аффинными преобразованиями. Закодировав какое-то изображение двумя аффинными преобразованиями, оно однозначно определяется с помощью 12-ти коэффициентов. Если определить начальную точку итерационного процесса (например, *X=0 Y=0*) и запустить этот процесс, то через несколько итераций совокупность полученных точек будет описывать закодированное изображение.

В качестве примера использования IFS для построения фрактальных структур, можно привести кривую "дракона" Хартера-Хейтуэя.



"Дракон" Хартера-Хейтуэя

Использование IFS для сжатия обычных изображений, например, фотографий основано на выявлении локального самоподобия (в отличие от фракталов, где наблюдается глобальное самоподобие). По алгоритму Барнсли происходит выделение в изображении пар областей, меньшая из которых подобна большей, и сохранение нескольких коэффициентов, кодирующих преобразование, переводящее большую область в меньшую. Требуется, чтобы множество таких областей покрывало все изображение. Восстанавливающий алгоритм должен применять каждое преобразование к некоторому подмножеству, принадлежащему области, соответствующей применяемому преобразованию.

Фракталы с большой точностью описывают многие физические явления и природные образования: облака, турбулентные течения, ветви деревьев, кровеносные сосуды. Мандельброт в свое время заметил: *"Почему геометрию часто называют холодной и сухой? Одна из причин заключается в ее неспособности описать форму облака, горы, дерева или берега моря. Облака - это не сферы, горы - не конусы, линии берега - это не окружности, и кора не является гладкой, и молния не распространяется по прямой. Природа демонстрирует нам не просто более высокую степень, а совсем другой уровень сложности."*

В машинной графике фрактальные подходы приходят на помощь, например, когда требуется, с помощью нескольких коэффициентов, задать линии и поверхности очень формы. Фрактальная геометрия сегодня незаменима при компьютерной генерации облаков, гор, поверхности моря, других сложных «неевклидовых» объектов, образы которых напоминают природные.

3) Фракталы в природе

В реальной жизни фрактальные объекты имеют вполне определенные границы фрактальности, в том числе и самоподобия. Тем не менее, фракталы – это очень удобная и

наглядная абстракция, которая сегодня уже широко применяется при моделировании естественных процессов. При этом спектр применения фракталов постоянно расширяется, сегодня он применяется и к моделированию информационного пространства.

Один из лучших примеров проявления фракталов в природе –структура береговых линий. Действительно, на километровом отрезке побережье выглядит столь же изрезанным, как и на стокилометровом. Опыт показывает, что длина береговой линии L зависит от масштаба l, которым проводятся измерения, и увеличивается с уменьшением последнего по степенному закону $L = \Lambda\, l^{1-\alpha}$, $\Lambda = const$. Так, например, для побережья Великобритании $\alpha \approx 1.24$, т.е., так называемая, фрактальная размерность береговой линии Великобритании равна 1.24 .

Береговая линия побережья Великобритании

Недавно Б. Саповаль из Политехнической школы в Палезо (Франция) и его коллеги создали компьютерную модель эрозии побережья. В модели вещество разрушалось либо под прямым воздействием волн, либо медленным "выветриванием", когда минералы растворялись в воде. Побережье было разделено на равные участки, причем в модели типы камней на этих участках выбирались случайным образом. Эрозионная сила моря зависит от того, насколько сильно глушатся волны. В узком заливе или бухте вода всегда спокойнее. Саповаль предположил, что глушение волн усиливается по мере того, как берег становится более изрезанным. Модель показала, что изначально гладкая береговая линия стремительно приобретает неровный профиль с выступами и множеством отделенных от берега островов. При моделировании береговых линий использовались двумерные стохастические фракталы, которые получаются в том случае, если в итерационном процессе случайным образом варьировать некоторые его параметры. Образовавшийся при моделировании берег очень напоминал Восточное побережье США.

В 2004 году в Ньюфаундленде биологом Ги Нарбонна из университета Кингстона (Канада) была открыта редкая ископаемая природная структура фрактального типа. Были найдены следы организмов, живших на Земле около 575 миллионов лет назад, и не относившихся ни к растениям, ни к животным, и называются рангеоморфами. Они были неспособны двигаться и не имели репродуктивных органов, а размножались, создавая новые ответвления. Организмы собирались во фрактальные структуры из разветвляющихся частей. Как выяснилось, каждый ветвящийся элемент фрактальных структур состоял их множества трубок, удерживаемых вместе полужестким органическим скелетом организмов. Нарбонн обнаружил рангеоморфы, собранные в несколько разных форм. Фрактальный рисунок представляется достаточно сложным, но, по словам исследователя, сходство организмов друг с другом обеспечивалось достаточно простым геном.

Уже около полувека в биологии известен закон, который утверждает, что многие свойства организмов, от продолжительности жизни и количества детенышей до скорости обмена веществ, пропорциональны массе тела в степени n/4, где n - целое. При этом сама природа закона более полувека оставалась загадкой. На первый взгляд, вместо четверки должна быть тройка, поскольку масса пропорциональна кубу размера тела.

Несколько лет назад объяснение, было найдено. Дело в том, что пронизывающие каждый организм сети - кровеносная у животных или капиллярная у растений - обладают свойствами фракталов. Фрактальность этих сетей как раз и приводит к добавлению еще одного "измерения" у живых организмов.

И наконец, вся Вселенная, в соответствии с гипотезой российского физика С. Хайтуна, является фракталом, причем единственным известным в природе, полностью удовлетворяющим классическому определению. В физике известен факт, что плотность космических объектов стремительно падает с их размерами. Еще в 50-х годах советские физики-теоретики пришли к выводу, что "бесконечная" плотность Вселенной равна нулю. Эта идея и новейшие представления о фрактальности Вселенной подтверждают друг друга. Дело в том, что плотность всякого фрактала, расположенного в трехмерном пространстве, тождественно равна нулю. Классические фракталы обладают "всюду пустой" структурой, которая при проникновении в нее "расступается" до бесконечности. Вместе с тем, реальные системы бесконечного углубления в свою структуру не позволяют; на каком-то конечном шаге структура, будь то, скажем, снежинка или кровеносная система человека, теряет свой "фрактальный" вид - реальные структуры лишь "фракталоподобны". В соответствии с гипотезой Хайтуна, позволяя - из-за своей бесконечности - бесконечное проникновение в свою структуру, Вселенная, по мнению многих исследователей, является единственным "настоящим" фракталом, имея нулевую бесконечную плотность.

4) Информационное пространство и фракталы

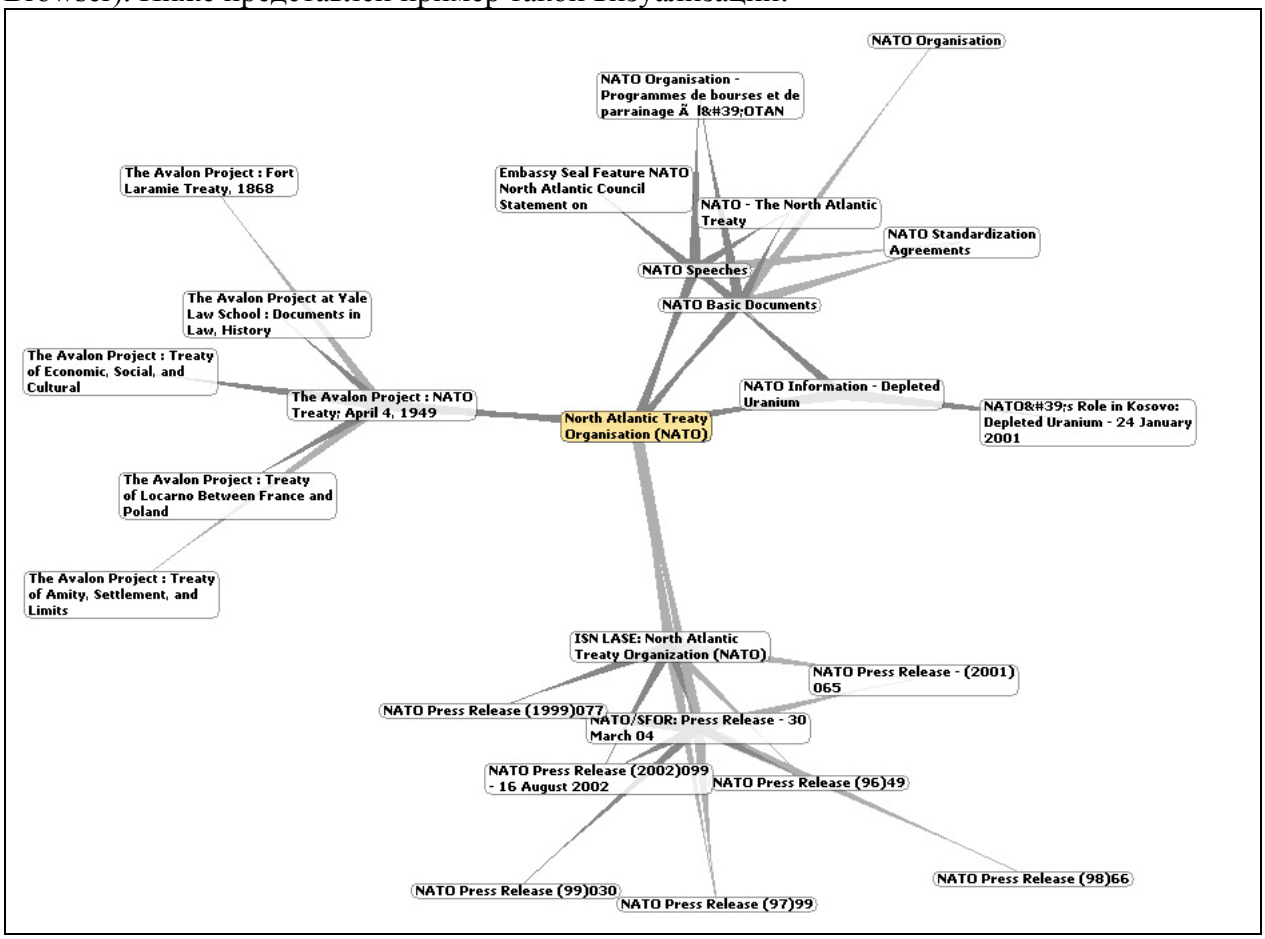
В настоящее время информационное пространство в целом, ввиду его объемов и динамики изменения, принято рассматривать как стохастическое. Во многих моделях информационного пространства изучаются структурные связи между тематическими множествами, входящими в это пространство. При этом численные характеристики этих множеств подчиняются гиперболическому закону (с возможными степенными поправками). Сегодня в моделировании информационного пространства все чаще используется фрактальный подход, базирующийся на свойстве самоподобия информационного пространства, т.е. сохранение внутренней структуры множеств при изменениях их размеров или масштабов их рассмотрения извне.

Самоподобие информационного пространства выражается, прежде всего в том, что при его лавинообразном росте в последние десятилетия, частотные и ранговые распределения, получаемые в таких разрезах, как источники, авторы, тематика практически не меняют своей формы. Поэтому применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, составляющие основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, которые по своей сути являются стохастическими фракталами, так как их самоподобие справедливо на лишь уровне математических ожиданий, например, распределения кластеров по размерам.

В информационном пространстве возникают, формируются, растут и размножаются кластеры – группы взаимосвязанных документов. Системы, основанные на

кластерном анализе, самостоятельно выявляют новые признаки объектов и распределяют объекты по новым группам.

Не так давно в Интернет появился сервис Touchgraph (www.touchgraph.com), который наглядно демонстрирует появление кластерных образований, сформированных подобием информационных объектов, в частности, Web-сайтов (Touchgraph Google Browser). Ниже представлен пример такой визуализации:



Объединение Web-сайтов по признаку подобия

Чем же определяется природа фрактальной структуры информационного пространства, порождаемым такими кластерными структурами? С одной стороны, параметрами ранговых распределений, а, с другой стороны, механизмом развития информационных кластеров, который отражает природу информационного пространства. Появление новых публикаций увеличивает размерность уже существующих кластеров и является причиной образования новых.

Фрактальные свойства характерны для кластеров информационных Web-сайтов, на которых публикуются документы, соответствующие определенным тематикам. Эти кластеры, как наборы тематических документов, представляют собой фрактальные структуры, обладающие рядом уникальных свойств. Например, российскими исследователями (С. Иванов и др.), определена фрактальная размерность подобных информационных массивов, изменяющаяся в пределах от 1.05 до 1.50, что свидетельствует о небольшой плотности заполнения кластеров документами по одной теме.

Как один из основных законов отражающих самоподобие информационного пространства можно назвать закон Зипфа. В 1949 году профессор филологии из Гарварда

Дж. Зипф собрал достаточный статистический материал, и экспериментально показал, что распределение слов естественного языка подчиняется закону: *"Если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, а затем ранжировать эти слова, т.е. расположить их в порядке убывания частоты встречаемости в данном тексте и пронумеровать в возрастающем порядке, то для любого слова произведение его порядкового номера (ранга) этом списке и частоты его встречаемости в тексте будет величиной постоянной."* Ученый описал обнаруженную им закономерность распределения слов в текстах на английском языке:

- небольшое количество слов, таких как "the", "and" в английском языке, которые имеют очень высокий ранг;
- среднее количество слов имеет средний ранг;
- большое количество слов имеет очень низкий ранг.

Таким образом: $f * r = c$, где f - частота встречаемости слова в тексте; r - ранг (порядковый номер) слова в списке; c - эмпирическая постоянная величина. Так, например, для английских текстов константа Зипфа равна приблизительно 0,1. Для русского и украинского языков коэффициенты Зипфа составляю приблизительно 0,06-0,07.

Существуют также закономерности, открытые другими учеными (прежде всего, Брэдфордом - для периодических изданий и Лотки – для распределения авторов), являющиеся уточняющими следствиями закономерностей Зипфа, и также свидетельствующими о самоподобии информационного пространства.

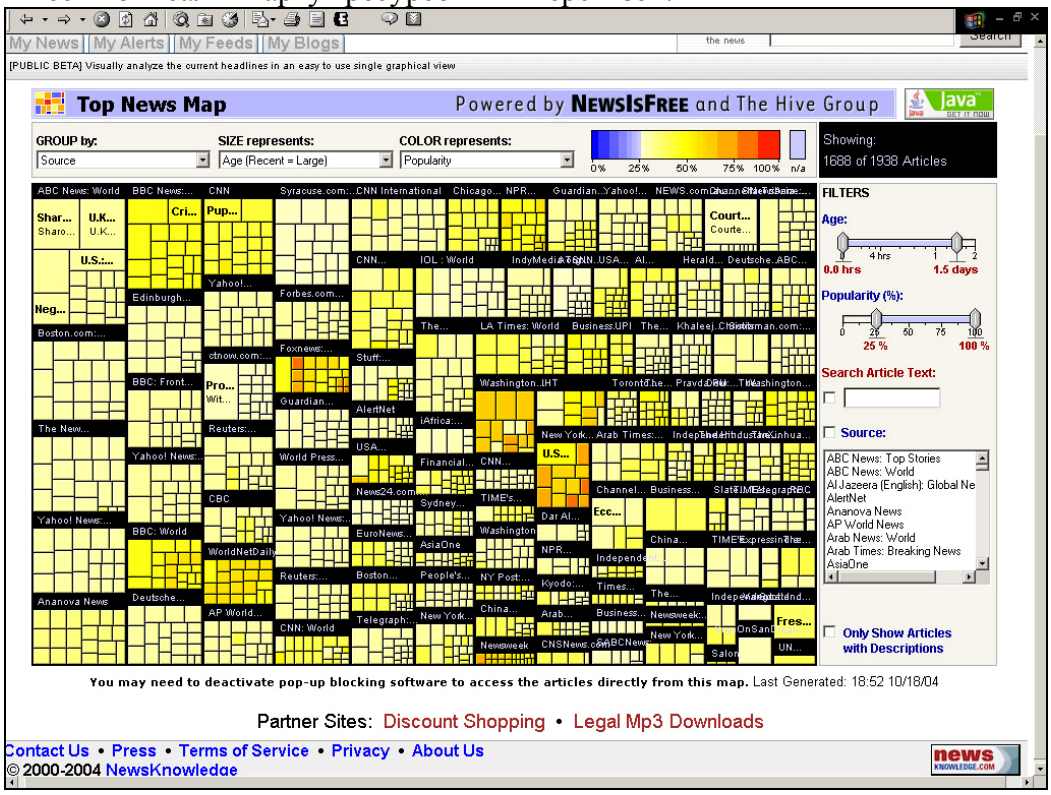
Теория фракталов тесно связана с кластерным анализом, решающим задачу выделения компактных групп объектов с близкими свойствами. Кластеризация сегодня применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, для упрощения процесса просмотра при поиске необходимой информации, нахождения уникальных документов из коллекции, выявления дубликатов или близких по содержанию документов.

Фрактальный принцип самоподобия предполагает бесконечное дробление набора объектов с сохранением их свойств. В тематических информационных потоках, например, можно наблюдать подобие сюжетных цепочек, получаемых при уточнении запроса (конечно в определенных рамках). Вместе с тем, сегодня многими исследователями рассматривается не дробление, а естественный рост размеров информационного пространства.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует новый интерфейс представленный на веб-сайте службы News Is Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. При этом учитывается два основных параметра отображения – ранг популярности и «свежесть» информации. В рамках этой модели можно наблюдать «дробление» групп источников при увеличении ранга популярности и «свежести» изданий. Когда этот ранг становится достаточно высоким, дробление не позволяет без особых усилий читать названия источников и идентифицировать отдельные документы.

Web-пространство, являясь, пожалуй, самой динамичной частью информационного пространства, характеризуется большом количестве скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. В ноябре 1999 года один из руководителей института поиска и анализа текстов, входящего в исследовательское подразделение IBM,

Андрей Брёдер (Andrei Broder) и его соавторы из компаний AltaVista, IBM и Compaq математически описали "карту" ресурсов и гиперсвязей.



Кластеры публикаций службы News Is Free

Исследования опровергли расхожее мнение, будто Internet - это единое густое пространство. Проследив с помощью поискового механизма AltaVista свыше 200 млн. Web-страниц и несколько миллиардов ссылок, размещенных на этих страницах, ученые пришли к выводу о структуре Web-пространства как ориентированного графа, в котором вершины соответствуют Web-страницам, а ребра – соединяющим эти страницы гиперссылкам. В рамках этой модели задача анализа структуры связей между отдельными Web-страницами было обнаружено:

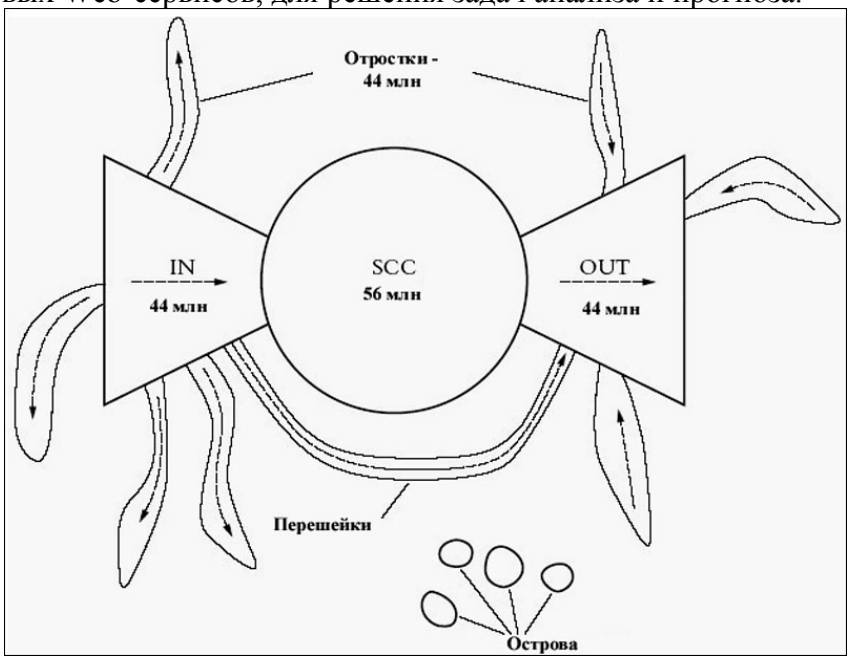
- центральное ядро (28% Web-страниц) - компоненты сильной связности (SCC).
- 22% Web-страниц - это "отправные Web-страницы" (IN). Они содержат гиперссылки, которые в конечном счете ведут к ядру, но из ядра к ним попасть нельзя.
- столько же - 22% - "оконечных Web-страниц" (OUT), к которым можно прийти по ссылкам из ядра, но нельзя вернуться назад.
- 22% Web-страниц - отростки - полностью изолированы от центрального ядра: это либо "мысы", связанные гиперссылками со страницами любой другой категории, либо "перешейки", соединяющие две Web-страницы, не входящие в ядро.

В модели учтены и "острова", которые вообще не пересекаются с остальными ресурсами Internet. Единственный способ обнаружить ресурсы этой группы - знать адрес.

Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств Web-пространства, подтверждая тем самым наблюдение о том, что "Web - это фрактал", т.е. свойства структуры всего Web-пространства Bow Tie также верны и его отдельных подмножеств.

Алгоритмы, использующие информацию о структуре Web-пространства, предположительно должно работать и на отдельных его подмножествах. Информация о структуре Web-пространства уже достаточно широко используется при решении многих

задач, например, для оптимизации эффективности механизмов сканирования, при построении новых Web-сервисов, для решения задач анализа и прогноза.



Модель Bow Tie

5) Фракталы и временные ряды

Новостная составляющая информационного пространства Интернет сегодня настолько значительна по своим объему и динамике, что может рассматриваться как мощный информационный поток. Причем поток достаточно неоднородный, который может характеризоваться большим количеством параметров, среди которых выделяются такие, как источники информации (веб-сайт) и тематики. Именно их можно рассматривать, как лежащие на поверхности основы для кластеризации.

В то время, как для традиционных средств научной коммуникации подходы к кластеризации с точки зрения теории фракталов были впервые исследованы Ван Рааном, анализировавшим массивы статей и связи, образуемые цитированием, информационные потоки сообщений из Интернет до последнего времени не ассоциировались с фракталами, что связано с проблемами идентификации информационных потоков как фрактальных множеств, а также с трудностью нахождения основ для построения кластеров - сообщений в политематических потоках, порождающих многократное цитирование.

По этой же причине исследуются количественные характеристики лишь тематических информационных потоков, которые характеризуются итеративностью при формировании и вполне доступны как для количественного, так и для качественного анализа.

Объемы сообщений в тематических информационных потоках образуют временные ряды. Временные ряды, порождаемые тематическими информационными потоками, также обладают фрактальными свойствами и могут рассматриваться как стохастические фракталы. Этот подход расширяет область применения теории фракталов на информационные потоки, динамика которых описывается средствами теории случайных процессов.

С другой стороны, теория фракталов рассматривается как подход к статистическому исследованию, который позволяет получать важные характеристики информационных потоков, не вдаваясь в детальный анализ их внутренней структуры и связей. Одним из основных свойств фракталов является самоподобие (скейлинг). Как

показано в работах С.А. Иванова, для последовательности сообщений тематических информационных потоков в соответствии со скейлинговым принципом, количество сообщений, резонансов на события реального мира пропорционально некоторой степени количества источников информации (кластеров) и итерационно продолжается в течение определенного времени. Точно так же, как и в традиционных научных коммуникациях, растущее множество сообщений в Интернет по одной тематике во времени представляет собой динамическую кластерную систему, возникающую в результате итерационных процессов. Этот процесс объясняется републикациями, прямой или совместной цитируемостью, различными публикациями – отражениями одних и тех же событий реального мира, прямыми ссылками и т.д. Кроме того, для большинства тематических информационных потоков наблюдается увеличение их объемов, причем на коротких временных интервалах – линейный рост, а на длительных – экспоненциальный.

Фрактальная размерность в кластерной системе, соответствующей тематическим информационным потокам, показывает степень заполнения информационного пространства сообщений в течение определенного времени:

$$N_{публ} = \varepsilon^{\rho} N_k(t)^{\rho},$$

где $N_{публ}$ – размер кластерной системы (общее число электронных публикаций в информационном потоке); N_k – размер - число кластеров (тематик или источников), ρ - фрактальная размерность информационного массива; ε - коэффициент масштабирования. В приведенном соотношении между количеством сообщений и кластеров проявляется свойство сохранения внутренней структуры множества при изменении масштабов его внешнего рассмотрения.

По мнению Иванова, все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Зипфа, могут быть обобщены именно в рамках теории стохастических фракталов, что было отмечено выше.

6) Показатель Херста

Сегодня в связи с развитием теории стохастических фракталов становится популярной такая характеристика временных рядов как, показатель Херста (H). Известно, что он связан с традиционной «клеточной» фрактальной размерностью (D) простым соотношением:

$$D + H = 2.$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью, определено Е. Федером следующим образом: *«… рассматривают клетки, размеры которых малы по сравнению как с длительностью процесса, так и с диапазоном изменения функции; поэтому соотношение справедливо, когда структура кривой, описывающая фрактальную функцию, исследуется с высоким разрешением, т.е. в локальном пределе».* Еще одним важным условием является самоаффинность функции. Не вдаваясь в подробности заметим, что для информационных потоков это свойство интерпретируется как самоподобие, возникающее в результате процессов их формирования. Можно заметить, что указанными свойствами обладают не все информационные потоки, а лишь те, которые характеризуются достаточной мощностью и итеративностью при формировании. При этом временные ряды, построенные на основании мощных тематических информационных потоков, вполне удовлетворяют этому условию. Поэтому при расчете показателя Херста, фактически определяется и такой показатель тематического информационного потока, как фрактальная размерность.

Известно, что показатель Херста представляет собой меру персистентности - склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в

прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность - броуновское движение.

Для изучения фрактальных характеристик тематических информационных потоков изучались значения показателя Херста за определенный период для временных рядов, составленных из количества относящихся к ним сообщений. Показатель Херста связывают с коэффициентом нормированного размаха (*R/S*), где *R* – вычисляемый определенным образом «размах» соответствующего временного ряда, а *S* – стандартное отклонение.

Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной (в нашем случае количество сообщений в информационном потоке) за *N* дней:

$$< \xi >_N = \frac{1}{N} \sum_{t=1}^{N} \xi(t)$$

Затем рассчитывается накопившееся отклонение ряда измерений $\xi(t)$ от среднего $< \xi >_N$:

$$X(t,N) = \sum_{u=1}^{t} (\xi(u) - < \xi >_N)$$

После этого рассчитывается разность максимального и минимального накопившегося отклонения, которая и называется “размахом”:

$$R(N) = \max_{1 \le t \le N} X(t,N) - \min_{1 \le t \le N} X(t,N)$$

Стандартное отклонение рассчитывается по известной формуле:

$$S = (\frac{1}{N} \sum_{t=1}^{N} (\xi(t) - < \xi >_N)^2)^{1/2}$$

В свое время Херст экспериментально обнаружил, что для многих временных рядов справедливо:

$$R / S = (N / 2)^H$$

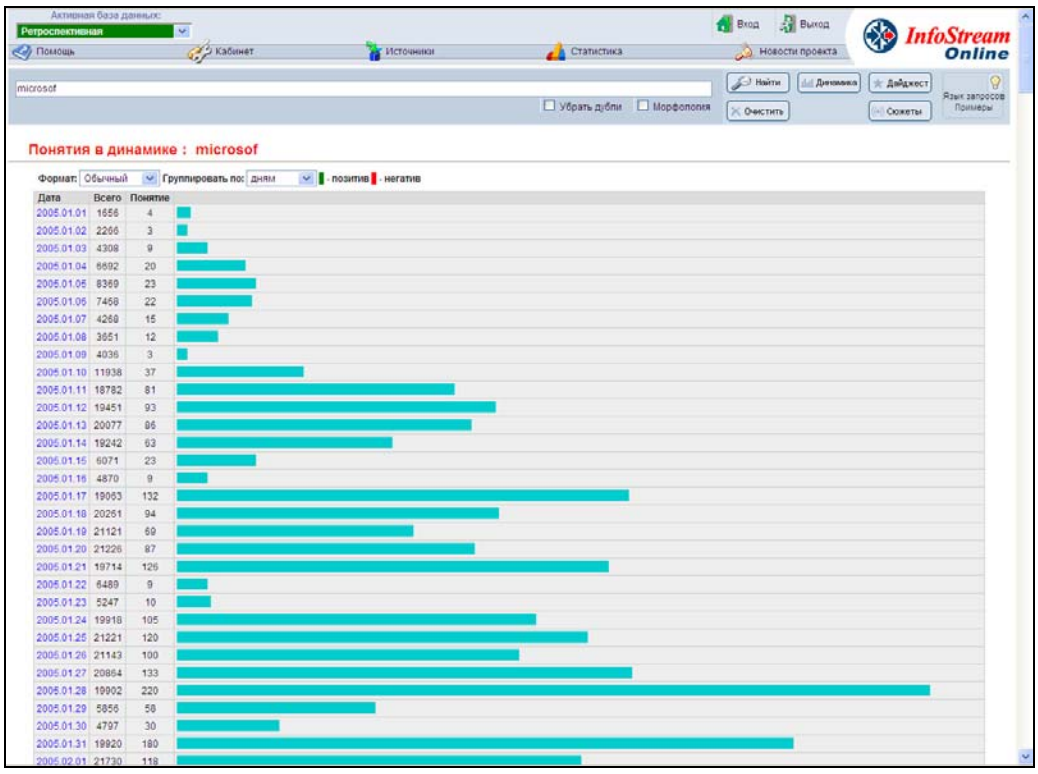
Именно коэффициент *H* и получил название показателя Херста.

7) Описание вычислительного эксперимента

В качестве экспериментальной базы для исследования фрактальных свойств тематических информационных потоков использовалась система контент-мониторинга InfoStream, разработанная в Информационном центре «ЭЛВИСТИ». Эта система, которая применяется для решения задач автоматизированного сбора новостной информации с открытых Web-сайтов и обеспечения доступа к ней в поисковых режимах, в настоящее время охватывает свыше 2000 источников информации - более 40000 уникальных новостных сообщений в сутки. В ретроспективных базах данных системы накоплено свыше 25 млн. сообщений.

Тематика исследуемого информационного потока определялась запросом к системе InfoStream, состоящим всего из одного слова «Microsoft». Ретроспективный период исследования составлял весь 2005 год и 2 месяца 2006 года, т.е. 424 дня (*N* = 424). В результате поиска было найдено 42357 релевантных документов.

Исходные данные были получены из интерфейса режима «Динамика появления понятий» (Рис. 1). На основании обработки этих данных была получена полная картина экспериментальных данных – временной ряд за указанный период (Рис.2).



Фрагмент диаграммы динамики встречаемости понятия «Microsoft»

Для этого временного ряда по формуле (6) было вычислено стандартное отклонение (S=43.71). Одновременно, с помощью механизма формирования основных сюжетов, входящего в состав системы InfoStream, были определены основные события, приведшие к возникновению пиковых значений на диаграмме.

Динамика накопления отклонения позволила определить «размах» этого параметра (R = 1207.64).

И наконец, для значения N = 424 был вычислен показатель Херста, который оказался равным 0,62, что свидетельствует о положительной персистентности всего временного ряда.

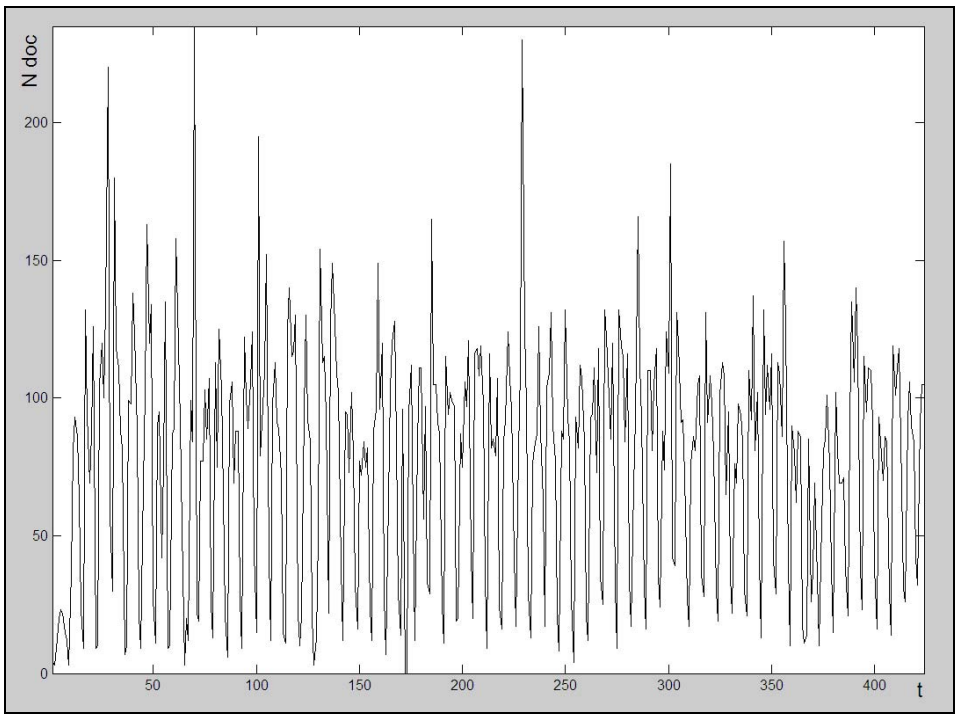
Кроме того, были выполнены расчеты показателей Херста для всех значений N, начиная с 5.

Изучение такой характеристики, как показатель Херста позволяет прогнозировать динамику информационных потоков, сообщения которых отражают процессы, происходящие в реальном мире.

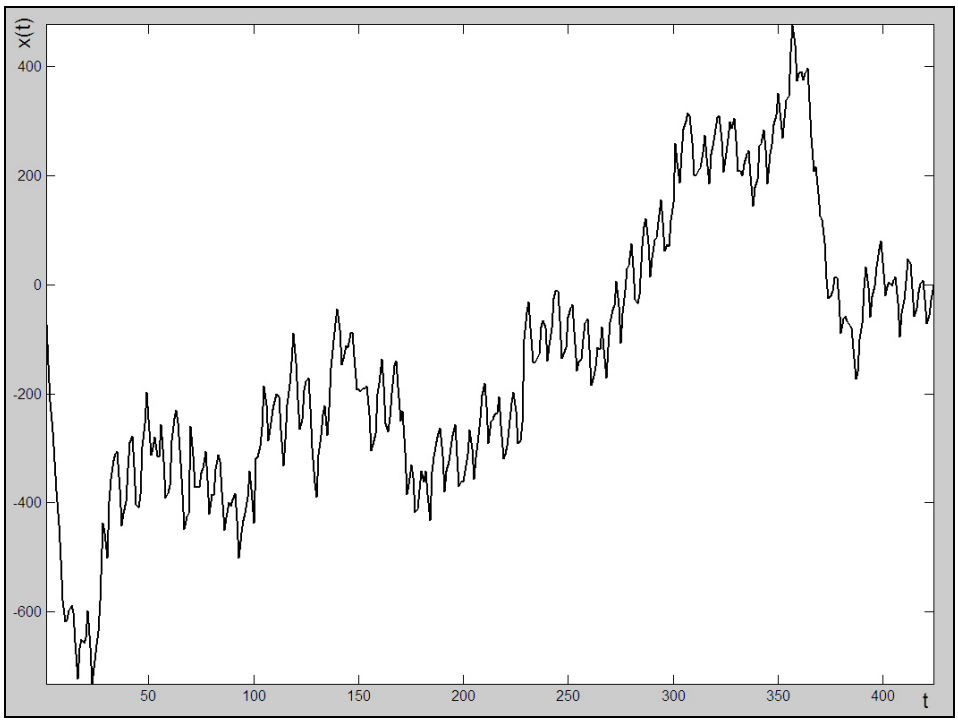
Приведенные в примере данные подтвердили лежащее в основе исследования предположение об итеративности процессов в информационном пространстве. Републикации, цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах. Скейлинговый принцип объясняется также сходством ментальности авторов, публикующих сообщения в Интернет. Вместе с тем различные маркетинговые, рекламные, PR-кампании ведут к скачкообразным изменениям в стабильных статистических закономерностях, резким скачкам и искажениям по сравнению со стандартными статистическими распределениями.

В результате эксперимента также подтверждено наличие статистической корреляции в информационных потоках на длительных временных интервалах.

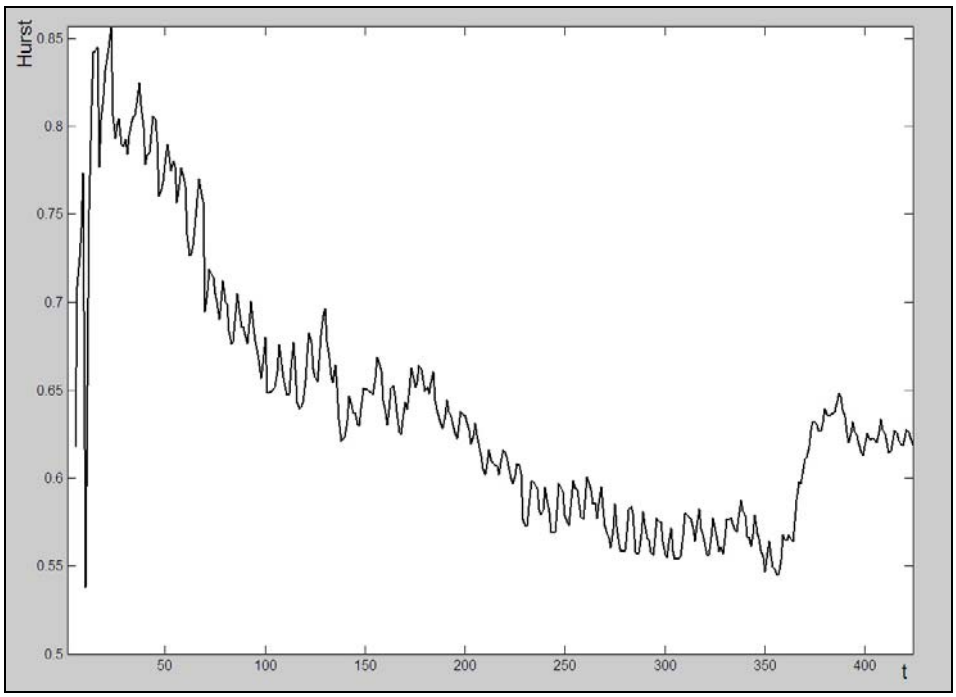
В частности, на рассматриваемом примере, показана персистентность процесса, что говорит, об общем среднем увеличении публикации о компании Microsoft, периодическом появлении пиков, связанных, как правило, с двумя подтемами-кластерами - личностью Билла Гейтса (четыре из пяти топ-кластеров) и отражениями вирусных атак (пятый топ-кластер).



Временной ряд встречаемости понятия за весь период. Пиковые значения: встречи в Давосе (конец января 2005 г.), признание журналом Forbes Б.Гейтса самым богатым человеком в мире (март 2005 г.), публикация журналом Time 100 самых влиятельных людей планеты (апрель 2005 г.), атака сетевого червя ZOTOB (август 2005 г.), 50-летний юбилей Б. Гейтса (конец октября 2005 г.)



Динамика накопления отклонения



Показатели Херста для различных временных интервалов

Естественно, описанные результаты исследований могут использоваться не только для приведенного тематического информационного канала. Своего исследования ждут кластеры, порождаемые в соответствии и с другими принципами, например, близкими по направлениям источниками информации (Web-сайтами, сетевыми СМИ, блогами и др.)

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Фракталы и кластеры в информационном пространстве. // Корпоративные системы. –2005. - №6, 2005. - С. 35-39. (http://dwl.visti.net/art/frak/)
2. Поиск знаний в Internet. Ландэ Д.В. 2005 г. (http://poiskbook.kiev.ua)
3. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет. // Регистрация, сбор и обработка данных, - К., 2006. – Т 8. - № 2. – С. 93 – 99 (http://dwl.visti.net/art/frak-ip/)

Публикации других авторов:

1. Б. Мандельброт. Фрактальная геометрия природы. – М.: Институт компьютерных исследований, 2002 г. - 656 с.
2. Б. Мандельброт. Фракталы, случай и финансы. – М.: Регулярная и хаотическая динамика, 2004 г. - 256 с.
3. *Федер Е.* Фракталы. —М.: Мир, 1991. — 254 с.
4. Э. Петерс. Хаос и порядок на рынках капитала. Новый аналитический взгляд на циклы, цены и изменчивость рынка: Пер. с англ. - М.: Мир. 2000. -333 с.
5. *Иванов С.А.* Стохастические фракталы в Информатике // Научно-техническая информация. — Сер. 2. — 2002. — № 8. — С. 7–18.
6. *Van Raan A.F.J.* Fractal Geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. —1991. — Vol. 20, N 3. — P. 439–449.