УДК 681.3

А. Г. Додонов[1], **Д. В. Ландэ**[2]
[1]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина
[2]Информационный центр «ЭЛВИСТИ»
ул. М. Кривоноса, 2а, 03037 Киев, Украина

Самоподобие массивов сетевых публикаций по компьютерной вирусологии

Описан подход к организации анализа потока тематических публикаций по компьютерной вирусологии, представленных в web-пространстве. Обоснована фрактальная природа информационных потоков, описаны основные алгоритмы, применяемые в процессе исследований, а также приведены прогнозные выводы на основе свойств персистентности временных рядов.

Ключевые слова: *Интернет-ресурсы, самоподобие, корреляции, фрактальный анализ, информационные потоки, компьютерные вирусы.*

Проблематика

В связи с наблюдающимся в последние годы ростом объемов и темпов обновления сетевой информации приобретает актуальность задача изучения статистических свойств сетевых документальных массивов [1–3]. Сложность и многоплановость этой задачи, в свою очередь, предполагает активное использование современных методов, позволяющих более глубоко понять специфику выбранной предметной области. В том числе перспективными представляются различные методы теории детерминированного хаоса [4, 5], получившие в настоящее время широкое распространение во многих областях науки. Применение таких методов в нашем случае представляется тем более интересным, что сетевые документальные массивы в этом плане остаются малоизученными.

Изучение явлений самоподобия, применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на эмпирические законы, составляющие теоретические основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, и могут рассматриваться как стохастические фракталы [6]. Известно, что все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Ципфа, могут быть обобщены именно в рамках теории стохастических фракталов [7].

© А. Г. Додонов, Д. В. Ландэ

Под компьютерной вирусологией принято понимать совокупность методов и приемов изучения компьютерных вирусов и разработки эффективных средств защиты от них [8]. Предметной областью данной статьи являются публикации в электронных СМИ по проблематике компьютерной вирусологии, отражающие такие события как вирусные атаки, создание нового программного обеспечения (ПО) как вирусного, так и антивирусного, а также мероприятия, посвященные данной тематике: конференции, выставки, презентации и т.п. (рис. 1).

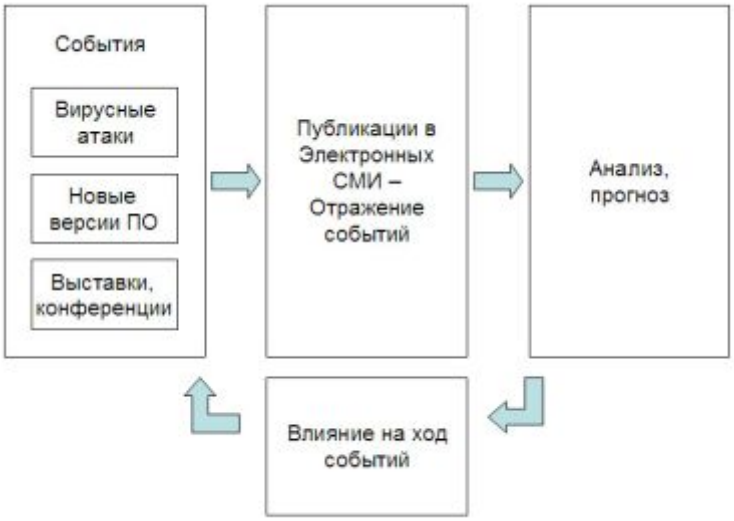

Рис. 1. Замкнутый цикл, связанный с публикациями в электронных СМИ

Очевидно, что резкие скачки в объемах потоков электронных публикаций по тематике компьютерной вирусологии свидетельствуют о некоторых реальных событиях, на которые возможна реакция соответствующих специалистов. То есть можно предположить, что как средство анализа вирусной ситуации само изучение параметров массивов электронных публикаций по данной теме относиться к средствам компьютерной вирусологии.

Описание задачи

Как известно, возникновение детерминированного хаоса в динамике объектов тесно связано с наличием у него фрактальных свойств, важность которых в последние годы широко обсуждается в самых различных областях науки. Теория фракталов широко применяется как подход к статистическому исследованию, который позволяет получать важные характеристики информационных потоков, не вдаваясь в детальный анализ их внутренней структуры. В частности, количество тематических сообщений в Интернет, резонансов на события реального мира, пропорционально некоторой степени количества тематических источников (web-сайтов). Точно так же, как и в традиционных научных коммуникациях, количество сообщений в Интернет по выбранной тематике представляет собой динамическую кластерную систему.

Как и в случае потоков энергии или вещества, проходящих через открытые системы, информационные потоки также во многих случаях обладают самоорганизацией, т.е. свойствами самоподобия, которое характеризуется сильными, подчиняющимися степенному закону, корреляциями. Если рассматривать информационные потоки как ряды публикаций в течение времени, то можно воспользоваться таким определением строгого самоподобия (масштабной инвариантности, скейлинга): процесс $X(t)$ является самоподобным, если $X(t)$ и $\alpha^{-H}X(\alpha t)$ имеют одинаковые распределения вероятностей для всех $\alpha > 0$.

В предлагаемой работе исследуются временные ряды, соответствующие количеству публикаций в сети Интернет по заданной проблематике. В наблюдаемых рядах выявлено самоподобие и устойчивые взаимные корреляции. На основании обработки данных наблюдений получены значения различных статистических показателей соответствующих рядов, а также показано, что они обладают фрактальной природой.

Экспериментальная база

Исследования проводились на наборе документальных корпусов, содержащих сообщения онлайновых СМИ различных объемов, сформированные системой InfoStream [9], которая в настоящее время позволяет осуществлять сканирование информации из нескольких тысяч web-сайтов. В ходе исследований обрабатывался информационный корпус, содержащий сообщения онлайновых СМИ — массив из 4317 документов, опубликованных за 486 суток с 1 января 2006 г. по 30 апреля 2007 г., по тематике компьютерной вирусологии, удовлетворяющих запросу **«компьютерный вирус» OR «вирусная атака»** (рис. 2).

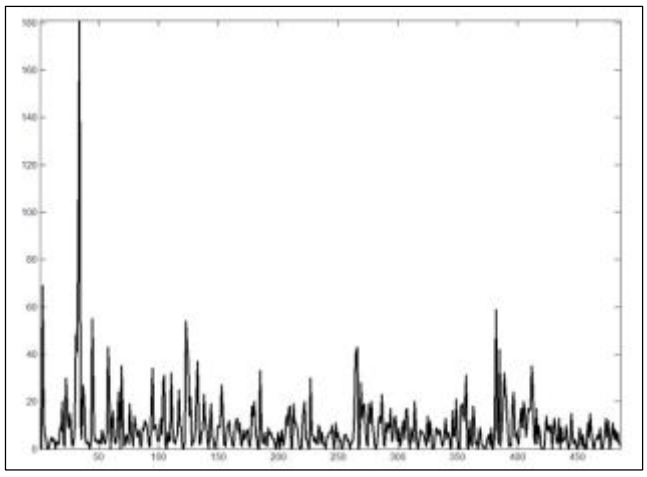

Рис. 2. Количество публикаций (ось Y) в разрезе дат (ось X)

Метод DFA

Один из подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) [10–12] — достаточно универсальном методе обработки временных рядов. Этот подход представляет собой вариант дисперсион-

ного анализа одномерных случайных блужданий, позволяющий исследовать эффекты длительных корреляций в нестационарных временных рядах. В рамках алгоритма DFA анализируется среднеквадратическая ошибка линейной аппроксимации в зависимости от размера аппроксимируемого участка. Этот метод был применен авторами к ряду значений количества публикаций по теме компьютерной вирусологии в разрезе дат. В методе DFA для различных участков ряда наблюдений одинаковой длины k исследуемой последовательности строится линейная аппроксимация, для которой затем вычисляется среднеквадратичная ошибка $D(k)$.

На рис. 3 представлена зависимость среднеквадратичной ошибки аппроксимации от длины участков аппроксимации в двойном логарифмическом масштабе. Наличие линейного тренда на этом графике позволяет говорить о наличии локального скейлинга.

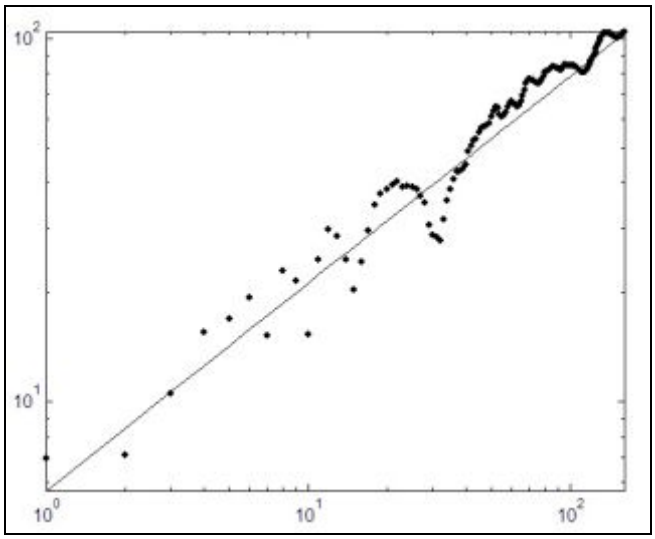

Рис. 3. Зависимость $D(k)$ ряда наблюдений (ось Y) от длины участка аппроксимации (ось X)

Коэффициенты корреляции

Как известно, коэффициенты корреляции для ряда измерений рассчитываются следующим образом:

$$R(k) = < (X_{k+t} - m)(X_k - m) >/\sigma^2,$$

где $R(k)$ — коэффициент корреляции; X_k — ряд измерений; m — математическое ожидание X_k; σ^2 — дисперсия.

Графическое представление коэффициента корреляции для исследуемого ряда наблюдений свидетельствует о разделении корреляционных свойств по дням недели (рис. 4). Вместе с тем, коэффициенты корреляции ряда наблюдений, усредненного по неделям, аппроксимируются гиперболической функцией, что свидетельствует о долгосрочной зависимости исходного ряда (рис. 5).

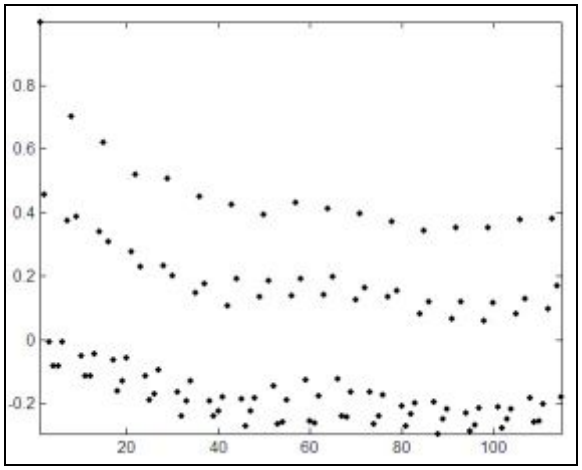

Рис. 4. Коэффициенты корреляции ряда наблюдений

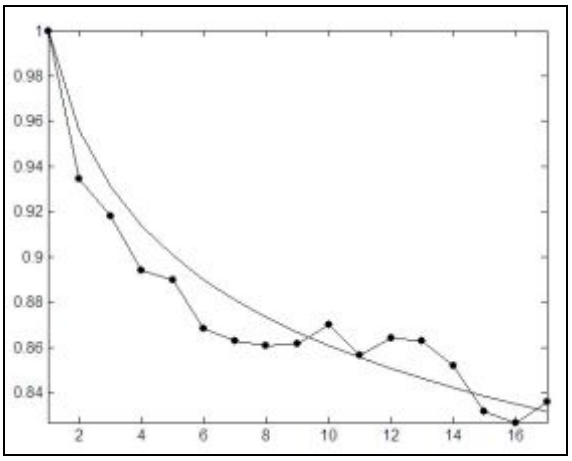

Рис. 5. Коэффициенты корреляции ряда наблюдений, усредненного по неделям

Фактор Фано

Для изучения поведения процессов и подтверждения их самоподобия принято использовать еще один показатель — индекс разброса дисперсии (IDC), так называемый, фактор Фано. Эта величина определяется как отношение дисперсии числа событий (в нашем случае — числа публикаций) временного ряда на заданном окне наблюдений k к соответствующему математическому ожиданию:

$$F(k) = \sigma^2(k) / m(k).$$

Для самоподобных процессов выполняется соотношение:

$$F(k) = 1 + Ck^{2H-1},$$

где C и H — константы. На рис. 6 приведен график значений $F(k)$ в двойном логарифмическом масштабе.

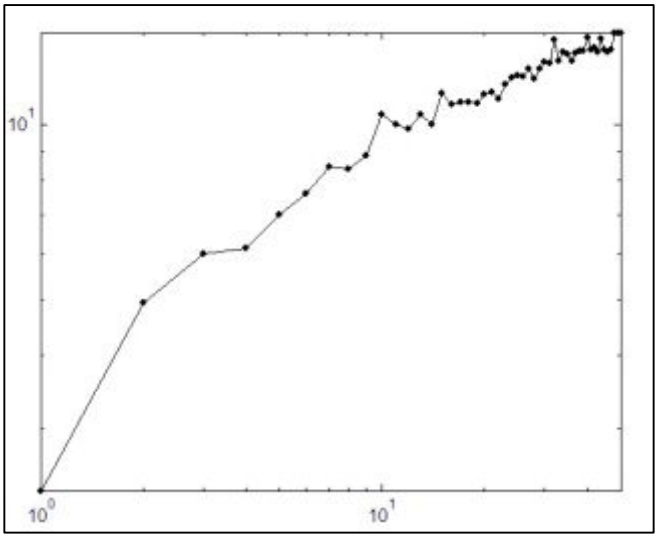

Рис. 6. Зависимость фактора Фано от ширины окна наблюдений

Показатель Херста

Основной характеристикой рядов, обладающих хаотическим поведением, является, как известно, показатель Херста [4]. Для его определения воспользуемся так называемым *R*/*S*-анализом, который успешно применялся авторами ранее в исследованиях фрактальной природы научных коммуникаций и информационных потоков [6]. Он позволяет достаточно эффективно исследовать свойства числовых рядов на основе отношения разброса значений к среднеквадратичному отклонению.

Пусть R — размах значений числового ряда, образуемого в нашем случае набором n значений ряда наблюдений, а S — его среднеквадратичное отклонение. Тогда имеет место соотношение:

$$R/S = (n/2)^H,$$

где H — показатель Херста, который для достаточно широкого класса рядов измерений связан с хаусдорфовой размерностью $D = 2 - H$.

Заключение

Численные значения H характеризуют различные типы коррелированной динамики (персистентности). При $H = 0{,}5$ наблюдается некоррелированное поведение ряда, а значения $0{,}5 < H < 1$ соответствуют антикорреляциям (чередование больших и малых величин в анализируемых данных).

Авторами были проведены исследования фрактальных свойств информационных потоков, для чего использовался документальный корпус системы мониторинга новостей из Интернет InfoStream [8]. Рассматривались ряды, соответствующие количеству публикаций в разрезе дат. На рис. 7 изображена зависимость нормированного размаха (*R*/*S*) от размерности подмножества документов в двой-

ном логарифмическом масштабе. Мы видим, что его значение хорошо аппроксимируется прямой. При этом показатель Херста стремится к величине ~0,67, что соответствует хаусдорфовой размерности ~1,33.

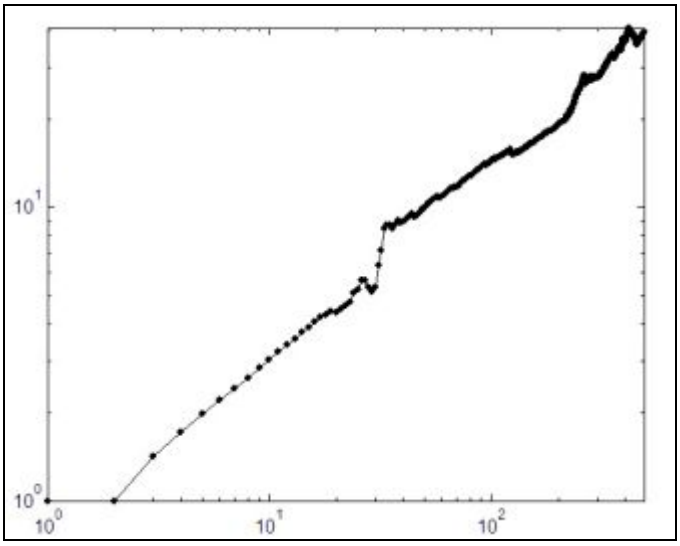

Рис. 7. Значения показателя нормированного размаха (ось Y) в зависимости от объема исследуемого массива (ось X)

Таким образом, проведенные авторами исследования тематических информационных потоков подтвердили предположение о самоподобии и итеративности процессов в информационном пространстве. Републикации, цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах.

В результате эксперимента было подтверждено наличие высокого уровня статистической корреляции в информационных потоках на продолжительных временных интервалах. В частности, на рассмотренном примере показана высокая персистентность процесса, что свидетельствует об общей тенденции увеличения публикаций по тематике компьютерной вирусологии, периодическое появление пиков, связанных, как правило, с новыми вирусными атаками (вирусы «Камасутра», Vbs_Valentine, Worm_Gift, «НИКСЕМ», мобильными, в частности, MMS-вирусами), а также общими антивирусными мероприятиями: выставками, конференциями (например, Softool), републикациями отчетов таких компаний как DrWeb, «Лаборатория Касперского», McAfee и др.

Анализ самоподобия информационных массивов, таким образом, может рассматриваться как технология, предназначенная для осуществления аналитических исследований с элементами прогнозирования, способная к экстраполяции полученных зависимостей.

1. *Додонов А.Г., Ландэ Д.В.* Организация сети информационных прокси-серверов // Реєстрація, зберігання і оброб. даних. — 2006. — Т. 8, № 3. — С. 24–31.

2. *Gianna M. Del Corso, Antonio, Francesco Romani.* Ranking a Stream of News // Proc. of the 14th International Conf. on World Wide Web. — Chiba (Japan). — 2005. — P. 97–106.

3. *Брайчевский С.М., Ландэ Д.В.* Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1. — 2005. — Вып. 11. — С. 21–33.

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

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

6. *Ландэ Д.В.* Фрактальные свойства тематических информацийнных потоков из Интернет // Реєстрація, зберігання і оброб. даних. — 2006. — Т. 8, № 2. — С. 93–99.

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

8. *Безруков Н.Н.* Компьютерная вирусология. — К.: УРЕ, 1991. — 416 с.

9. *Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н.* InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: Научно-методическое пособие. — К.: ООО «Старт-98», 2007. — 40 с.

10. *Peng C.-K., Havlin S., Stanley H.E., Goldberger A.L.* Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series // CHAOS. — 1995. — Vol. 5. — P. 82.

11. *Stanley H.E., Amaral L.A.N., Goldberger A.L., Havlin S., Ivanov P.Ch., Peng C.-K.* Statistical Physics and Physiology: Monofractal and Multifractal Approaches // Physica A. — 1999. — Vol. 270. — P. 309.

12. *Павлов А.Н., Сосновцева О.В., Зиганшин А.Р.* Мультифрактальный анализ хаотической динамики взаимодействующих систем // Изв. ВУЗов. Прикладная нелинейная динамика. — 2003. — Т. 11, №. 2. — С. 39–54.

Поступила в редакцию 12.06.2007