УДК 681.3

Д. В. Ландэ
Информационный центр «ЭЛВИСТИ»
ул. М. Кривоноса, 2а, 03037 Киев, Украина

Фрактальные свойства тематических информационных потоков из Интернет

Рассмотрены фрактальные свойства информационных потоков из Интернет. В качестве базы данных для вычислительного эксперимента выбрана система мониторинга сетевых новостей InfoStream. Представлена методика вычисления показателей Херста для кластера, определенного тематикой запроса, приведена качественная интерпретация результатов.

Ключевые слова: *информационные потоки, стохастические фракталы, Интернет, показатель Херста, размерность фрактальная.*

Фракталы и временные ряды

Новостная составляющая информационного пространства Интернет сегодня настолько значительна по своим объему и динамике, что может рассматриваться как мощный информационный поток [1]. Причем поток достаточно неоднородный, который может характеризоваться большим количеством параметров, среди которых выделяются такие как источники информации (web-сайт) и тематики. Именно их можно рассматривать как лежащие на поверхности основы для кластеризации [2].

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 в рамках данной статьи 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. Для исследования временных рядов сегодня все шире используется теория фракталов, традиционная область применения которой — фрактальная геометрия, обработка изображений и т.п. [3]. Вместе с тем временные ряды, порождаемые тематическими информационными потоками, также обладают фрактальными свойствами [4] и могут рассматриваться как стохастические фракталы [5, 6]. 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:

$$N_{публ} = \varepsilon^{\rho} N_k(t)^{\rho}, \qquad (1)$$

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 С.А. Иванова, все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Зипфа, могут быть обобщены именно в рамках теории стохастических фракталов.

Показатель Херста

Сегодня в связи с развитием теории стохастических фракталов становится популярной такая характеристика временных рядов как показатель Херста (H). В

книге Е. Федера [4] показано, что он связан с традиционной «клеточной» фрактальной размерностью (Θ) простым соотношением:

$$\Theta = 2 - H. \tag{2}$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью в соответствии с формулой (2), 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 отметить, 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:

$$< \xi >_N = \frac{1}{N} \sum_{t=1}^{N} \xi(t) . \tag{3}$$

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

$$X(t, N) = \sum_{u=1}^{t} (\xi(u) - < \xi >_N) . \tag{4}$$

После этого определяется разность максимального и минимального накопившегося отклонения, которая и называется «размахом»:

$$R(N) = \max_{1 \le t \le N} X(t, N) - \min_{1 \le t \le N} X(t, N). \quad (5)$$

Стандартное отклонение рассчитывается по известной формуле:

$$S = \left(\frac{1}{N}\sum_{t=1}^{N}(\xi(t) - <\xi>_N)^2\right)^{1/2}. \quad (6)$$

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

$$R/S = (N/2)^H. \quad (7)$$

Именно коэффициент H и получил название показателя Херста.

Вычислительный эксперимент

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).

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.

На рис. 3 представлена динамика накопления отклонения, которая была вычислена в соответствии с формулой (4) и позволила в соответствии с формулой (5) определить «размах» этого параметра ($R = 1207{,}64$).

И наконец, для значения $N = 424$ по формуле (7) был вычислен показатель Херста, который оказался равным 0,62, что свидетельствует о положительной персистентности всего временного ряда.

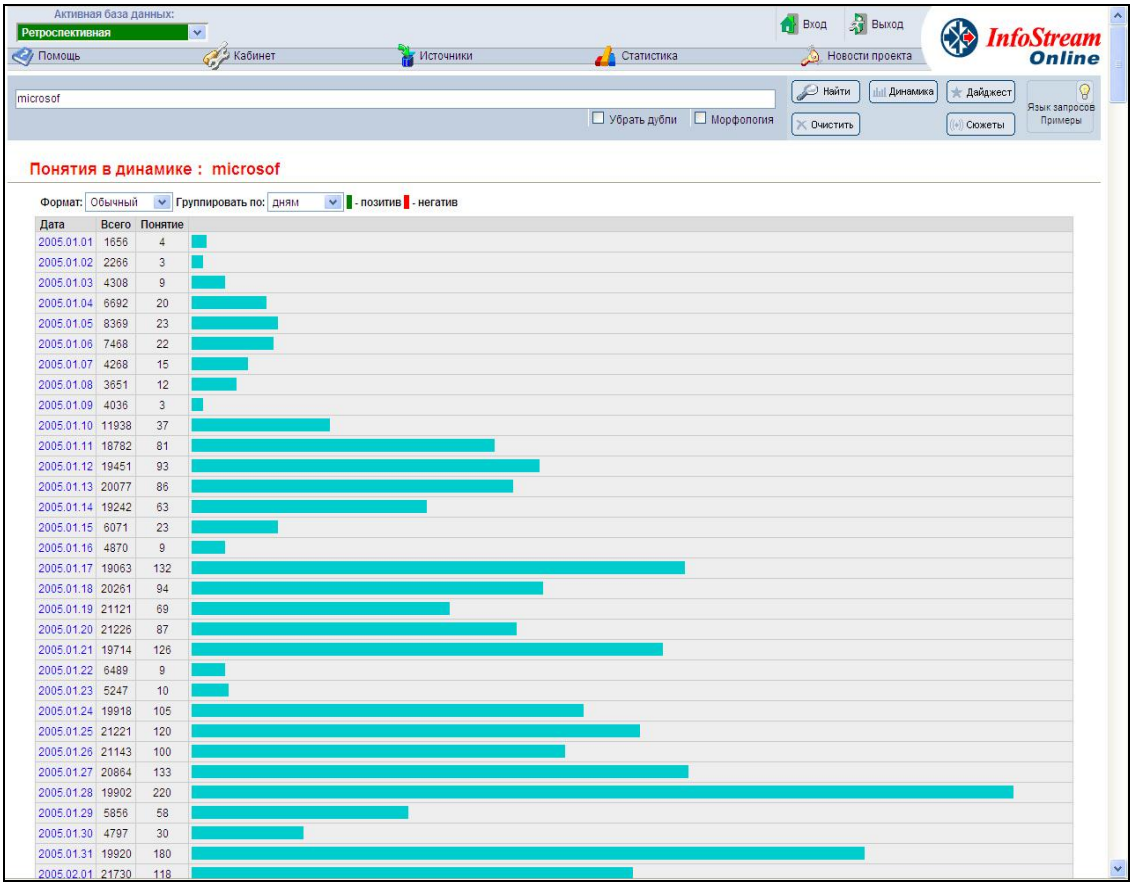

Рис. 1. Фрагмент диаграммы динамики встречаемости понятия «Microsoft»

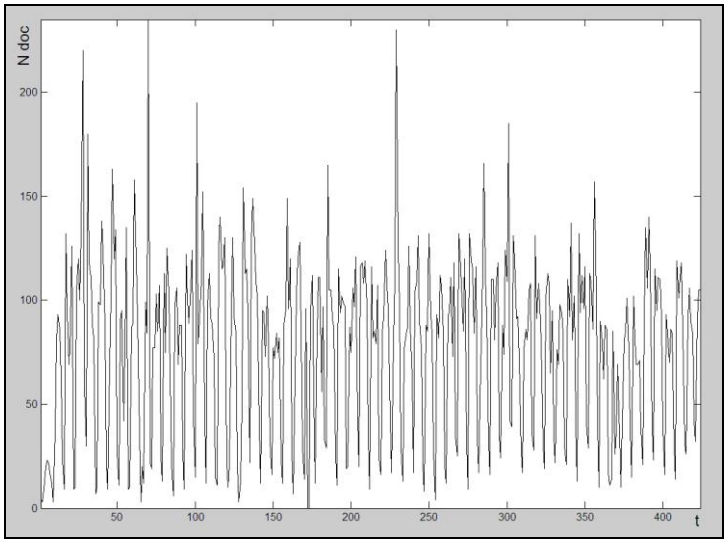

Рис. 2. Временной ряд встречаемости понятия за весь период. Пиковые значения: встречи в Давосе (конец января 2005 г.); признание журналом Forbes Б. Гейтса самым богатым человеком в мире (март 2005 г.); публикация журналом Time 100 самых влиятельных людей планеты (апрель 2005 г.); атака сетевого червя ZOTOB (август 2005 г.); 50-летний юбилей Б. Гейтса (конец октября 2005 г.)

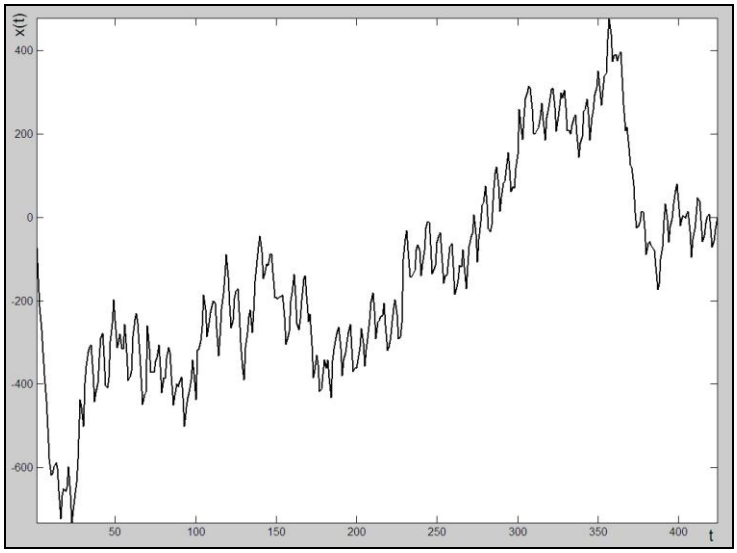

Рис. 3. Динамика накопления отклонения

Кроме того, были выполнены расчеты показателей Херста для всех значений *N*, начиная с 5, результаты которых приведены на рис. 4.

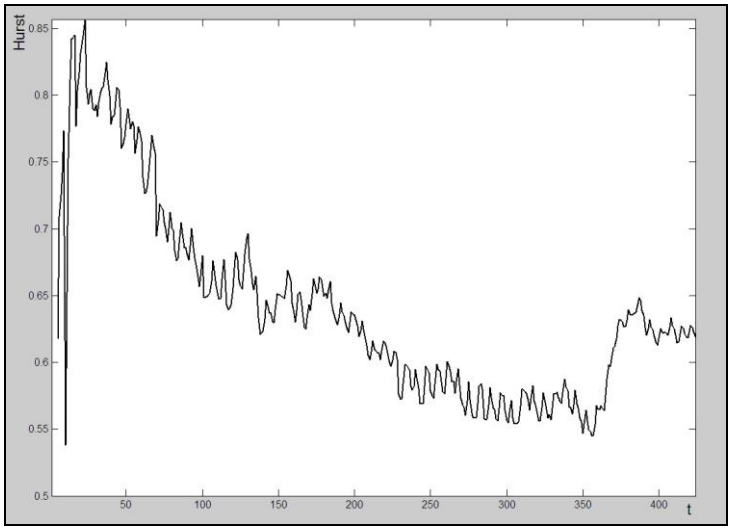

Рис. 4. Значения показателя Херста для различных временных интервалов

Интерпретация результатов

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 эмпири-

ческих 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).

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.)

1. *Брайчевский С.М., Ландэ Д.В.* Современные информационные потоки: актуальная проблематика // Научно-техническая информация. — Сер. 1. — 2005. — № 11. — С. 21–33.

2. *Van Raan A.F.J.* Fractal Geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. —1991. — Vol. 20, N 3. — P. 439–449.

3. *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа. — М.: Вильямс, 2005. — 272 с.

4. *Федер Е.* Фракталы. —М.: Мир, 1991. — 254 с.

5. *Иванов С.А.* Стохастические фракталы в Информатике // Научно-техническая информация. — Сер. 2. — 2002. — № 8. — С. 7–18.

6. *Иванов С.А., Круковская Н.В.* Статистический анализ документальных информационных потоков // Научно-техническая информация. Информ. процессы и системы. — Сер. 2. — 2004. — № 2. — С. 11–14.

Поступила в редакцию 15.03.2006