О.Г. Додонов, Д.В. Ланде, В.Г. Путятін

ІНФОРМАЦІЙНІ ПОТОКИ В ГЛОБАЛЬНИХ КОМП'ЮТЕРНИХ МЕРЕЖАХ

Національна академія наук України
Інститут проблем реєстрації інформації
(ІПРІ НАН України)

Додонов О.Г., Ланде Д.В., Путятін В.Г.

ІНФОРМАЦІЙНІ ПОТОКИ В ГЛОБАЛЬНИХ КОМП'ЮТЕРНИХ МЕРЕЖАХ

Київ 2009

УДК 681.3+519.83
ББК 22.18, 32.81, 60.54
І95

Рекомендовано до видання
Вченою радою Інституту проблем реєстрації інформації
НАН України
(протокол № 13 від 24 листопада 2009 року)

Рецензенти:
О.Є. Литвиненко - доктор технічних наук, професор
В.В. Мохор - доктор технічних наук, професор

І95 Додонов О.Г., Ланде Д.В., Путятин В.Г.
Інформаційні потоки в глобальних комп'ютерних мережах. – К.: Наукова думка, 2009. – 295 с.

Монографія присвячена питанням дослідження, моделювання та аналізу документальних інформаційних потоків, які формують змістовне наповнення сучасних глобальних мереж, насамперед, мережі Інтернет. Розглядаються теоретичні засади теорії інформаційних потоків, математичні моделі, застосування технологій інформаційного пошуку та змістовного аналізу текстів до інформаційних потоків, які є базою для побудови сучасних інформаційних сервісів.

В роботі досліджуються основні характеристики та властивості сучасних інформаційних потоків. Представлені сучасні методи аналізу часових рядів, що відповідають тематичним інформаційним потокам, застосування кореляційного, фрактального, дисперсійного та вейвлет-аналізу до параметрів інформаційних потоків. Наведені теоретичні засади інформаційного пошуку, змістовного аналізу інформації (Text Mining), методів агрегування даних та екстрагування понять.

Книга орієнтована на широке коло фахівців у галузі інформаційних технологій, студентів старших курсів, аспірантів.

ISBN 978-966-00-0973-9
ІПРІ НАН України
Замовлення № 9616
Тираж 350 прим.

УДК 681.3+519.83
ББК 22.18, 66.4, 60.54

ЗМІСТ

ПЕРЕДМОВА

Цю монографію написано для тих, кого цікавлять процеси, які протікають у сучасних інформаційних мережах, де безперервно породжуються та гинуть тематичні інформаційні потоки.

Орієнтація даної книги спрямована на змістовну складову інформаційного середовища, а саме на дослідження та моделювання інформаційних потоків, пов'язаних з деякими тематиками з реального світу.

Обсяги інформаційних потоків, в яких доводиться шукати крупиці необхідної, актуальної, готової до безпосереднього використання інформації для вирішення проблем, обумовлюють актуальність і значущість самого процесу пошуку, особливостям та основним моделям якого у книзі приділено багато уваги.

Треба відзначити, що разом із зростанням об'ємів інформації зростає і кількість інформаційних джерел. Одним з класів таких джерел виступає інформаційна складова мережі Інтернет. В рамках даної монографії інформаційні потоки в Інтернет розглядаються як полігон, інформаційний масив, динаміка і об'єми якого, зокрема, зумовили на даний час появу проблеми орієнтації в його динамічній частині. Тому як база для вирішення актуального завдання інтеграції сучасних інформаційних потоків вибрана саме новинна складова мережі Інтернет, саме бурхливий розвиток якої останнім часом породив ряд специфічних проблем, зв'язаних, в першу чергу, з швидким зростанням об'ємів даних, що підлягають зберіганню і обробці.

У ході написання даної роботи були розглянуті різні підходи, у тому числі динамічне та багатоагентне моделювання. Через велику кількість факторів (параметрів), що впливають на інформаційні потоки у глобальних мережах, у тому числі соціальних та психологічних, які важко піддаються формальному математичному опису, питання моделювання інформаційних потоків залишається і на даний час відкритим.

Відомо, що відповідність результатів реальності - це нагальна проблема будь-якого моделювання. На цей час основний метод перевірки адекватності моделей - ретроспективний аналіз, приклади застосування якого наведені у відповідних розділах цієї монографії. За допомогою доступної авторам системи моніторингу та інтеграції новинної інформації з Інтернет InfoStream були отримані дані, які відображують деякі тематичні інформаційні потоки за актуальними або вже застарілими тематиками.

На даний час підходи, властиві технології інтеграції інформаційних потоків, застосовуються адміністраторами веб-сайтів при формуванні колонок новин, студентами при написанні оглядових курсових робіт, маркетологами при аналізі ринків, політиками, бізнесменами, ученими - всіма, хто активно бере участь в сучасних інформаційних, політичних і бізнес-процесах.

Книга орієнтована на досить широке коло читачів: фахівців в області інформатики, безпеки, соціологів; вона буде також корисна й аналітикам, які при рішенні задач хочуть враховувати підходи, які застосовуються для моделювання та прогнозування інформаційних процесів. Сподіваємося, що ця книга виявиться також корисною при підготовці спеціальних навчальних курсів з питань теорій інформації та інформаційного пошуку.

ВСТУП

На даний час розвиток інформаційних комп'ютерних технологій призвів до різкого зростання обсягів інформації, яка має зберігатися, оброблятися, поширюватися у середовищі глобальних мереж. Зростання обсягу і динаміки інформаційного простору супроводжується багатократним дублюванням інформації, слабким її структуруванням, надлишком «інформаційного сміття», зростанням в цілому рівня інформаційного шуму. Тому завдання аналізу та узагальнення інформаційних потоків у комп'ютерних мережах для створення інформаційних ресурсів, інтеграції сучасних інформаційних потоків можна вважати найактуальнішим в умовах стрімкого розвитку економічних, політичних та суспільних процесів. Інформаційних ресурсів мережі Інтернет, з яких складаються інформаційні потоки, досить багато, до того ж вони систематично оновлюються. Загальний рівень їх впорядкованості набагато нижчий, ніж для випадку традиційних ЗМІ. Якщо до цього додати те, що значна частина мережних ресурсів практично не контролюється, стає ясно, що процеси, які відбуваються, достатньо складні та вимагають для свого вивчення застосування розвинених сучасних методів.

Основним змістом цієї роботи є опис взаємозалежної сукупності теоретичних та технологічних засад, досліджень та практичних робіт в областях збору, обробки, систематизації інформації, створення інформаційних ресурсів, орієнтованих на вирішення аналітичних задач, які мають входити до сучасних систем підтримки прийняття рішень на основі узагальнення змісту потоків даних з комп'ютерних мереж, зокрема з мережі Інтернет.

Цю книгу присвячено інформаційним потокам в комп'ютерних мережах. У кібернетиці інформаційні потоки вже традиційно розглядаються як потоки даних, які спрямовуються від джерела до приймача інформації через канали

передавання. Дійсно, саме таким чином здійснюється телекомунікаційні процедури. Але розглядаючи потоки даних, що протікають в глобальних комп'ютерних мережа, зокрема в Інтернет, в рамках цієї книги основну увагу приділено дослідженню процедур агрегації інформаційних ресурсів, які зберігаються на мережних концентраторах інформації (веб-серверах, ftp-серверах, базах даних тощо), їх подальшого накопичення у сховищах даних та забезпечення до них узагальненого доступу з боку кінцевих користувачів. Саме з цих позицій розглядаються інформаційні потоки у першій главі, де поряд з формальним математичним визначенням таких понять, як ентропія, кількість інформації, розглядаються питання, пов'язані інформаційними ресурсами, інформаційними мережами, наводяться моделі сучасного інформаційного простору, зокрема її динамічної частини.

Другий розділ присвячений математичним моделям інформаційних потоків, серед яких багато місця приділено, логістичній моделі тематичних потоків. Ця модель враховує «конкуренцію» тематик реального миру, відбивану у віртуальному просторі. В цьому розділі представлено так званні емерджентні підходи до моделювання, описано моделі, побудовані на базі теорії клітинних автоматів, зокрема модель дифузії інформації. Незважаючи на те, що абстрактні моделі інформаційних потоків багато у чому залежать від суб'єктивних уявлень розробників цих моделей, навіть такі результати можуть пояснити реальність у багатьох випадках краще, ніж звичайний життєвий досвід. Математичне моделювання інформаційних потоків є, безумовно, важливим і цікавим, особливо враховуючи те, що це питання до теперішнього часу залишається в науці відкритим.

Аналіз динаміки інформаційних потоків, що генеруються у веб-просторі, стає сьогодні одним з найбільш інформативних методів дослідження актуальності тих або інших тематичних напрямків. Розвиток інформаційної складової мережі Інтернет (точніше, веб-простору), динаміка та обсяги мережної складової обумовили на цей час появу проблеми орієнтації у інформаційному просторі, зокрема, у його динамічній частині. Разом з тим,

до цього часу не існувало визнаного теоретичного обґрунтування та технологічних засад процедур створення інформаційних ресурсів, орієнтованих на вирішення аналітичних задач, забезпечення інтегрованого оперативного доступу користувачів до великої кількості розрізненої інформації в мережі Інтернет.

Третій розділ охоплює опис сучасних методів аналізу інформаційних потоків, для чого розглядаються ряди з відповідних їхніх параметрів. Особу увагу приділено дослідженню самоподібності інформаційних потоків, їхніх фрактальних характеристик. Детально розглядаються застосування до аналізу інформаційних потоків кореляційного, дисперсійного, фрактального та вейвлет-аналізу.

Четвертий розділ присвячено питанням застосування основ традиційної теорії інформаційного пошуку до сучасних інформаційних потоків. Також розглянуті основні характеристики та основні методи ранжирування результатів пошуку, які базуються на мережному представленні масивів гіпертекстових документів. У цьому ж розділі представлені основні алгоритми пошуку в децентралізованих мережах, один з видів яких, а саме пірингові мережі на даний час є найбільшими мережами за обсягами інформації та трафіку.

У п’ятому розділі висвітлюється проблема узагальненого доступу до інформаційних ресурсів, яка вирішується шляхом агрегування мережної інформації, створення своєрідних концентраторів інформаційних ресурсів, побудови розподілених інформаційних сховищ.

У шостому розділі розглядаються засади аналізу інформації, зокрема, концепція глибинного аналізу текстів – Text Mining, яка включила технологічні та методологічні підходи контент-анализу, комп'ютерної лінгвістики, штучного інтелекту. Як окремі складові цієї концепції розглядаються алгоритми класифікації та кластеризації інформації. Приведені підходи до вирішення таких завдань, як узагальнення інформаційних потоків шляхами автоматичного реферування, виявлення і

аналізу взаємозв'язків понять, виявлення нових подій. У цьому ж розділі розглядаються також засади анотованого пошуку, яка базується на засобах глибинного аналізу текстів.

Лише поєднання інформаційного пошуку у великих мережних структурах із змістовним аналізом даних у єдиному технологічному ланцюжку дозволить підвищити якість опрацювання поточної інформації та ефективність інформаційної підтримки процесів прийняття рішень. На цей час основою для створення інтелектуального середовища вирішення аналітичних міждисциплінарних проблем можуть стати спеціальні автоматизовані системи опрацювання та узагальнення інформаційних потоків в комп'ютерних мережах.

Сьомий розділ присвячений опису технології контент-моніторингу новинної інформації з мережі Інтернет. У цьому розділі детально розглядаються архітектурні рішення систем, що базуються на цій технології, технологічних засад моніторингу інформації з мережі, побудови інформаційних сховищ, баз даних, інформаційно-пошукових систем, засобів глибинного аналізу текстів. У цьому розділі також наведений невеликий огляд наявних рішень.

1. ЗАСАДИ АНАЛІЗУ ІНФОРМАЦІЙНИХ ПОТОКІВ

Бурхливий розвиток інформаційних мережних технологій, зокрема, мережі Інтернет, у останній час породив ряд специфічних проблем, пов'язаних, у першу чергу, із швидким ростом обсягів даних, що підлягають обробці та зберіганню, та їх динамікою [3].

На початку існування World-Wide Web невелика кількість веб-сайтів оприлюднювала інформацію окремих авторів для відносно великої кількості відвідувачів (цей стан умовно позначається як веб-1). Сьогодні ситуація різко змінилася. Самі відвідувачі веб-сайтів приймають активну участь у створенні контенту (становлення веб-2.0), що, серед іншого, привело до різкого зростання обсягів і динаміки інформаційного простору.

Цей процес супроводжується:

- непропорційним зростанням рівня інформаційного шуму;
- надлишком паразитної інформації (що несанкціоновано додається до дійсно потрібної);
- слабим структуруванням інформації;
- многократним дублюванням інформації.

Традиційному веб-простору до того ж притаманні такі недоліки, як неможливість гарантування цілісності документів, обмеженість змістовного пошуку, домінування «прихованого» веб, доступ до якого обмежений.

Надалі у цій роботі буде часто застосовуватися поняття «інформаційного простору». На відміну від математичного визначення поняття «простір», у цій книзі під поняттям «інформаційні потоки» розуміється сукупність інформаційних ресурсів, технологій їх супроводження та використання, інформаційних і телекомунікаційних систем, що створюють інформаційну інфраструктуру.

На відміну від звичайного сховища інформації, інформаційний простір Інтернет характеризується великою кількістю прихованих у ньому неявних експертних оцінок, реалізованих у вигляді гіпертекстових посилань. Саме гіперпосилання є базою для побудови основних моделей Інтернет-простору. Для досягнення достатнього охоплення інформаційних ресурсів засобами інформаційно-пошукових систем необхідно враховувати наявну архітектуру всього Інтернет-простору. До кінця XX століття цією інформацією ніхто не володів. Першу «математичну карту» ресурсів і гіперзв'язків існуючого простору World Wide Web [23] було створено лише у листопаді 1999 року в Інституті пошуку та аналізу текстів, який входить у дослідницький підрозділ IBM, при співпраці з компаніями AltaVista, IBM та Compaq.

1.1. Інформаційні потоки в комп'ютерних мережах

Для дослідження сучасних інформаційних потоків в Інтернет, тобто потоків повідомлень, що публікуються на сторінках веб-сайтів, в соціальних мережах, блогах, тощо, має застосовуватися принципово новий інструментарій, тому що класичні методи узагальнення інформаційних масивів (класифікації, фазового укрупнення, кластерного аналізу тощо) не завжди здатні адекватно відбивати ситуацію щодо динамічної складової інформаційного простору. В цьому випадку мова йде не стільки щодо аналізу документального масиву фіксованого розміру, нехай навіть дуже великого, скільки про узагальнення динамічного потоку гіпертекстових даних.

Звичайно, більша частина інформації, яка представлена в Інтернет, знаходить свого споживача. Однак якщо розглядати всю сукупність мережних публікацій як якусь спільність по відношенню до конкретного користувача (або групи користувачів), то можна побачити ряд проблем, пов'язаних з повнотою, релевантністю та оперативністю одержання даних. Пошук, фільтрація, збір інформації в Інтернет вимагають достатньої кваліфікації персоналу та, на жаль, не можуть ураховувати всіх особливостей інформаційної структури мережі та представлення в ній даних. Це, у свою

чергу, веде до того, що жодного разу вибірка інформації з Інтернет не може вважатися репрезентативною.

При цьому інформаційний потік, що "споживається" організацією з Інтернет носить, як правило, виражену предметну спрямованість, яка характеризується областю інтересів даної організації. Пошук і попередня обробка інформації в ручному режимі - досить трудомісткий, а найголовніше, тривалий процес, який найчастіше не дає бажаного результату. Вирішення проблеми на практиці можливо шляхом створення автоматизованих систем збору, фільтрації та аналізу інформації, так званих «інтелектуальних посередників» між користувачем або корпоративною інформаційною системою та мережею Інтернет. Подібна система повинна здійснювати збір і селекцію інформації з Інтернет і створювати документальну базу даних, специфіковану предметною областю користувача, тобто виконувати функції інтеграції інформаційних потоків. Завантаження інформації в базу даних має супроводжуватися її класифікацією та структуризацією. Для подальшої інформаційно-аналітичної роботи користувачеві мають надаватися ефективні засоби навігації, пошуку та узагальнення інформації, яка зберігається у відповідній динамічній документальній базі даних.

Для початку назвемо деякі припущення, спрощення загальні для всього подальшого викладу.

Введемо формальне визначення інформаційного потоку, яке кореспондується з класичним визначенням з теорії інформації [38].

Розглянемо відрізок (a,τ) дійсної вісі (вісі часу), де $\tau > a$. Припустимо, що на цьому відрізку часу у відповідності з деякими закономірностями, характер яких будемо розглядати нижче, в мережі публікується деяке число інформаційних документів - k. На вісі часу їх координати позначимо як $\tau_1,\ \tau_2\ ,...,\ \tau_k\ \ (a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$. Інформаційним потоком будемо називати процес $N_\alpha(\tau)$, реалізація якого характерізує кількість точок

(документів), що з'явилися на інтервалі (a,τ), як функцію правого кінця відрізку τ. Відповідно цьому визначенню реалізація інформаційного потоку є неубутна щаблева функція. Функція $N_\alpha(\tau)$ завжди цілочисельна.

Наведене визначення на локальних часових областях відповідає дійсності, але не враховує такий ефект, як старіння інформації, яке протиречить «накопичувальній» здатності інформаційного потоку $N_\alpha(\tau)$ на великих проміжках часу.

Так визначений інформаційний поток враховує лише кількість інформаційних повідомлень, не залежно від їх змісту. Взагалі, визначення змісту, тематики окремих документів є досить суб'єктивним процесом. Для строгого моделювання тематичних інформаційних потоків використовують моделі, які відрізняють документи за окремими словами або словосполученнями (зазвичай їх називають термами, від англ. *Terms*).

У цій роботі фактично аналізуються сукупність елементарних одиниць змістовного наповнення інформації. Як така одиниця будемо використати документ. У вузьких рамках даної роботи не розрізняються поняття «документ», «повідомлення» або «публікація». Надалі буде переважно використатися термін «документ», оскільки він більше звичний в областях досліджень, пов'язаних з інформаційними потоками. Позначимо множину документів як:

$$D(\tau)=\{D_i,\ \ i=1,\ \ ...,N(\tau)\},$$

де D_i – документ із номером i, τ - час, $N(\tau)$ – кількість документів у потоці на момент τ. $D_i = \{ w_{ij} \}$, де w_{ij} - множина термів, що входять у документ D_i.

Передбачається, що новинні повідомлення мають властивіст застарівають, втрачаючи свою актуальність. Цей процес у деяких випадках можна моделювати експонентною залежністю, що досить часто підтверджується на практиці. Якщо припустити, що ступінь актуальності $\alpha\in[0,1]$ в момент публікації документа дорівнює 1 і λ - коефіцієнт

напіврозпаду рангу актуальності, тобто, що $\alpha = \exp(-\lambda \Delta t) = 1/2$, то Δt – проміжок часу (наприклад, у годинах), за який документ в інформаційному потоці через своє старіння втрачає актуальність наполовину. Якщо припустити, що документ у деякому тематичному новинному потоці за добу втрачає половину своєї актуальності, те маємо: $\exp(-\lambda \cdot 24) = 1/2$, і, відповідно, $\lambda \approx 0,029$. Актуальним може вважатися документ, у якого ступінь актуальності перевищує задане заздалегідь експертним шляхом граничне значення, наприклад, 0.01 [17].

Весь Інтернет-простір можна з достатньою часткою умовності розділити на дві складові - стабільну та динамічну, які мають дуже різні характеристики розвитку. Зокрема, процес старіння інформації у відомій моделі Бартона-Кеблера [17] описується рівняням, яке складається з двох компонент:

$$m(t) = 1 - ae^{-T} - b\,\mathrm{e}^{-2T},$$

де $m(t)$ – доля корисної інформації у загальному потоці через час *T*, перший від'ємник відповідає стабільним ресурсам, а другий – динамічним, новинним. Це рівняння також у повній мірі відповідає обсягам інформації, що публікуються в Інтернет за певними тематиками, які час від часу виникають і зникають. Стабільна складова Інтернет містить інформацію "довгострокового" плану, у той час, як динамічна складова містить ресурси, які постійно оновлюються. Деяка частина останьої складової згодом вливається в стабільну, однак більша частина "зникає" з Інтернет або попадає у сегмент так званого "прихованого" веб, не доступного користувачам за допомогою звичайних інформаційно-пошукових систем (ІПС).

У традиційній мережній інформаційно-пошуковій системі інформаційний простір, що складається зі стабільної та динамічної частин, та індексується за допомогою цієї ІПС, змінює своє наповнення через певну кількість днів: деякі новинні документи йдуть до стабільної частини у вигляді архівів, а інші зникають. У цьому випадку користувач при звертанні

до ІПС знаходить релевантні запиту документи зі стабільної частини, посилання з динамічної частини, які застаріли, та нічого не знаходить з обновленої динамічної частини.

На цей час жодна із традиційних пошукових систем у достатньому обсязі не допомагає в пошуку актуальної новинної інформації, що знаходиться в динамічній частині мережі Інтернет. Вирішення цієї задачи вимагає застосування системи-посередника між користувачем та мережею. Подібний посередник (або агент новин) повинен виконувати роботу зі збору, селекції інформації та забезпечувати передумови (здійснювати попередньу обробку) для створення документальної бази даних.

Принцип індексування, яке має здійснюватися цим посередником, трохи відрізняється від індексування традиційними пошуковими системами: індексується не весь контент мережі Інтернет, а тільки його динамічна частина. При цьому, за рахунок необхідності сканування відносно невеликого обсягу даних, частота індексування вибирається досить малою - від декількох хвилин до декількох годин (залежно від джерела). У результаті застосування подібного посередника виникає така ситуація: користувач одержує необхідні відповіді з новинної та із "застарілої" новинної частини, підтверджених документами зі створенної архівної бази даних, але не одержує повної вибірки документів зі стабільної частини мережі Інтернет.

Таким чином, проблема одержання повної інформації з Інтернет на даний час в ідеалі може бути вирішена шляхом використання двох інструментів - традиційних ІПС (для стабільної частини веб-простору) і системи інтеграції інформаційних потоків новин.

1.2. Інформаційні ресурси

Інформація виникає в мережі Інтернет не сама по собі. Її публікують, розміщують на веб-сайтах, сторінках соціальних мереж, вузлах пірингових мереж тощо. Надалі такі інформаційні ресурси будемо називати інформаційними джерелами. Забігаючи наперед, зазначимо, що система

контент-моніторингу InfoStream, яка є одним з практичних «полігонів» цього дослідження, споживає політематичний інформаційний потік з понад 4000 джерел. На рис. 1.1 наведено фрагмент динаміки цього потоку (вісь 0Z – кількість документів) у розрізі часу (діб, вісь 0Y) і найбільш продуктивних джерел (вісь 0X). На цьому графіку явно виражені тижневі коливання обсягів інформаційних потоків.

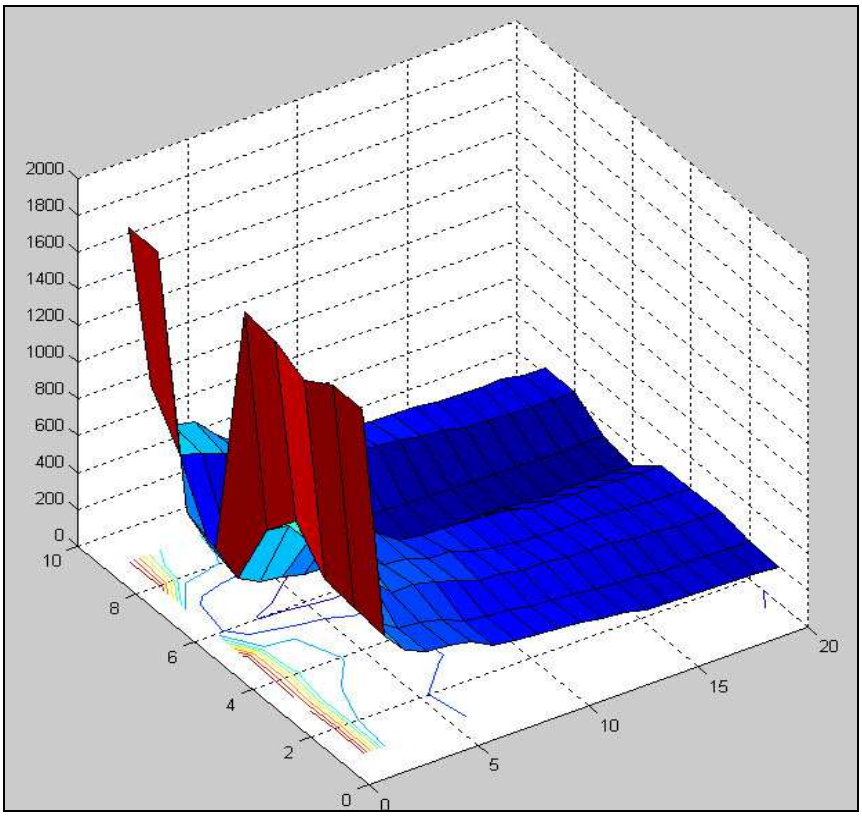


Рис. 1.1. Динаміка оприлюднення інформації найбільш рейтинговими джерелами (отримано за допомогою системи InfoStream)

На даний час потужні можливості Інтернет породжують проблему оптимізації складу та кількості джерел, які можуть використовуватися корпоративною інформаційною системою з метою забезпечення необхідної кількості та якості документів, які задовольняють потребам користувачів. В зв'язку з цим актуальними виявляються питання ранжирування і вибору джерел новинної інформації – веб-сайтів, до яких потрібно забезпечити доступ через один інтерфейс як в пошуковому режимі, так і в режимах перегляду та аналітичного узагальнення.

Принципам ранжирування веб-документов присвячена велика кількість наукових робіт і практичних розробок [17, 21]. Посилальне ранжирування веб-сайтів сьогодні є окремим напрямом інтернет-бізеса – SEO (search engine optimization). Разом з тим, питанням ранжирування і відбору інформаційних ресурсів з урахуванням їхнього контенту, об'ємів і стабільності тематики публікацій приділяється значно менша увага. Безумовно, основним критерієм при виборі джерел для таких систем моніторингу новин є їхній зміст. Розподіл джерел за контентом, який відповідає тематичним потребам корпоративних користувачів, задовольняє закону Бредфорда, відповідно, при відборі джерел обов'язково повинне враховуватися їх ранжирування за ступенем відповідності тематиці.

Проте, реалізація такого вибору веде до деяких складнощів. На практиці таке ранжирування здійснюється експертами шляхом оцінки кількості документів, релевантних наперед визначеному пакету тематичних запитів, які адресуються до фрагмента бази даних, складеної з документів джерела, щоаналізується. А це неминуче приводить до певного суб'єктивізму зі всіма витікаючими наслідками.

Тому представляється перспективним доповнити традиційний підхід об'єктивнішими та більш суворими методами, що дозволяють оптимізувати процес формування інформаційної бази систем інтеграції інформаційних потоків.

На рис. 1.2 приведено графік розподілу (у напівлогарифмічному масштабі) кількості документів, опублікованих джерелами, що скануються системою InfoStream, ранжируваними за параметром – кількістю документів, які опубліковано джерелом. Нижче приведені розподіли, що відносяться до масиву документів за березень 2008 року об'ємом понад 1.2 млн. документів із понад 2500 джерел – відкритих веб-сайтів. Центральна частина графіка добре апроксимується прямою, що свідчить про близькість представленої залежності до гіперболічної (тобто про дію узагальненого закону Ципфа). На рис. 1.3 приведено загальну кількість документів, що охоплюються системою

моніторингу залежно від джерел, що враховуються в ній, також ранжируваних по кількості опублікованих документів. Оскільки закон Ципфа припускає апроксимацію щільності розподілу гіперболічною залежністю вигляду a / x, то функція розподілу кількості документів

$$f(x) \sim \int \frac{a}{x} dx = a \ln x + C$$

у дозволеному наближенні описується логарифмічним законом.

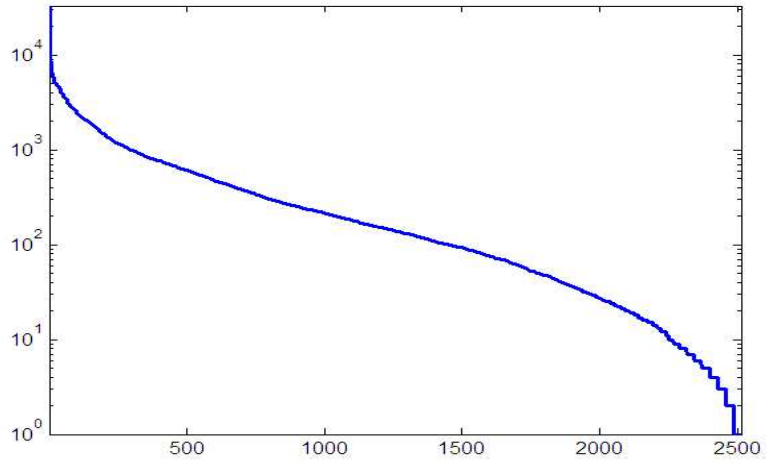


Рис.1.2. Розподіл кількості публікацій (вісь 0Y) за ранжируваним списком джерел (вісь 0X)

Приведена залежність дозволила побудувати критерій вибору необхідної частини джерел для різних корпоративних застосувань із загального списку, які задовільняють потреби користувачів.

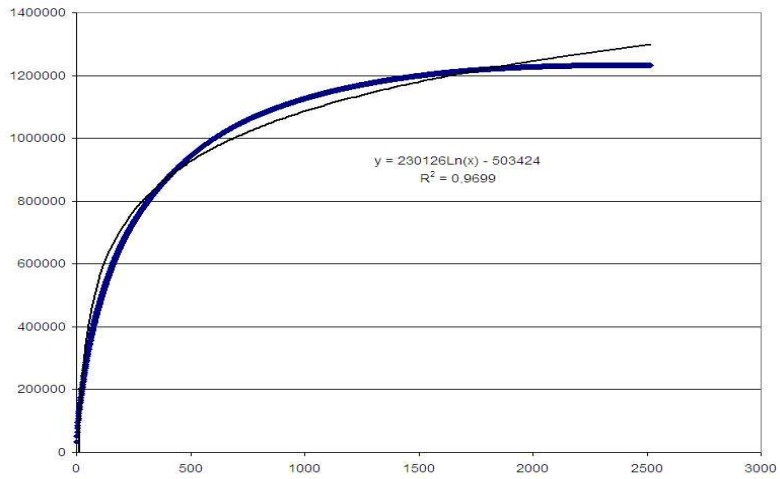


Рис.1.3. Кількість публікацій в системі моніторингу (вісь 0Y) залежно від джерел, ранжируваних за кількістю документів (вісь 0X)

Якщо припустити, що всі джерела давали б однаковий внесок за кількістю опублікованих документів, то дана залежність була б лінійною і виражалася б формулою:

$$f_{lin}(n) = n\frac{f_{\max}}{N},$$

де $f_{\max}$ - максимальний об'єм охоплюваних документів, N - загальна кількість джерел, n - номер поточного джерела.

Очевидно, що відхилення реальної залежності від лінійної спочатку зростає, а потім зменшується до нуля. Називатимемо кількість джерел пороговою n_p, коли значення реальної залежності максимально відхиляється від наведеної лінійної:

$$n_p = \arg\max \{f(n) - f_{lin}(n)\}.$$

На рис. 1.4 наведена ілюстрація значень n_p для різних значень N, тобто коли вибирається N найбільш продуктивних джерел. Що цікаво (і цілком відповідає характеру функції $f(n)$), значення n_p практично лінійно залежать від N (рис. 1.5): $n_p \sim 0.24N$, при цьому кількість охоплюваних документів, відповідних n_p при максимальній кількості джерел (2514, рис. 1.6) досягає 80 відсотків від $f_{\max}$.

При цьому можна зауважити, що побудована залежність задовольняє принципу Парето: приблизно 20% найбільш продуктивних джерел публікують 80% документів.

Як вже було відмічено, цитованість окремих документів і веб-сайтів сьогодні є одним з основних критеріїв оцінки рангів документів в мережевих пошукових системах (PageRank, HITS, TrustRank, ТІЦ тощо). Ідея оцінки рівня цитування дозволила реалізувати одну з перших моделей динамічної частини веб-простору [17]. Слід зазначити, що оцінка рівня джерела інформації як «автора» переважно по кількості веб-сайтів, з яких на нього

ведуть гіперпосилання, цілком узгоджується із запропонованим Мораном і Лемпелем алгоритмом Salsa.

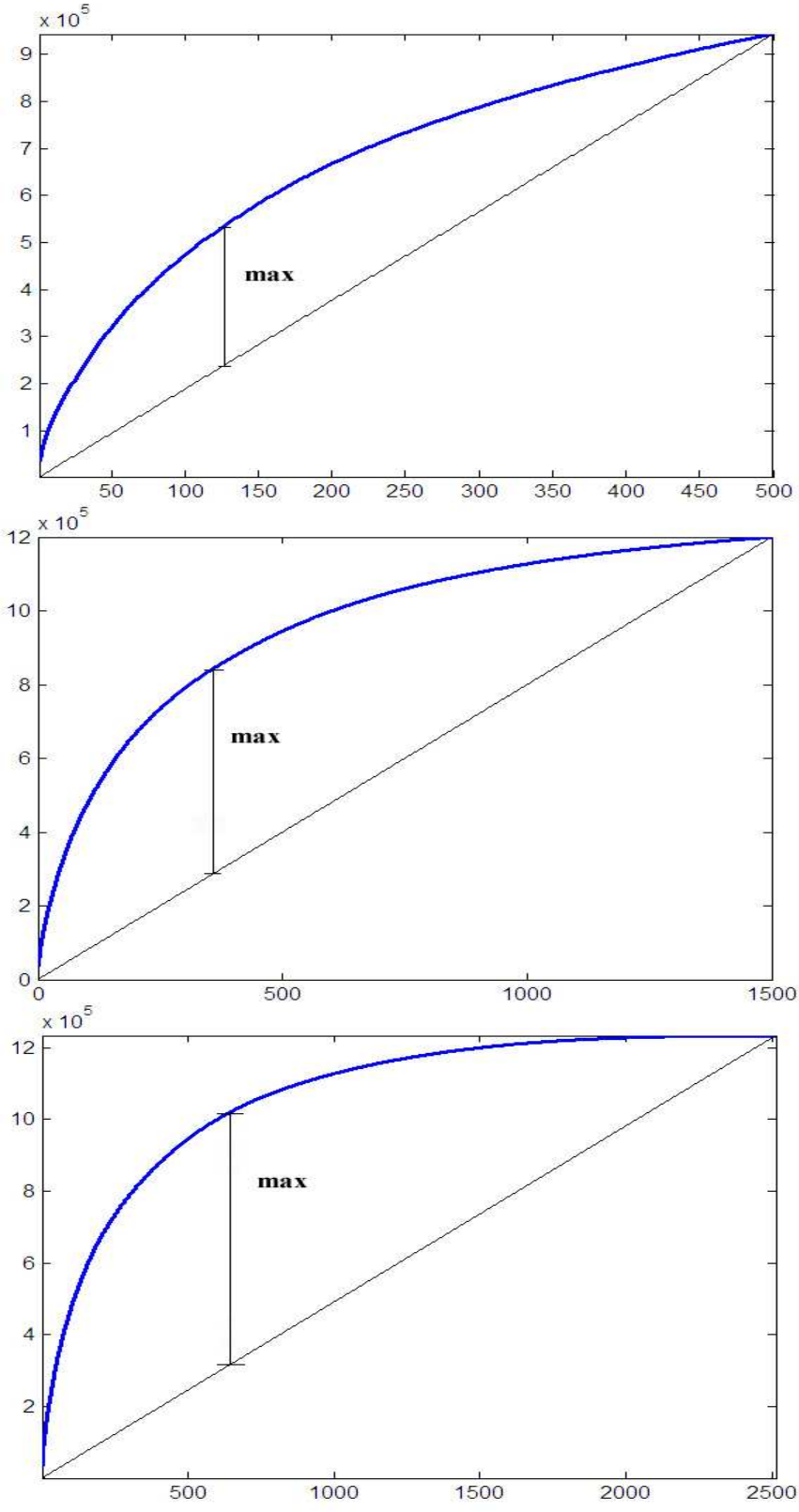


Рис.1.4. Кількість публікацій в системі моніторингу при підключенні нових найбільш інтенсивних джерел (500, 1500, 2500)

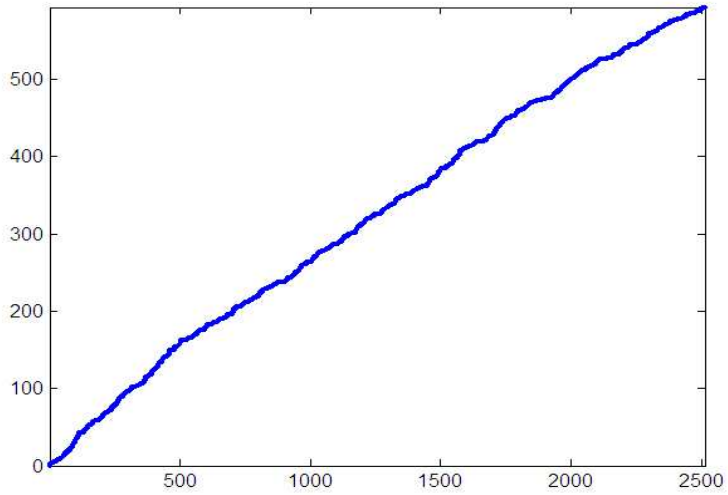


Рис.1.5. Зміна порогового значення (вісь 0Y) при зміні початкової кількості джерел (вісь 0X)

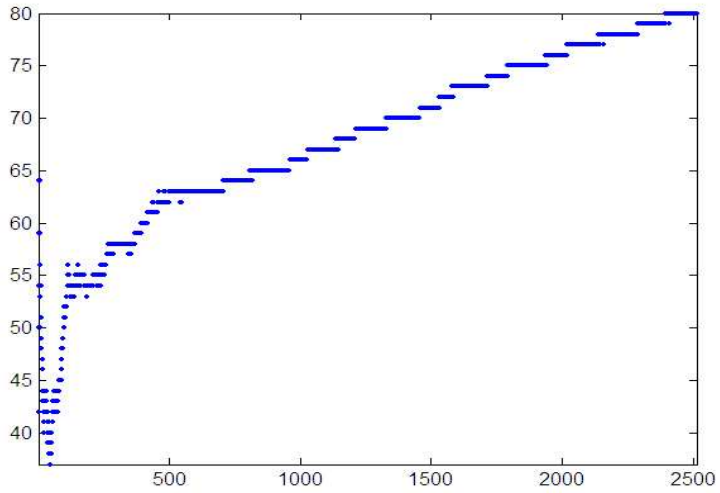


Рис.1.6. Питома кількість документів, що охоплюються системою (вісь 0Y), при зміні початкової кількості джерел (вісь 0X)

Спеціальне місце в дослідженні займало вивчення змістовного дублювання інформації. При цьому слід зазначити, що відсоток документів, що дублюються по сенсу, в системі моніторингу InfoStream значно менший, ніж у всьому новинному Web-просторі. Це пояснюється підбором джерел для сканування, до числа яких не входять багато новинних інтеграторів.

Як вже наголошувалося нами раніше, однією з головних особливостей новинної інформації є наявність великої кількості повідомлень, дублюючих

один одного. Так, про подію світового значення напишуть всі засоби масової інформації (ЗМІ), причому, швидше за все, на одній з перших сторінок. Споживач же (за винятком деяких специфічних напрямів аналітичних досліджень інформаційного простору) бажає отримувати по кожній події одне повідомлення.

Тому дослідження характеру і властивостей дублювання інформації набуває в сучасних технологіях виключно важливого значення. Зокрема, украй актуальним стає завдання відбору найбільш оригінальних джерел, що дозволяють (принаймні статистично) виключити не тільки формальне, але і змістовне дублювання інформації. Дублювання повідомлень на веб-сайтах залежить від різних причин, тому проведені вимірювання для ранжируваного по кількості публікацій списку джерел показують різний рівень, при цьому інформація не носить наочного характеру. Разом з тим, згладжування за допомогою методу ковзаючою середньою (з вікном спостереження, рівним 20), дозволив отримати графік (рис. 1.7), що наочно свідчить про стійку тенденцію: чим продуктивніше джерело інформації, тим більше він містить запозичень з інших джерел.

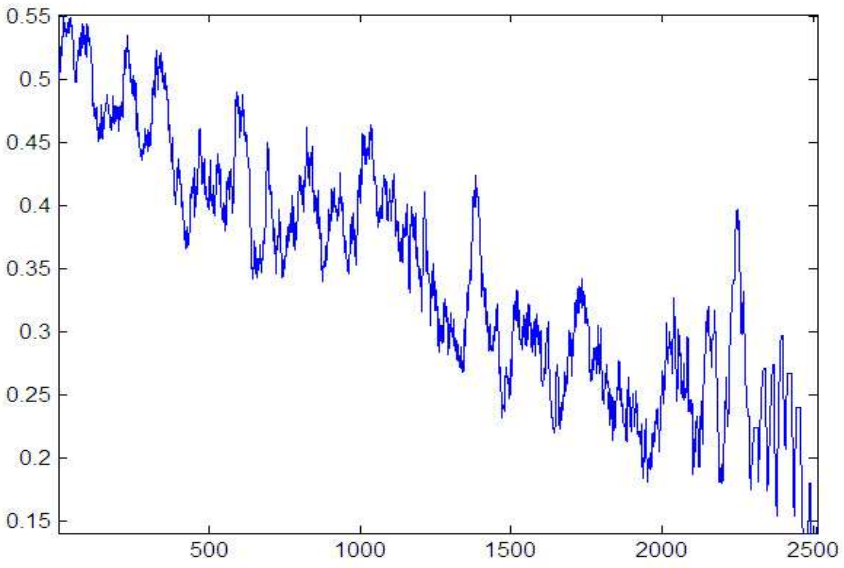


Рис.1.7. Усереднена питома кількість документів (вісь 0Y), що дублюються, за ранжируваним за кількістю публікацій списком джерел (вісь 0X)

Одній з важливих характеристик інформаційних джерел в новинному сегменті Інтернет є їх стабільність, яка розуміється як генерація постійного числа документів в одиницю часу. Прикладом стабільних джерел можуть служити провідні інформаційні агентства, що регулярно поставляють споживачам приблизно однакові об'єми інформації впродовж тривалого часу, а прикладом нестабільних – блоги, багато з яких активно діють протягом декількох днів, а потім згасає.

Природно, джерело, що регулярно випускає свою продукцію, з більшою вірогідністю відобразить в своїх публікаціях важливі події, чим джерело, що виходить нерегулярно, від випадку до випадку (він може просто «проскочити мимо події»).

З іншого боку, крупні видання, що забезпечують повноцінне освітлення подій, як правило на перше місце виводять масштабні, значні в суспільному розумінні події, про які без переваг можна дізнатися з будь-яких засобів масової інформації. Події ж меншої суспільної ваги, але при цьому, можливо, цікаві і важливі для окремих груп споживачів, або взагалі відсутні, або втрачаються «на останніх сторінках». Тому завдання оптимального урахування стабільності джерел зовсім не тривіальне і вимагає серйозних досліджень.

Нижче ми звернемося до одного з важливих її аспектів. Як було показано, щоденна загальна кількість документів, що публікуються на основних інформаційних веб-сайтах приблизно постійна, і коливається в основному залежно від дня тижня. Разом з тим тематикам публікацій притаманнаі істотни коливання.

Один з можливих підходів до вирішення проблеми ранжирування джерел інформації грунтується на підході, що полягає у вивченні динаміки тематичних інформаційних потоків, які породжуються ними. На практиці серед багатьох проблем вибору і аналізу джерел контента велике значення, зокрема, має урахування параметрів їх тематичної стабільності. При цьому тематична стабільність і стабільність публікації інформації джерелами часто

грають вирішальну роль при проведенні аналітичних досліджень. Наприклад, такі важливі властивості інформаційних джерел, як тематичну кореляцію і повноту, має сенс враховувати тільки для джерел стабільної тематичної спрямованості.

Тематичну стабільність джерела можна визначити як кореляцію наборів тематичних рубрик, яким відповідають документи з цього джерела в різні періоди часу. Природно, конкретний набір рубрик мало впливає на пропонований нижче метод розрахунку стабільності джерел (під тематичною рубрикою в даному випадку розуміється тематика, семантика якої, зокрема, знаходить своє віддзеркалення у вигляді запиту інформаційно-пошуковою мовою). Передбачається, що документу приписується та або інша рубрика, якщо він відповідає певному запиту. Перелік рубрик і відповідних ним запитів був вибраний авторами на підставі досвіду роботи з політематічними новинними ресурсами мережі Інтернет. Ці рубрики і запити встановлені і апробовані протягом тривалого часу в системі контент-моніторінга InfoStream. В даний час система включає 35 основних тематичних рубрик.

При дослідженні тематичної спрямованості деяких джерел інформації були виявлені документи, що відхиляються від основної спрямованості цих джерел. Такі документи, якщо їх кількість відносно невелика, не повинні впливати на рівень стабільності джерел, що розраховується нижче. Звичайно, автоматична рубрикація багато в чому залежить від якості запитів, проте деякими погрішностями в рубрикації при статистичному дослідженні можна нехтувати.

Для обчислення рівня стабільності джерела інформації використовувалася формула, заснована на так званому, R/S - аналізі [33]. Слід зазначити, що цей підхід має безпосереднє відношення до фрактального аналізу. R/S-анализ дозволяє досліджувати «порізанність» кривої, що утворюється, на основі відношення розкиду значень часового ряду до середньоквадратичного відхилення.

Був запропонований параметр K тематичної стабільності часового ряду інтенсивності публікацій на веб-сайтах (джерелах), який виглядає таким чином:

$$K = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i}{R_i},$$

де N - кількість тем (рубрик), що відповідають джерелу; S_i – середньоквадратичне відхилення за рубрикою i; R_i – розмах значень за рубрикою i.

Значення S_i обчислюється за формулою:

$$S_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\{r_j^{(i)} - \frac{1}{M}\sum_{k=1}^{M} r_k^{(i)}\right\}^2},$$

де $r_j^{(i)}$ – кількість входження рубрики і за день j, M - кількість значень ряду вимірювання (тижнів, наприклад).

Значення R_i обчислюється таким чином:

$$R_i = \max_{1<k<M} X_k^{(i)} - \min_{1<k<M} X_k^{(i)},$$

де $X_k^{(i)}$ – накопичене до моменту k відхилення за рубрикою i, яке обчислюється за формулою:

$$X_k^{(i)} = \sum_{j=1}^{k}(r_j^{(i)} - \frac{1}{M}\sum_{l=1}^{M} r_l^{(i)}).$$

На рис. 1.8 представлена крива значень коефіцієнтів стабільності для джерел, ранжируваних по цих значеннях. Зокрема, самими тематично стабільними документами (значення правої верхньої частини діаграми), опинилися періодичні професійні видання, такі як «Континент Сибирь», «Дзеркало тижня», «Російський Вісник», «Політичний журнал», «Влсть денег» тощо, які з певною періодичністю друкують постійну кількість повідомлень з тематик, розподілених в приблизно в однакових пропорціях. Підтвердилася гіпотеза стосовно того, що саме професіоналізм інформаційного джерела корелює з тематичною стабільністю. Зокрема, практично всі провідні інформаційні агентства, які продукують

політематічну інформацію, увійшли до складу найбільш тематично стабільних.

Окрім приведеної тематичної, досліджувалася простіша, діаграма внетематічного розподілу джерел, ранжирувана за коефіцієнтами стабільності. Отримані дані ще раз підтвердили той факт, що електронні видання більш схильні змінювати тематику публікацій, чим свої об'єми, виражені загальною кількістю публікацій.

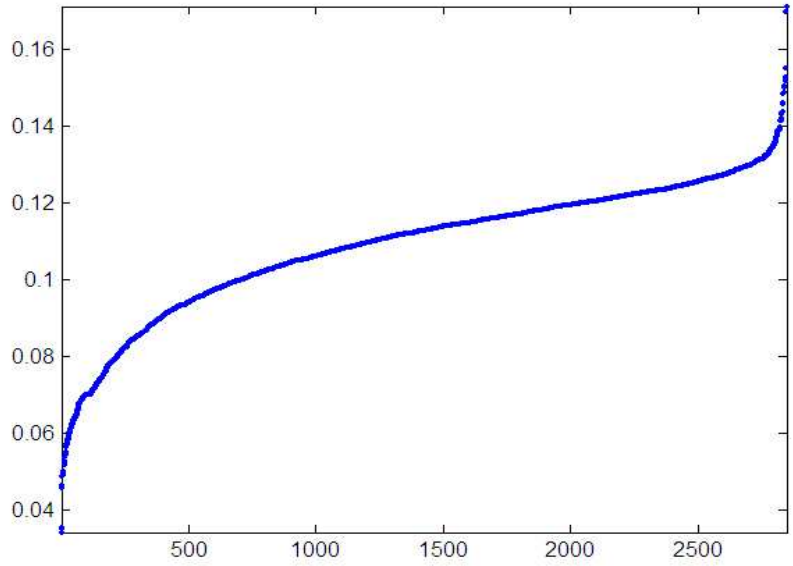


Рис.1.8. Ранжируваний список джерел (вісь 0X) за параметром тематичної стабільності (вісь 0Y)

Сьогодні стає ясно, що розробка якісно нових засобів роботи з мережними ресурсами переходить до розряду пріоритетних завдань. Зокрема, без розвинених засобів спостереження за мережевими інформаційними джерелами (моніторингу) неможливо забезпечити отримання відповідних репрезентативних вибірок, а це завдання сьогодні є одним з найактуальніших при відборі джерел для корпоративних застосувань систем контент-моніторінга.

Відзначимо лише декілька практичних застосувань ранжирування інформаційних джерел.

По-перше, це дасть можливість виявлення першоджерел інформації, наприклад, для розміщення в них рекламних матеріалів, матеріалів інформаційного впливу тощо.

По-друге, можна скоротити витрати часу і засобів шляхом ігнорування, виключення з пошуку та аналізу дійсно слабких, «шумових» джерел. Крім того, для оперативного знаходження актуальної інформації коректне ранжирування може сприяти знаходженню дійсно корисних першоджерел і служб синдикації інформації.

Результати даних досліджень джерел інформації можуть використовуватись при ранжируванні видачі інформаційно-пошукових систем, підрахунку медіа-рейтингів, дозволяють рекомендувати користувачам найбільш тематично стабільні і оригінальні джерела інформації, наприклад, для включення їх в список «персональних» в інтерфейсах систем контент-моніторингу інформаційних ресурсів.

Слід зазначити, що не дивлячись на те, що в даній роботі приведено декілька критеріїв ранжирування джерел інформації, остаточний «універсальний» критерій не приводиться. Теоретично його можна було б записати, наприклад, як лінійну комбінацію приведених критеріїв з деякими коефіцієнтами, які можуть визначатися експертно. Проте практика, що диктується інформаційними потребами корпоративних користувачів, показує, що при виборі джерел інформації зупиняються на одному з приведених критеріїв, доповнюючи його деякими неформальними міркуваннями (смакові переваги, облік регіональних чинників тощо).

1.3. Інформаційні мережі

Останнім часом все велику популярність отримує область дискретної математики, яка має назву «теорія складних мереж» (complex networks) [54]. Ця теорія охоплює вивчення параметрів мереж, враховуючи не тільки їх топологію, але і статистичні феномени, розподіл ваги окремих вузлів і ребер,

ефекти протікання і провідності в таких мережах струму, рідині, інформації тощо. Виявилось, що властивості багатьох реальних мереж істотно відрізняються від властивостей класичних випадкових графів.

Не дивлячись на те, що в розгляд теорії складних мереж потрапляють різні мережі – електричні, транспортні, інформаційні, найбільший внесок у розвиток цієї теорії внесли дослідження соціальних мереж. Термін «соціальна мережа» позначає зосередження соціальних об'єктів, які можна розглядати як мережа (або граф), вузли якої - об'єкти, а зв'язки - соціальні відносини. Цей термін був введений в 1954 році соціологом з «Манчестерської школи» Дж. Барнсом (J. Barnes) в роботі «Класи і збори в норвезькому острівному приході». У другій половині XX сторіччя поняття «Соціальна мережа» стало популярним у західних дослідників, при цьому як вузли соціальних мереж стали розглядати не тільки представників соціуму, але і інші об'єкти, яким властивий соціальні зв'язки. У теорії соціальних мереж отримав розвиток такий напрям, як аналіз соціальних мереж (Social Network Analysis, SNA). Сьогодні термін «соціальна мережа» позначає поняття, яке виявилось ширше за свій соціальний аспект, воно включає, зокрема, багато інформаційних мереж, у тому числі й WWW.

В рамках теорії складних мереж розглядають не тільки статистичні, але динамічні мережі, для розуміння структури яких необхідно враховувати принципи їх еволюції [59].

У теорії складних мереж виділяють три основні напрями: дослідження статистичних властивостей, які характеризують:

- поведінку мереж;
- створення моделей мереж;
- прогнозування поведінки мереж при зміні структурних властивостей.

У прикладних дослідженнях зазвичай застосовують такі типові для аналізу мереж характеристики, як розмір мережі, мережева щільність, ступінь центральності тощо.

При аналізі складних мереж як і в теорії графів досліджуються параметри окремих вузлів; параметри мережі в цілому; мережеві підструктури.

Для окремих вузлів виділяють наступні параметри:

- вхідний ступінь вузла - кількість ребер графа, які входять у вузол;
- вихідний ступінь вузла - кількість ребер графа, які виходять з вузла;
- відстань від даного вузла до кожного з інших;
- середня відстань від даного вузла до інших;
- ексцентричність (eccentricity) - найбільше з геодезичних відстаней (мінімальних відстань між вузлами) від даного вузла до інших;
- посередництво (betwetnness), що показує, скільки найкоротших шляхів проходить через даний вузол;
- центральность - загальна кількість зв'язків даного вузла по відношенню до інших.

Для розрахунку індексів мережі в цілому використовують такі параметри, як: число вузлів, число ребер, геодезична відстань між вузлами, середня відстань від одного вузла до інших, щільність - відношення кількості ребер в мережі до можливої максимальної кількості ребер при даній кількості вузлів, кількість симетричних, транзитивних і циклічних тріад, діаметр мережі - найбільша геодезична відстань в мережі і т.д..

Існує декілька актуальних завдань дослідження складних мереж, серед яких можна виділити наступні основні:

- визначення клік в мережі. Кліки - це підгрупи або кластери, в яких вузли зв'язані між собою сильніше, ніж з членами інших клік;
- виділення компонент (частин мережі), які зв'язані всередині і не зв'язані між собою;
- знаходження блоків і перемичок. Вузол називається перемичкою, якщо при його вилученні мережа розпадається на незв'язані частини;

– виділення угрупувань - груп еквівалентних вузлів (які мають максимально схожі профілі зв'язків).

Важливою характеристикою мережі є функція розподілу ступенів вузлів $P(k)$, яка визначається як вірогідність того, що вузол має ступінь $k_i = k$. Мережі, що характеризуються разними $P(k)$, демонструють вельми різну поведінку. $P(k)$ в деяких випадках може бути розподілами Пуассона ($P(k) = e^{-m} m^k / k!$, де m – математичне очікування), експоненціальним ($P(k) = e^{-k/m}$) або ступеневим ($P(k) \sim 1/k^{\gamma}$, $k \neq 0$, $\gamma > 0$).

Мережі із ступеневим розподілом ступенів вузлів називаються безмасштабними (scale-free). Саме безмасштабниі розподіли часто спостерігаються в реально існуючих складних мережах. При ступеневому розподілі можливе існування вузлів з дуже високим ступенем, чого практично не спостерігається в мережах з пуассоновим розподілом.

Відстань між вузлами визначається як кількість кроків, які необхідно зробити, щоб по існуючих ребрах добратися від одного вузла до іншого. Природно, вузли можуть бути сполучені прямо або опосередковано. Шляхом між вузлами d_{ij} назвемо найкоротшу відстань між ними. Для всієї мережі можна ввести поняття середнього шляху, як середнє по всіх парах вузлів найкоротшої відстані між ними:

$$l = \frac{2}{n(n-1)} \sum_{i>j} d_{ij},$$

де n - кількість вузлів, d_{ij} – найкоротша відстань між вузлами i та j.

Угорськими математиками П. Эрдешем (P. Erdős) і А. Реньи (A. Rényi) було показано, що середня відстань між двома вершинами у випадковому графові росте як логарифм від числа вершин [46].

Деякі мережі можуть виявитися незв'язковими, тобто знайдуться вузли, відстань між якими виявиться нескінченною. Відповідно, середній шлях може опинитися також рівним нескінченності. Для обліку таких випадків

вводиться поняття середнього інверсного шляху між вузлами, що розраховується за формулою:

$$il = \frac{2}{n(n-1)} \sum_{i>j} \frac{1}{d_{ij}}.$$

Мережі також характеризуються таким параметром як діаметр або максимальний найкоротший шлях, рівний максимальному значенню зі всіх d_{ij}.

Д. Уаттс (D. Watts) і С. Строгатц (S. Strogatz) в 1998 році визначили такий параметр мереж, як коефіцієнт кластерності [63], який відповідає рівню зв'язності вузлів в мережі. Цей коефіцієнт характеризує тенденцію до утворення груп взаємозв'язаних вузлів, так званих клік (clique). Крім того, для конкретного вузла коефіцієнт кластеризації показує, скільки найближчих сусідів даного вузла є також найближчими сусідами один для одного.

Коефіцієнт кластерності для окремого вузла мережі визначається таким чином. Хай з вузла виходить k ребер, які сполучають його з k іншими вузлами, найближчими сусідами. Якщо припустити, що всі найближчі сусіди сполучені безпосередньо один з одним, то кількість ребер між ними складала б $\frac{1}{2}k(k-1)$. Тобто це число, яке відповідає максимально можливій кількості ребер, якими могли б з'єднуватися найближчі сусіди вибраного вузла.

Відношення реальної кількості ребер, які сполучають найближчих сусідів даного вузла до максимально можливому (такому, при якому всі найближчі сусіди даного вузла були б сполучені безпосередньо один з одним) називається коефіцієнтом кластерності вузла i – $C(i)$. Природно, ця величина не перевищує одиниці.

Коефіцієнт кластерності може визначатися як для кожного вузла, так і для всієї мережі. Відповідно, рівень кластерності всієї мережі визначається як нормована по кількості вузлів сума відповідних коефіцієнтів окремих вузлів. Розглянутий нижче феномен «малих світів» безпосередньо пов'язаний з рівнем кластерності мережі.

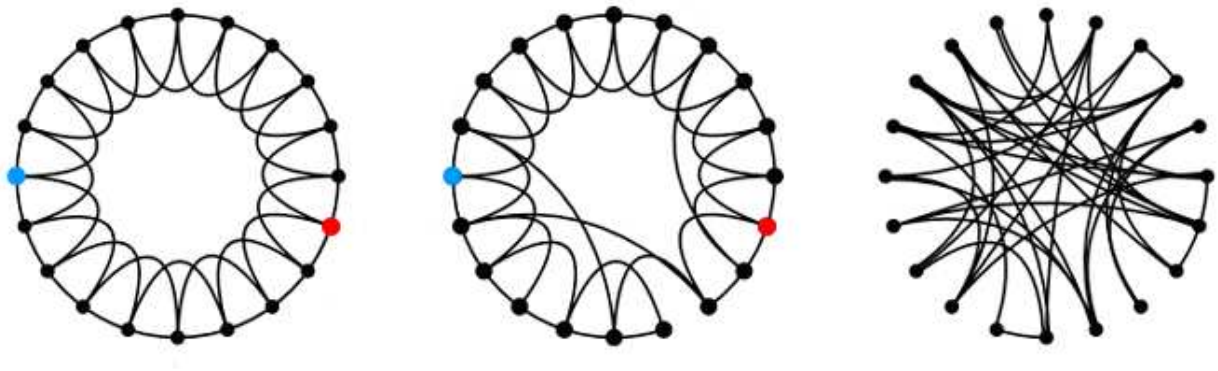
Посередництво (betweenness) – це параметр, що показує, скільки найкоротших шляхів проходить через вузол. Ця характеристика відображає роль даного вузла у встановленні зв'язків в мережі. Вузли з найбільшим посередництвом грають головну роль у встановленні зв'язків між іншими вузлами в мережі. Посередництво b_m вузла m визначається по формулі:

$$b_m = \sum_{i \neq j} \frac{B(i,m,j)}{B(i,j)},$$

де $B(i,j)$ - загальна кількість найкоротших шляхів між вузлами i та j, $B(i,m,j)$ - кількість найкоротших шляхів між вузлами i та j, що проходять через вузол m.

Не дивлячись на величезні розміри деяких складних мереж, в багатьох з них (і в WWW, зокрема) існує порівняно короткий шлях між двома будь-якими вузлами – геодезична відстань. У 1967 р. психолог С. Милгран в результаті виконаних масштабних експериментів обчислив, що існує ланцюжок знайомств, в середньому завдовжки шість, практично між двома будь-якими громадянами США [53].

Д. Уаттс і С. Строгатц виявили феномен, характерний для багатьох реальних мереж, названий ефектом малих світів (Small Worlds) [63]. При дослідженні цього феномена ними була запропонована процедура побудови наочної моделі мережі, якою властивий цей феномен. Три стани цієї мережі представлено на рис. 1.9: а) - регулярна мережа - кожен вузол якої сполучений з чотирма сусідніми; б) - та ж мережа, у якої деякі «ближні» зв'язки випадковим чином замінені «далекими» (саме в цьому випадку виникає феномен «малих світів») і в) - випадкова мережа, в якій кількість подібних замін перевищила деякий поріг.



а) *б)* *в)*

Рис. 1.9. Модель Уаттса-Строгатца

На рис. 1.10 приведені графіки зміни середньої довжини шляху і коефіцієнта кластеризації штучної мережі Д. Уаттса і С. Строгатца від вірогідності встановлення «далеких зв'язків» p (у напівлогарифмічній шкалі).

У реальності виявилось, що саме ті мережі, вузли яких мають одночасно деяку кількість локальних і випадкових «далеких» зв'язків, демонструють одночасно ефект малого миру і високий рівень кластеризації.

WWW є мережею, для якої також підтверджений феномен малих світів. Аналіз топології веб, проведений Ши Жоу (S. Zhou) і Р. Дж. Мондрагоном (R.J. Mondragon) з Лондонського університету, показав, що вузли з великим ступенем витікаючих гіперпосилань мають більше зв'язків між собою, чим з вузлами з малим ступенем, тоді як останні мають більше зв'язків з вузлами з великим ступенем, чим між собою. Цей феномен був названий "клубом багатих" (rich-club phenomenon). Дослідження показало, що 27% всіх з'єднань мають місце між всього 5% найбільших вузлів, 60% доводиться на з'єднання інших 95% вузлів з 5% найбільших і лише 13% - це з'єднання між вузлами, які не входять в ті, що лідирують 5%.

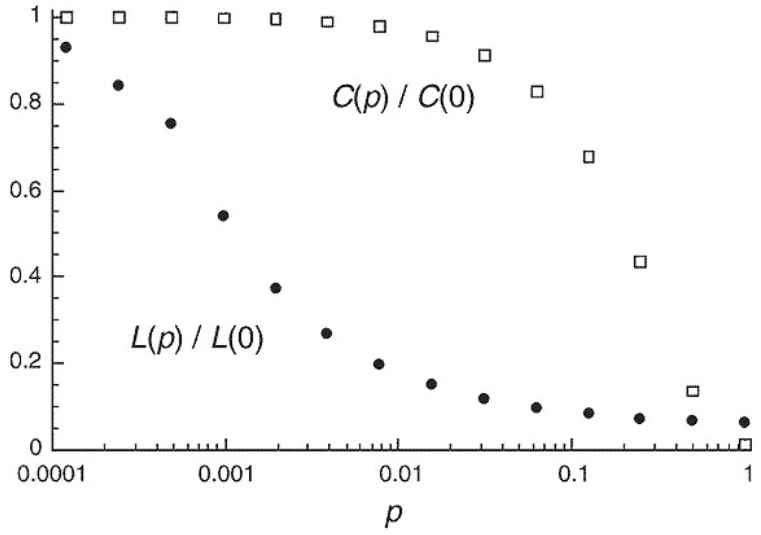


Рис. 1.10 Динаміка зміни довжини шляху і коефіцієнта кластерності

Ці дослідження дають підстави вважати, що залежність WWW від великих вузлів значно істотніша, ніж передбачалося раніше, тобто вона ще чутливіша до зловмисних атак. З концепцією «малих світів» зв'язаний також практичний підхід, званий «мережевою мобілізацією», яка реалізується над структурою «малих світів». Зокрема, швидкість розповсюдження інформації завдяки ефекту «малих світів» в реальних мережах зростає на порядки в порівнянні з випадковими мережами, адже більшість пар вузлів реальних мереж сполучені короткими шляхами.

Крім того, сьогодні досить успішно вивчаються "малі світи", що масштабуються, статичні, ієрархічні, і інші мережі, досліджуються їх фундаментальні властивості, такі, як стійкість до деформацій і перколяция. Недавно було показано, що найбільшу інформаційну провідність має особливий клас мереж, званих "заплутаними" (entangled networks). Вони характеризуються максимальною однорідністю, мінімальною відстанню між будь-якими двома вузлами і дуже вузьким спектром основних статистичних параметрів. Вважається, що заплутані мережі можуть знайти широке застосування в області інформаційних технологій, зокрема, в нових поколіннях веб, дозволяючи істотним чином понизити об'єми мережевого трафіку.

1.4. Моделювання інформаційного простору

Одне з найактуальніших завдань, що стоїть перед сучасним науковим співтовариством складається у побудові чіткої моделі сучасного інформаційного простору, яка має базуватися на досягненнях в області лінгвістики та інформатики, а також на відповідному математичному інструментарії.

Разом з тим дослідження сучасних інформаційних потоків можуть становити чималий інтерес як для лінгвістів, математиків, так навіть і для фізиків, наприклад, у плані аналогового моделювання статистичних процесів, у тому числі складних нелінійних систем з елементами самоорганізації. Перспективи охоплення інформаційного простору також будуть залежати від створення та розвитку ефективної інфраструктури, у рамках якої будуть функціонувати програмні продукти з боку Web-серверів і користувачів.

Слід зазначити, що як вся інформаційна мережа WWW, так і її окремі фрагменти і навіть сайти несуть значне соціальне навантаження, яке дозволяє порівнювати їх змістовному рівні з соціальними мережами, освіченими відносинами людей або цитуванням в науці. Веб, будучи, напевно, найдинамічнішою частиною інформаційного простору, характеризується великою кількістю прихованих в нім неявних експертних оцінок, реалізованих у вигляді гіперпосилань.

Тому WWW можна з повним правом вважати соціальною мережею, дослідження якої можна проводити, базуючись на існуючому підході аналізу таких мереж - SNA. Багато мережевих служб, які дозволяють людям встановлювати зв'язки в Мережі, автоматично формують соціальні мережі. Крім того, сьогодні бурхливо розвинувся спеціальний сервіс по цілеспрямованій побудові соціальних мереж у веб-просторі.

У листопаді 1999 долі один з керівників Інституту пошуку й аналізу текстів, що входити у дослідницький підрозділ IBM, А. Бредер (Andrei Broder) та його співавтори з компаній AltaVista, IBM та Compaq зробілі прорів, математічно описавши "карту" ресурсів і гіперзв'язків існуючого простору World Wide Web [43]. Дослідження спростувалі розхожу думання, начебто Інтернет - це єдиний густій простір.

Простеживши за допомогою пошукового механізму AltaVista понад 200 млн. Web-сторінок і декілька млрд. посилань, розміщених на цих сторінках, учені прійшлі до наступніх висновків про структуру Web-простору, який відповідає, на їхню думання, орієнтованому графу з топологією "краватки-метеліка" (Bow Tie), у якому вершині відповідають сторінкам, а ребра - з'єднуючим сторінки гіперпосиланням (рис.1.11). У рамках цієї моделі структурі зв'язків між окремімі Web-сторінками було віявлене:

- центральне ядро (28% Web-сторінок) - компоненту сильної зв’язності (SCC).
- 22% Web-сторінок - це "відправні Web-сторінки" (IN).
- 22% - "кінцевих Web-сторінок" (OUT).
- 22% Web-сторінок - відростки - повністю ізольовані від центрального ядра.

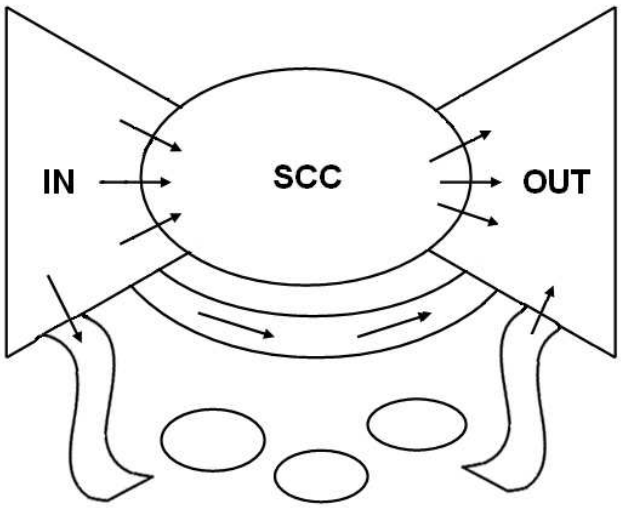


Рис. 1.11. Модель веб-простору Bow Tie

Існують також і "острови", які взагалі не перетинаються з іншими ресурсами Інтернет. Єдиний спосіб виявити ресурси цієї групи - знати адреси. Ніякі пошукові машини не зможуть знайти ці острови, якщо вони у минулому якимсь образом не з'єднувалися з іншими частинами Інтернет.

Були досліджені такі параметри даної моделі, як середня кількість сайтів, через які зв'язуються будь-які два сайти гіперпосиланнями, а також розподіл вхідних і вихідних посилань.

Топологія й характеристики моделі виявилися приблизно однаковими для різних підмножин Web-простору, підтверджуючи тим самим спостереження про те, що властивості структури всього Web-простору «Bow Tie» також вірні і його окремі підмножини. Таким чином, алгоритми, що використовують інформацію про структуру Web-простору, приблизно будуть працювати й на окремих його підмножинах.

Були досліджені такі параметри моделі Bow Tie, як середня кількість сайтів, через які зв'язуються будь-які два сайти гіперпосиланнями, а також розподіл вхідних і витікаючих посилань. Виявилось, що розподіл ступенів вузлів (вхідних і витікаючих гіперпосилань) веб-простору (досліджувалися сайти домена edu в кількості 325729) підкоряється статечному закону, тобто вірогідність того, що відповідний ступінь вершини рівний i, пропорційна $1/i^k$ (для вхідних посилань $k \approx 2.1$, а для витікаючих $k \approx 2.45$). Крім того, виявилось, що мережа WWW є «малим миром» з середньою довжиною найкоротшого шляху, рівною 11 і відносно великим значенням коефіцієнта кластерності, приблизно рівним 0.15 (для класичного випадкового графа це значення склало б 0.0002).

Останнім часом взяла веліку популярність теорія фракталів і детермінованого хаосу, що знаходіть свої застосування в різних областях, у тому числі й при аналізі простору WWW. Топологія та характеристики моделі Bow Tie віявіліся пріблізно однаковімі для різних підмножин Web-простору, побічно підтверджуючи тім самим його фрактальну природу у

тому плані, що властивості структурі всього Web-простори виявіліся також вірні і для його окреміх підмножин.

Таким чином, алгорітмі, що вікорістають інформацію щодо структурі Web-простору, повинні працюваті також на окреміх його підмножинах. Ця властивість структурі Web-простору сьогодні вже досить широко використовується при рішенні багатьох завдань, напріклад, для оптимізації ефективності механізмів ськанування, при побудові новіх Web-сервісів, для рішення завдань аналізу та прогнозу.

Разом з тім, моделі "Bow Tie" притаманні суттєві недоліки, такі як слабке урахування ресурсів «пріхованого» Web, динамічної складової Web-простору ігнорування поняття змістовного дублювання документів, контекстніх (не гіпертекстових посилань). До таких ресурсів, зокрема, відносяться деякі динамічно формовані веб-сторінки і документи з баз даних.

В зв'язку з цим необхідно підкреслити деяку некоректність розрахунку об'ємів «островів» по Бредеру через те, що список веб-ресурсов був отриманий з бази даних системи AltaVista, отриманий в результаті роботи програми-робота, скануючого веб-ресурси, переходячи від одного до іншого по гіперпосиланнях.

В даний час широкого поширення набули каталоги «прихованого» веб. Також здійснюються спроби доступу до об'єктів «прихованого» веб через спеціалізовані системи пошуку.

Ці недоліки обумовили потребу у створенні моделі динамічної складової Web-простору, більш конкретно - моделі новинного Web-простору, яку наведено нижче.

Л. Бйорнеборном (L. Björneborn) була запропонована модель «пом'ятого веб», яка асоціюється з пом'ятим папером. При цьому шлях між вибраними крапками на пом'ятому папері найчастіше коротше, оскільки протилежні частини листа паперу сполучені разом. Відповідно до цієї моделі кожне нове гіперпосилання змінює всі існуючі зв'язки, створюючи нові деформації

«пом'ятої» мережі. Тобто кожен новий гіперзв'язок - «гачок», який розтягує або деформує форму існуючої мережі WWW.

Модель динамічної частини веб-простору [19] базується на ідеї оцінки рівня цитування окремих інформаційних джерел, урахування як гіпертекстових так і контекстних посилань. Ця модель природним чином поєднує в собі змістовний аспект з можливістю урахування кількісних параметрів, значення яких визначаються цілком об'єктивно.

При вивченні цитованості новинних веб-ресурсов як інформаційних джерел необхідно враховувати ряд умов, актуальних для таких джерел. Можна було б просто застосувати модель А. Бредера і критерій типа PageRank до новинної складової Web-простору, проте такий підхід не можна вважати коректним з ряду причин:

- новинні потоки характеризуються підвищеною динамікою, що сильно впливає на природу гіперпосилань. Наприклад, на найбільш актуальні повідомлення протягом певного часу посилань може взагалі не існувати;
- модель Бредера слабо враховує особливості «прихованого» Web, тобто тих інформаційних Web-ресурсів, на які не існує прямих гіперпосилань (свого часу їм розглядалися лише ресурси, вже охоплені пошуковою системою AltaVista);
- у новинних потоках необхідно враховувати не тільки гіперпосилання, але і посилання контекстні, причому не тільки на об'єкти з відкритої частини Web-простору (це можуть бути часто посилання на ресурси, доступні тільки по паролю, або навіть оффлайновиє публікації видань, можливо і присутніх в Інтернет);
- крім того, модель Бредера не включає такого поняття, як змістовне дублювання інформації.

Було отримано розподіл новинних джерел по кількості веб-сайтів, що мають на них посилання. Всього за місяць посилання вказували на 1459 джерел (без самоцитування). Виявилось, що на перші 100 джерел ведуть

посилання з більше 80% веб-сайтів. На рис 1.12 представлений графік ранжируваного розподілу новинних джерел веб-сайтів по кількості сайтів, що мають на них посилання. Слід звернути увагу на те, що наведений графік дозволяє достатньо чітко виділити дві зони, що мають різні статистичні характеристики (кути нахилу на графіці): перша зона (ліва на рис.1.12), що включає інформаційні агентства і найбільші видання і друга, яка відповідає сайтам госорганів, спеціальних видань, компаній, що публікує прес-релізи.

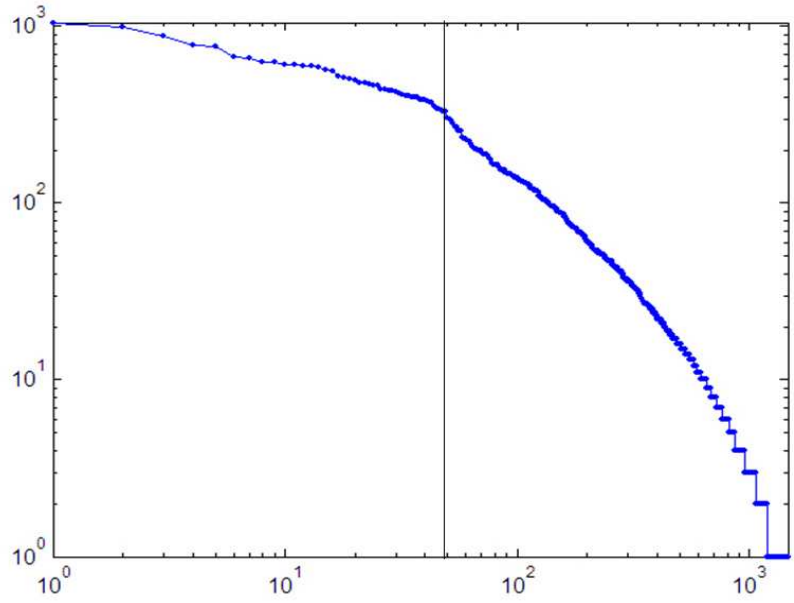


Рис.1.12. Залежність кількості веб-сайтів (вісь 0Y), з яких йдуть посилання, від рангу джерел (вісь 0X) в логарифмічній шкалі

Для кожного із джерел було складено запит у наступному форматі [19]:

<код джерела>#<шаблон для пошуку>[#...#<шаблон для пошуку>],

сукупність цих запитів було об'єднано в конфігураційному файлі, фрагмент якого представлений нижче:

srd00001#укроп#ukrop.com

srd00002#BBC#bbc.co.uk

srd00004#Crashes.ru#crashes.ru

srd00006#"Немецкая волна#Deutsche Welle#dwelle.de

srd00011#InoPressa#Инопресса#inopressa.ru

srd00015#Lenta.Ru#Лента.ру#lenta.ru

У результаті спеціальної обробки такого пакета запитів для кожного повідомлення, що входить до певного джерела - веб-сайту, були виявлені вхідні посилання на інші джерела (посилання на власне джерело виключалися). Було виявлено, що вхідні контекстні посилання були присутні у 484945 повідомленнях з 2323 Web-сайту. Було виявлено також 54 джерела, що не входять у цей список, тобто тих, на які не вело жодне з контекстних посилань і повідомлення яких не посилалися ні на один з досліджуваних Web-сайтів.

Такі Web-сайти («абсолютні острови») були винесені за рамки моделі. У таблиці 1.1 наведено список Web-сайтів, з яких веде максимальна кількість посилань.

Таблиця 1.1. Веб-сайти, з яких йде найбільше посилань

Web-сайт	Кількість посилань
RAMBLER	363
VLASTI.NET	271
RosInvest	270
"Обозреватель"	231
ИА "REGNUM"	217
«Деловая пресса"	202
“Россия-Он-Лайн"	193
"Оглядач"	191
RNews	183
Fin.org.ua	166
PRESIDENT.ORG.UA	164
"Промислово-торговельні новини"	160
"4 ВЛАДА"	159
"Україна промислова"	156

Також був отриманий розподіл новинних Web-сайтів за кількістю вхідних посилань. Усього за лютий посилання вказували на 1470 джерел (без самоцитування). Виявилося, що на 100 джерел веде понад 80% посилань.

У таблиці 1.2 наведений початковий фрагмент ранжируваного списку джерел, на які веде максимальна кількість посилань.

Крім того, були виявлені джерела, на які не посилаються, але які мають вихідні посилання (393) і джерела, що цитуються та не посилаються ні на кого (332).

Спеціальне місце в дослідженні займало вивчення змістовного дублювання інформації. При цьому слід зазначити, що відсоток повідомлень, що дублюються, у системі InfoStream значно менше, ніж у всьому новинному Web-просторі. Це обумовлюється підбором джерел для сканування, у число яких не входять багато з новинних інтеграторів.

Виявлення дублюючих за змістом новинних повідомлень у системі InfoStream виконується на основі лінгвостатистичних методів, що базуються на виявленні найбільш вагомих слів у документах, які виступають своєрідними ключами. Досвід показав, що в російсько- та україномовних потоках новин збіг 6 найбільш вагомих ключових слів у документах, з більш ніж 95% імовірністю свідчить про змістовне дублювання.

Таблиця 1.2. Найпопулярніші новинні джерела

Web-сайт	Кількість веб-сайтів, що посилаються
ИА «Интерфакс»	1051
«РосБизнесКонсалтинг»	983
«Reuters»	882
ИТАР-ТАСС	787
РИА «Новости»	773
УНИАН	675
Радио «Свобода»	662
НТВ	631
«Коммерсантъ»	623
BBC	598
«Комсомольская правда»	595

У результаті проведених досліджень була сформована модель новинного Web-простору, яку наведено на рис. 1.13. Ця модель включає такі зони:

- вхідний півострів. Web-сайти, яким відповідають менше граничного значення вхідних посилань і будь-яка перевищуюча гранична

кількість вихідних посилань (таких Web-сайтів виявилося 312 або 16,7%);

- вихідний півострів. Web-сайти, яким відповідають менше граничного значення вихідних посилань і будь-яка перевищуюча гранична кількість вхідних посилань (таких Web-сайтів виявилося 513 або 27,5%);
- острів. Web-сайти, яким відповідають менше граничного значення вихідних і вхідних посилань (таких Web-сайтів виявилося 358 або 19,3%);
- ядро, що складається з трьох областей: вхідної, вихідної та комунікаційної зони (таких Web-сайтів виявилося 680 або 36,5%). Зона ядра характеризується середніми та більшими значеннями рівнів вихідних і вхідних зв'язків, однак, як бачимо, допускає ранжирування за рівнем цих комунікацій.

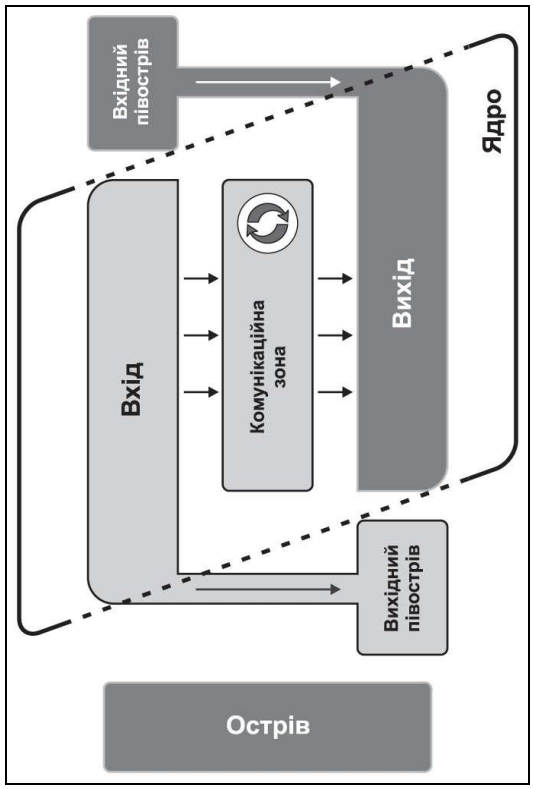


Рис. 1.13. Архітектура новинного Web

Основа моделі була побудована шляхом аналізу повної картини розподілу вхідних і вихідних посилань. При цьому враховувалися підходи до

побудови матриці інціденцій і відповідних графів зв'язку [23]. Разом з тим виявилося, що саме відношення кількості вхідних і вихідних посилань для кожного із джерел досить точно характеризує його влучення в названі кластери. Це значно спростило підрахунки та прискорило час побудови моделі.

Наприклад, для поділу області ядра на вхідну, вихідну та комунікаційну зони можна розглянути ранжируваний графік логарифма відносини кількості вихідних і вихідних посилань для кожного із джерел із цієї області (рис. 1.14).

Центральна зона цього графіка відповідає комунікаційній, ліва - вихідний, а права - вхідній зоні.

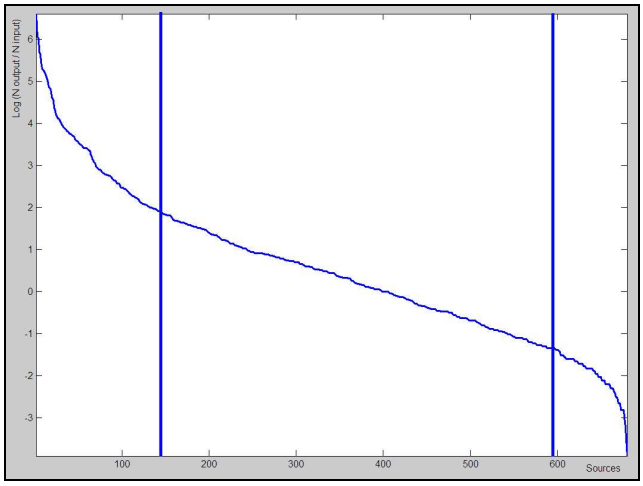


Рис. 1.14. Графік відносної кількості посилань від номера джерела

Цікавим виявився графік двовимірного перетину значень $log(N_{out} + 1)$, $log(N_{in} + 1)$, де N_{out} - кількість вхідних посилань, N_{in} - кількість вихідних посилань для кожного із джерел (рис. 1.15). Цей графік послужив основою ідеальної схеми подання областей моделі залежно від кількості вихідних і вхідних посилань (рис. 1.16).

У результаті проведених досліджень було побудовано модель новинного Web-простору (рис. 1.15, рис.1.16), що базується на контекстних посиланнях. Також запропоновані підходи до виявлення основних зон моделі новинного Web-простору та розраховані числові співвідношення різних зон моделі.

Разом з тим, дана модель припускає подальше вдосконалювання в наступних напрямках: більш точної ідентифікації контекстних посилань, удосконалювання критерію визначення зон на основі повного урахування структури посилань та методів кластерного аналізу, удосконалювання механізму визначення змістовного дублювання інформації (у тому числі за рахунок механізмів настроювання сканерів системи контент-моніторингу, врахування авторитетності джерел і можливих навмисних затримок публікації в Інтернет).

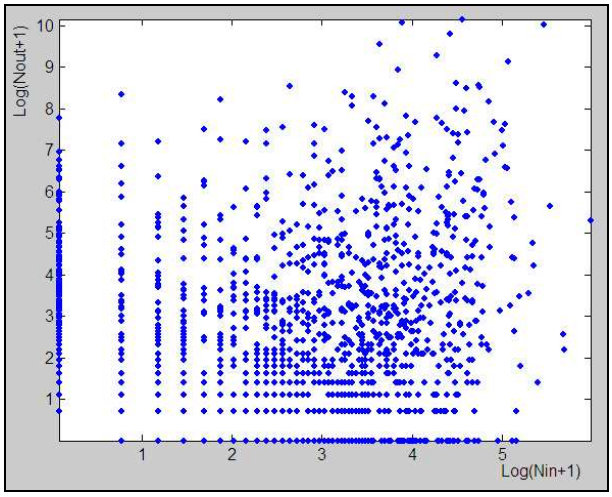


Рис. 1.15. Графік розподілу зони ядра у координатах «логарифм кількості вихідних повідомлень - логарифм кількості вхідних повідомлень»

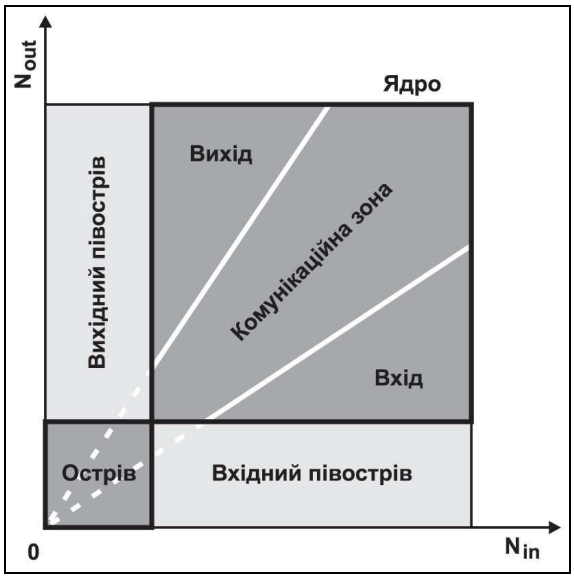


Рис. 1.16. Представлення областей моделі у залежності від кількості вихідних і вхідних посилань

1.5. Ентропія і кількість інформації

Однією з основ дослідження інформаційних потоків в мережах є класична теорія інформації, що оформилася в 40-х роках XX століття завдяки роботам К. Шенона (C.E. Shannon) [37].

Поняття ентропії спочатку виникло у фізиці, наприклад, у статистичній фізиці вводяться поняття макроскопічного та мікроскопічного станів. Останній визначається так званими мікроскопічними параметрами, наприклад, значеннями в даний момент часу всіх імпульсів і координат всіх частинок, з яких складається система. За асоціацією до цього, мікроскопічним станом інформаційного потоку на деякий час можуть бути окремі терми, їх взаємні змістовні зв'язки, взаємне розміщення.

Природно, одному макростану може відповідати декілька мікростанів. Чим більше мікростанів відповідають даному макростану, тим більша ймовірність цього макростану.

Як класичний приклад розглядається уявний експеримент - закритий ящик, розділений на дві рівні частини. За всім обсягом ящика рівномірно розподілені частинки, кожна з яких рівноімовірно може знаходитися як в лівій, так і в правій частині. Припускається, що кількість частинок дорівнює 100. Перший з даних макростанів такий: всі частинки розташовані в лівій частині, цьому макростану відповідає тільки одне мікростан ($N = 1$). Другий макростан такий – в лівій частині знаходиться тільки одна частинка - такому макростану відповідає вже сто мікростанів ($N = 100$). Як третій макростан вибирається такий, коли кількість частинок в лівій частині складає половину від всієї кількості частинок. Такому макростану, природно, повинно відповідати найбільше число мікростанів. Дійсно, як показує просте обчислення, кількість поєднань з 100 елементів по 50 складає $N \approx 10^{29}$. При такий гігантський відмінності кількості мікростанів для різних макростанів зрозуміло, що ймовірністю зустріти перший макростан, в порівнянні з ймовірністю зустріти третій ($\sim 10^{-29}$) можна нехтувати.

Для того, щоб не оперувати з великими числами розглядають логарифм від кількості мікростанів, відповідних даному макростану, який і називають ентропією:

$$S = k \ln N ,$$

де k - деяка константа (наприклад, у фізиці постійну Больцмана).

У разі, коли всі мікростани рівноімовірні $p_i = p = 1/N = const$ вираз для ентропії може бути записаний як:

$$S = k \ln N = k \sum_{i=1}^{N} p \ln N = -k \sum_{i=1}^{N} p \ln p .$$

Взагалі кажучи, ймовірність мікростанів може бути різною, тому вираз для ентропії у загальному випадку записується таким чином:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i .$$

У теорії інформації константу k прийнято вибирати рівною $k = 1/\ln 2$ (інформація вимірюється в бітах!), тобто:

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i ,$$

це і є якраз ентропія, яку було запропоновано К. Шеноном. Заснована ним класична теорія інформації була орієнтована перш за все на дослідження процесів передачі даних по каналах зв'язку. Завдяки використанню таких понять, як інформаційна ентропія, кількість інформації, взаємна інформація тощо, теорія інформації набула універсального характеру, і її методи почали широко використовуватися в багатьох областях науки і технологій. Багато ефективних методів вирішення завдань глибинного аналізу текстів базуються на понятті взаємної інформації (mutual information), яка широко застосовується, зокрема, в області статистичної обробки природних мов [23], дозволяючи визначати близькість між словами або якими-небудь іншими мовними явищами.

К. Шеннон розглядав ентропію як міру невизначеності ансамблю $U = \{u_1, \ ..., \ u_N\},$ яка визначається таким виразом:

$$H(U) = -K\sum_{i=1}^{N} p_i \log_2 p_i,$$

де p_i - ймовірність стану u_i, K - позитивна константа.

У випадку, якщо всі стани джерела інформації рівноімовірні, формула для ентропії приймає вигляд:

$$H(U) = -\sum_{i=1}^{N} p_i \log_2 p_i = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = -\log_2 \frac{1}{N} = \log_2 N,$$

яка співпадає з мірою так званою Хартлі, таким чином підтверджуючи той факт, що вона є окремим випадком ентропії Шенона.

Для пояснення поняття інформаційної ентропії можна розглянути процес отримання повідомлення довжиною N символів (букв або знаку пропуску).

Отже, нехай передається повідомлення, що складається з n різних символів, - u_1, u_2, ..., u_n. Дане повідомлення можна представити у вигляді таблиці:

u_{10}	u_5	u_{21}	.	u_3
1	2	3	.	N

де перший рядок - це символи повідомлення, а другий - відповідні цим символам номера місць в повідомленні.

Нехай для будь-якого значення i символ u_i ($i = 1, 2, ..., n$) генерується з ймовірністю p_i, причому це значення не залежить від попередніх символів. Тоді при досить великому N кількість символів u_i буде з великою точністю відповідати значенню Np_i. Таким чином ймовірність p отримати повідомлення, в якому міститься Np_1 символів u_1, Np_2 символів u_2 тощо (без урахування їх місцезнаходження у повідомленні), дорівнює:

$$p = p_1^{Np_1} p_2^{Np_2} ... p_n^{Np_n}.$$

Двійковий логарифм від цієї ймовірності можна записати таким чином:

$$\log_2 p = N\sum_{i=1}^{n} p_i \log_2 p_i .$$

Другий співмножник цього виразу із зворотним знаком – це ентропія Шенона:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i ,$$

Таким чином ймовірність появи повідомлення довжиною N символів з вказаними вище властивостями, дорівнює:

$$p = 2^{-NS} .$$

Оскільки всі подібні повідомлення рівноімовірні (з ймовірністю p), то їх кількість K дорівнює:

$$K = \frac{1}{p} = 2^{NS} .$$

Таким чином, інформаційна ентропія (або ентропія Шенона) визначає кількість повідомлень, в яких символи зустрічаються з «вірною» із статистичних міркуваннях частотою (u_1 з p_1, u_2 з p_2, і так далі).

Слід відмітити, що введена Шеноном ентропія - це та ж ентропія з фізики, хоча і використовується вона для інших цілей. З фізичної термінології, макростан задається набором $\{p_1,\ p_2,\ ...,\ p_n\}$. Кожному макростану відповідає $K\left(\{p_1, p_2, ...p_n\}\right) = 2^{NS(\{p_1, p_2, ... p_n\})}$ мікростанів. Для пояснення, приведемо два приклади.

Макростану $\{1,\ 0,\ ...,\ 0\}$ (з ймовірністю 1 зустрічається символ u_1) відповідає тільки одне повідомлення, тобто тільки один мікростан. Ентропія такого макростану (з урахуванням того, що $\lim_{x\to 0} x\log x = 0$) дорівнює:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i = -1\cdot\log_2 1 - 0\cdot\log_2 0 - ... = 0 .$$

Такий результат цілком відповідає інтуїції, невизначеності немає, повідомлення, яке ми можемо отримати, повністю визначене (передбачено) – ентропія мінімальна.

А ось , наприклад, макростану, в якому кожен символ зустрічається з однією і тією ж ймовірністю $p_i = 1/n$, - $\{1/n,\ 1/n,\ ...,\ 1/n\}$ відповідає набагато більша ентропія:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i = \log_2 n.$$

Кількість різних повідомлень довжиною N, в яких кожен символ зустрічається N/n раз, природно, набагато більша одиниці:

$$K\left(\{1/n, 1/n, ...1/n\}\right) = 2^{NS\left(\{1/n, 1/n, ...1/n\}\right)} = 2^{N \log_2 n} = n^N.$$

Ентропія, введена Шеноном як міра невизначеності стану дискретного джерела інформації, має наступні властивості:

1) Ентропія є дійсною раціональною ненегативною величиною в інтервалі [0, 1].
2) Ентропія - величина обмежена.
3) Ентропія дорівнює нулю лише тоді, коли ймовірність одного із станів дорівнює одиниці, тобто стан джерела точно визначений.

Розглянемо графік залежності ентропії джерела з двома станами, що характеризуються ймовірностями p і $1-p$, відповідно (рис. 1.17).

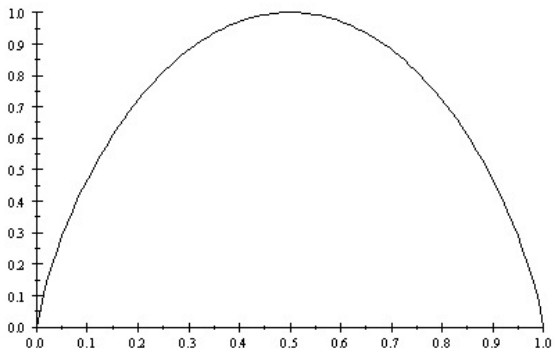


Рис. 1.17. Ентропія системи з двома станами

(вісь абсцис - p, вісь ординат – ентропія H)

Ентропія в цьому випадку дорівнює:

$$H(U) = -\left[p \log p + (1-p) \log(1-p)\right].$$

Для побудови графіка ентропії системи з трьома станами (рис. 1.18), що характеризуються ймовірностями p, q, $1-p-q$, застосовується формула:

$$H(U) = -\left[p\log p + q\log q + (1-p-q)\log(1-p-q) \right].$$

На практиці дуже важливою є і така властивість ентропії, як повне ігнорування змістовної сторони станів джерела, а лише урахування ймовірності цих станів. При цьому розглядається лише ступінь невизначеності. Так, наприклад, якщо розглянути повну множину результатів лікування хворого, що складається з двох подій, – сприятливого і несприятливого, то результат сприятливого результату вилікування з вірогідністю 0.9 і з вірогідністю 0.1 мають однакове значення ентропії.

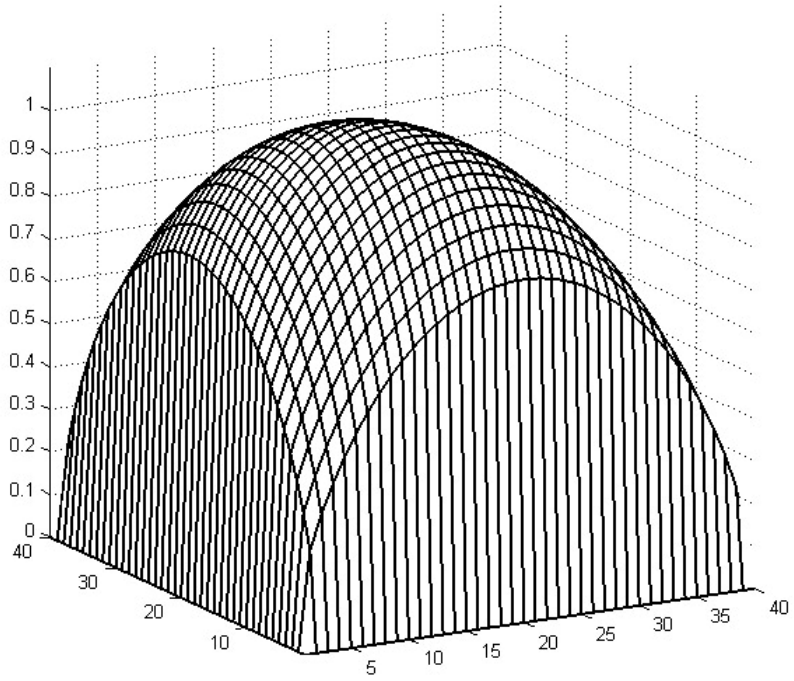


Рис. 1.18. Ентропія системи з трьома станами
(вісь OX - p, вісь OY – q, вісь OZ - ентропія) p,

Розглянемо поняття умовної ентропії. Для цього визначимо ентропію об'єднання двох статистично зв'язаних ансамблів U і V, яке характеризується матрицею вірогідності $p(U,V) = \left\| p(u_i, v_j) \right\|$, $(i = 1, \ldots, N; \quad j = 1, \ldots, M)$.

З теорії ймовірностей відомо: $p(u_i, v_j) = p(u_i)p(v_j / u_i) = p(v_j)p(u_i / v_j)$, відповідно, ентропія об'єднання подій виражається формулою:

$$H(U,V) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(u_i v_j)\log p(u_i v_j) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(u_i)p(v_j / u_i)\log[p(u_i)p(v_j / u_i)] =$$

$$= -\sum_{i=1}^{N} p(u_i)\log p(u_i) - \sum_{i=1}^{N} p(u_i)\cdot\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Назвемо $H_{u_i}(V)$ частковою умовною ентропією ансамблю V по відношенню до стану $u_i \in U$:

$$H_{u_i}(V) = -\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Відповідно, ступінь невизначеності, що доводиться на один стан ансамблю V при відомих станах ансамблю U, або умовна ентропія V по відношенню до U визначається як:

$$H_U(V) = \sum_{j=1}^{N} p(u_i)H_{u_i}(V) = -\sum_{i=1}^{N} p(u_i)\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Умовна ентропія має такі основні властивостями:

1) Ентропія об'єднання двох ансамблів V і U дорівнює сумі безумовної ентропії одного ансамблю та умовної ентропія іншого щодо першого:

   $H(UV) = H(U) + H_U(V)$,

   $H(UV) = H(V) + H_V(U)$.

2) Наявність відомостей щодо результатів реалізації стану одного ансамблю ніяк не може збільшити невизначеність вибору стану з іншого ансамблю:

   $H_U(V) \le H(V)$,

   $H_V(U) \le H(U)$.

3) У разі відсутності статистичного зв'язку в реалізаціях станів з ансамблів U і V:

$$H_U(V) = H(V),$$

$$H_V(U) = H(U).$$

Для визначення поняття кількості інформації будемо розглядати дискретне джерело Z як множину можливих повідомлень $Z = \{z_1, \ldots, z_N\}$. Допустимо, повідомлення передаються по каналу зв'язку і приймаються як деяка нова, множина (можливо, спотворена перешкодами) $W = \{w_1, \ldots, w_N\}$.

Середня невизначеність щодо будь-якого стану джерела, що залишається у адресата після отримання повідомлення w_j характеризується умовною ентропією:

$$H_{w_j}(Z) = -\sum_{i=1}^{N} p(z_i / w_j) \log p(z_i / w_j).$$

Тоді середня невизначеність по всьому ансамблю повідомлень, що приймаються, дорівнює сумі по всіх j:

$$H_W(Z) = -\sum_{i=1}^{N} p(w_j) H_{w_j}(Z).$$

Визначимо кількість інформації, що міститься в кожному прийнятому елементі повідомлення щодо будь-якого переданого повідомлення, таким чином:

$$I(ZW) = H(Z) - H_W(Z).$$

Очевидно, для кількості інформації справедливе співвідношення:

$$I(ZW) = \sum_{i=1}^{N} \sum_{j=1}^{N} p(z_i w_j) \log \frac{p(z_i w_j)}{p(z_i)\, p(w_j)}.$$

Розглянемо деякі властивості так визначеної кількості інформації:

1) Кількість інформації величина ненегативна. Дійсно

$$H(Z) \geq H_W(Z) \quad \Rightarrow \quad I(ZW) = H(Z) - H_W(Z) \geq 0.$$

2) За відсутності статистичного зв'язку між Z і W:

$$H(Z) = H_W(Z) \quad \Rightarrow \quad I(ZW) = 0.$$

3) $I(ZW) = I(WZ)$. Дійсно:

$$I(ZW) = H(Z) - H_W(Z) = H(ZW),$$

$$I(WZ) = H(W) - H_Z(W) = H(WZ).$$

При цьому $H(ZW) = H(WZ)$.

4) При взаємно однозначній відповідності між Z і W:

$I(ZW) = H(Z)$.

Тобто це максимальна кількість інформації щодо стану дискретного джерела.

Взаємна інформація визначається аналогічно поняттю кількості інформації, що міститься в кожному прийнятому елементі повідомлення, щодо будь-якого переданого повідомлення. Разом з тим, приведені вище формули для кількості інформації характеризують інформаційні властивості одного дискретного джерела або ансамблю. Проте при розгляді інформаційних потоків особливий інтерес представляє виявлення кількості інформації в ансамблі категорій V, кількість яких рівна N, щодо іншого – словника тексту U, що містить M ключових слів (термів).

Для визначення даної інформаційної характеристики розглянемо умовну ентропію $H_U(V)$, яка визначає середню кількість інформації, що видається повідомленнями ансамблю V за умови, що повідомлення ансамблю U вже відоме. Ця умовна ентропія задається формулою:

$$H_U(V) = \sum_{j=1}^{M} P(u_j) H_{u_j}(V).$$

Підставляючи в цю формулу значення часткової умовної ентропії $H_{u_i}(V)$, отримуємо:

$$H_U(V) = -\sum_{j=1}^{M}\sum_{k=1}^{N} P(v_k, u_j) \cdot \log(P(v_k, u_j) / P(u_j)).$$

Взаємна інформація між V і U визначається як:

$I(V,U) = H(V) - H_U(V)$.

Взаємна інформація вимірюється в тих же одиницях, що і ентропія (наприклад, в бітах). Величина $I(V,U)$ показує, скільки в середньому біт

інформації щодо реалізації ансамблю V дає спостереження щодо реалізації ансамблю U.

Виразимо взаємну інформацію через вірогідність:

$$I(V,U) = H(V) - H_U(V) =$$

$$= -\sum_{k=1}^{N} P(v_k) \cdot \log P(v_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log(P(v_k,u_j)/P(u_j)) =$$

$$= -\sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log P(v_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log(P(v_k,u_j)/P(u_j)) =$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log \frac{P(v_k,u_j)}{P(v_k) \cdot P(u_j)}.$$

Взаємна інформація має такі властивості:

1) $I(V,U) \geq 0$, причому рівність має місце тільки у тому випадку, коли V і U взаємно незалежні.
2) $I(V,U) = I(U,V)$, тобто U містить стільки ж інформації відносно V, скільки V містить відносно U. Тому можна так само записати $I(V,U) = H(U) - H_V(U)$.
3) $I(V,U) \leq H(V)$, $I(V,U) \leq H(U)$, причому рівність має місце, коли по реалізації U можна точно відновити реалізацію V або навпаки.
4) $I(V,V) = H(V)$, що дозволяє інтерпретувати ентропію джерела як інформацію ансамблю V саму про себе.

У випадку інформаційних потоків взаємна інформація описує кількість інформації про приналежність окремого документа (або цілого інформаційного потоку) до певної категорії c, яка наприклад, пов'язана з наявністю деякого терму t.

$$I(t,c) = P(t,c) \log \frac{P(t,c)}{P(t)P(c)},$$

де $P(t,c)$ - емпірично оцінена вірогідність тієї події, що одночасно зустрічається терм t і документ (потік) належить до категорії c; $P(t)$ -

ймовірність появи терму t, $P(c)$ - вірогідність приналежності документа (потоку) до категорії c.

Таким чином, взаємна інформація між термом і категорією описує ступінь асоціації терму t і категорії c.

На інформаційній теорії базуються досить багато інформаційно-пошукових і аналітичних систем. Так, зокрема, компанія Autonomy створила аналітичний сервер IDOL (Intelligent Data Operating Layer), ідеологія якого базується на використанні байєсовських ймовірностей і теорії Шенона, яка розглядається як математична основа побудови комунікаційних систем, що дозволяє визначати і інтерпретувати чисельні значення кількості інформації. На думку розробників сервера IDOL, природні мови мають високу ступень надмірності, неістотного змісту. За допомогою аналізу ентропії, а точніше, використовуючи методологію взаємної інформації, сервер IDOL забезпечує витягання «суті» з «надмірних» текстів. На думку ідеологів системи IDOL, чим рідше контекст зустрічається в процесі комунікації, тим він важливіший, тим більше інформації він несе. Завдяки такому підходу забезпечується знаходження найбільш інформативних понять в документах.

2. МОДЕЛІ ІНФОРМАЦІЙНИХ ПОТОКІВ

Аналіз динаміки інформаційних потоків, що генеруються у веб-просторі стає сьогодні одним з найбільш інформативних методів дослідження актуальності тих або інших тематичних напрямків [3]. Ця динаміка обумовлена факторами, багато з яких не піддаються точному аналізу. Однак загальний характер часової залежності кількості тематичних публікацій в Інтернеті все ж таки допускає побудову математичних моделей.

У поведінці інформаційних потоків спостерігаються дві характерні риси: по-перше, виразна тенденція до постійного зростання їхніх обсягів, а по-друге, ускладнення динамічної структури. Спостереження часових залежностей числа повідомлень в мережних інформаціних потоках переконливо свідчать про те, що механізми їхньої генерації та поширення, очевидно, зв'язані зі складними нелінійними процесами загальної мережної динаміки.

У літературі традиційними вважаються два класи моделей інформаційних потоків: лінійні й експонентні. Серед останніх виділяється модель Бартона-Кеблера, запропонована у свій час для опису процесу старіння інформаційних ресурсів:

$$m(t) = I - ae^{-t}(-t) - be^{-2t},$$

де $m(t)$ – частка корисної інформації в загальному потоці I; перша експонента відповідає статичним ресурсам, а друга – динамічним (новинним).

Обидва класи мають істотну обмеженість - монотонний характер часової залежності. Тобто вони мало придатні для вивчення реальної динаміки мережних інформаційних потоків.

2.1. Тематичні інформаційні потоки

Під тематичним інформаційним потоком будемо розуміти послідовність повідомлень, що відповідають певному тематичному запиту.

Отже, під тематичним інформаційним потоком будемо розуміти кількість документів, що у деякому змісті відносяться до заданої теми. Розглянемо загальну картину динаміки тематичних інформаційних потоків, обмеживши механізмами, типовими для динамічного сегмента Інтернет.

Численні факти свідчать про те, що в дійсності динаміка тематичних інформаційних потоків визначається комплексом внутрішніх нелінійних механізмів, які лише частково корелюють з об'єктивним оточенням. Очевидно, що ця динаміка в принципі не може бути пояснена деяким одним фактором, який повністю відповідає за всю розмаїтість ефектів, що спостерігаються. Саме ця обставина й надає особливу актуальність проблемі моделювання динаміки мережних тематичних потоків.

Ми виходимо з того, що загальний інформаційний потік, який вимірюється в кількості повідомлень, є величиною відносно стабільною. Змінюються в часі лише об'єми повідомлень, які відповідають тій або іншій тематиці. Іншими словами, зростання кількості публікацій по одній темі супроводжується зменшенням публікацій на інші теми [25], так що для кожного проміжку часу T маємо:

$$\int_0^T \sum_{i=1}^{M} y_i(t)dt = NT ,$$

де $y_i(t)$ – кількість публікацій в одиницю часу, а M – загальна кількість всіх можливих тим. Звичайно, передбачається, що частина $n_i(t)$ завжди дорівнює нулю. Тобто для локальних часових проміжків можна спостерігати так званий «тематичний баланс».

Основний інтерес в такому формулюванні представляє вивчення динаміки окремого тематичного потоку, який описується щільністю $n_i(t)$.

Теоретично можна припустити, що множини публікацій, асоційованих з певним набором тематик, перетинаються, тобто існують публікації, які можуть бути віднесені одночасно до декількох різних тем. Загалом кажучи, така політематичність дійсно спостерігається, вона є ефектом, який

необхідно враховувати, але в першому наближенні будемо вважати, що його внесок не спотворює загальну картину.

Подібне розуміння мережевих тематичних інформаційних потоків, мабуть, дозволяє більш менш адекватно описувати загальні закономірності їхній динаміці.

У практичному плані часто виявляється цілком задовільним спрощене розуміння інформаційного потоку як деякої залежної від часу величини *X(t)*, яка описується рівнянням:

$$\frac{dX(t)}{dt} = F\left(X(t), t\right).$$

Далі, кожна тематика також має ряд характерних властивостей, які допускають деяку класифікацію, наприклад, на основі особливостей її утворення та відтворення в часі:

- публікації на «разову» тему, часова залежність числа яких різко зростає, виходить на насичення, а потім убуває та асимптотично спрямовується до нуля;
- публікації за темами, що періодично з'являються у загальному інформаційному потоці, які після закінчення обмеженого проміжку часу практично зникають з нього;
- публікації за темою, часова залежність кількості яких коливається біля деякого значення та ніколи не зникає повністю.

Відповідно до цього повідомлення можуть підрозділятися на аналогічні категорії, причому кожна з них має власну специфіку розвитку в часі.

Ще складніше виглядає синхронна зміна кількості повідомлень з декількох тематичних інформаційних потоків. Їхня поведінка чітко нагадує процеси взаємодії популяцій у біоценозах. Так, наприклад, у ряді випадків збільшення числа публікацій за однією темою супроводжується скороченням числа публікацій за іншою. Загальна динаміка у цьому випадку може описуватися системою рівнянь, кожне з яких відноситься до окремого монотематичного потоку. Підкреслимо, що загальні політематичні потоки є

стаціонарними по кількості публікацій, динаміка ж в основному визначається «конкурентною боротьбою» окремих тематик.

У літературі описано багато різновидів систем «конкурентної боротьби» для різних модифікацій моделі в залежності від цілого ряду припущень щодо реальних умов протікання процесів. У найпростішому вигляді такі рівняння можуть мати такий вигляд:

$$\frac{dm_i(t)}{dt} = p_i \cdot m_i(t) - \sum_{j=1}^{N_m} r_{ij} \cdot m_i(t) \cdot m_j(t),$$

де N_m – кількість тематик.

Приведена система рівнянь описує перерозподіл публікацій між тематиками, які утворюють фіксований набір. Але в реальному житті тематики (сюжети) з'являються і з часом зникають, тому необхідно ввести в ці рівняння відповідні корективи. Це можна зробити по-різному, наприклад, визначивши коефіцієнти p_i і r_{ij} залежними від часу так, щоб кожен сюжет мав власний максимум активності на певному проміжку часу.

2.2. Традиційні моделі інформаційних потоків

Лінійна модель

У деяких випадках динаміка тематичних інформаційних потоків, що може бути виражена кількістю публікацій за певний період, її інтенсивністю, обумовленою, наприклад, зміною активності тематики (її підвищенням або старінням), відбувається лінійно, тобто кількість повідомлень у момент часу *t* можна, відповідно, представити формулою:

$$y(t) = y(t_0) + v(t - t_0),$$

де t_0 - деякий стартовий час відліку, $y(t)$ – кількість повідомлень на час t, v – середня швидкість збільшення (зменшення) інтенсивності тематичного інформаційного потоку.

Важливі характеристики інформаційного потоку можуть бути кількісно оцінені флуктуацією цього потоку – зміною середньоквадратичного відхилення $\sigma(t)$, що обчислюється за формулою:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left[y(t_i) - \left(y(t_0) + v(t_i - t_0)\right)\right]^2}.$$

Якщо ця величина змінюється пропорційно квадратному кореню від часу, то процес зміни кількості публікацій по обраній темі можна вважати процесом з незалежними приростами. При цьому зв'язками з попередніми тематичними публікаціями можна зневажити.

У випадку, коли середньоквадратичне відхилення пропроційно деякому ступеню часу: $\sigma(t) \propto t^{\mu}$ $(1/2 \le \mu \le 1)$, чим більше значення μ, тим вище кореляція між поточними та попередніми повідомленнями в інформаційному потоці.

Експоненціальна модель

У деяких випадках процес зміни актуальності тематики (збільшення або зменшення кількості тематичних повідомлень в інформаційному потоці в одиницю часу) апроксимується експонентною залежністю, яку можна виразити формулою:

$$y(t) = y(t_0)e^{\lambda\cdot(t-t_0)},$$

де λ - середня відносна зміна інтенсивності тематичного інформаційного потоку.

У реальності актуальність тематики є дискретною величиною, вимірюваною в моменти часу t_0, ..., t_n, яка лише апроксимується наведеною вище залежністю. У рамках даної моделі справедливо:

$$y(t_i) = y(t_0)e^{\lambda\cdot(t_i-t_0)} = y(t_0)\,\mathrm{e}^{\lambda\cdot(t_i-t_{i-1}+t_{i-1}-t_0)} = y(t_{i-1})\,\mathrm{e}^{\lambda\cdot(t_i-t_{i-1})}.$$

Звідки:

$$\frac{y(t_i)}{y(t_{i-1})} = \mathrm{e}^{\lambda\cdot(t_i - t_{i-1})}.$$

Введемо позначення: $\lambda(t_i)$ - відносна зміна інтенсивності тематичного інформаційного потоку в момент часу t_i:

$$\lambda(t_i) = \lambda \cdot (t_i - t_{i-1})$$

і прологарифмуємо наведене вище рівняння:

$$\lambda(t_i) = \ln\frac{y(t_i)}{y(t_{i-1})}.$$

Відносна зміна інтенсивності в момент часу t_i на практиці також часто обчислюється як співвідношення:

$$\lambda(t_i) = \ln\frac{y(t_i)}{y(t_{i-1})} \approx \frac{y(t_i) - y(t_{i-1})}{y(t_{i-1})}.$$

Зміна флуктуацій величини $\lambda(t_i)$ щодо середнього значення може оцінюватися за стандартним відхиленням:

$$\sigma(t_n) = \sqrt{\tfrac{1}{n}\sum_{i=0}^{n}\left(\lambda(t_i) - \lambda\right)^2}.$$

У цьому випадку також, якщо $\sigma(t)$ змінюється пропорційно кореню квадратному від часу, то можна говорити про процес із незалежними збільшеннями - кореляції між окремими повідомленнями несуттєві. У випадку наявності значної залежності повідомлень спостерігається співвідношення: $\sigma(t) \propto t^{\mu}$, причому μ перевищує ½, але обмежено 1.

Значення μ, що перевищує ½, свідчить щодо наявності довгострокової пам'яті в інформаційному потоці. Такий клас процесів одержав назву автомодельних, для яких передбачається кореляція між кількістю повідомлень, що публікуються у різні моменти часу.

Логістична модель

На відмінність від моделі Бартона-Кеблера у реальній динаміці інформаційних потоків мають місце як процеси росту, так і спаду кількості

документів. Тому для побудови реалістичної картини, безумовно, потрібно застосовувати більш гнучку модель.

Насамперед, варто сказати, що документи в інформаційному потоці в багатьох відносинах нагадують популяції живих організмів. Вони в певному сенсі «народжуються», «вмирають» і дають «потомство» (документи, що містять інформацію, що раніше з'явилася в декількох інших документах). У сучасній науковій літературі поняття популяції часто використовується в широкому тлумаченні, і тому цілком обґрунтовано введення його і при моделюванні інформаційних потоків.

У другій половині XX-го століття були досягнуті значні успіхи в побудові різних математичних моделей динаміки популяцій [6, 8]. Найбільш перспективною, на наш погляд, варто вважати логістичну модель, запропоновану у свій час П. Ферхюльстом для опису динаміки народонаселення [62] і згодом Р. Пером для біологічних співтовариств [55]. Надалі вона виявилася вкрай плідною в багатьох галузях науки та техніки.

Логістичну модель [1, 17, 25] можна розглядати як узагальнення експонентної моделі Мальтуса, яка передбачає пропорційність швидкості росту функції $y(t)$ в кожен момент часу її значенню:

$$\frac{dy(t)}{dt} = ky(t),$$

де $k-$ деякий коефіцієнт.

У реальному житті, як правило, динамічні системи мають досить ефективні зворотні зв'язки, які дозволяють коригувати характер процесів, що відбуваються в них, і тим самим утримувати їх у певних рамках. Інформаційні операції, коригуючи ці зворотні зв'язки в певні періоди еволюційного процесу, можуть досить ефективно вплинути на характер поведінки всієї системи.

Найбільш простим узагальненням закону Мальтуса, яке дозволяє вирішити (принаймні, принципово) проблему необмеженого зростання розв’язку, є заміна постійного коефіцієнта k деякою функцією часу $k(t)$.

Природно, ця функція повинна бути обрана таким чином, щоб дотримувалися такі умови:

- розв'язок рівняння мав би прийнятну поведінку;
- структура функції мала б певний зміст з погляду на явище, яке досліджується.

Головна ідея логістичної моделі складається в тому, що для обмеження швидкості росту на функцію $y(t)$ накладається додаткова умова, відповідно до якого її значення не повинне перевищувати деякої величини. Найпростіший спосіб обмежити зростання експонентної залежності розв'язку наведеного вище рівняння полягає у введенні для неї граничного значення. Для цього виберемо $k(t)$ такого вигляду:

$$k(t) = k \cdot [N - ry(t)],$$

де N – граничне значення, яке функція $y(t)$ не може перевищити, r – коефіцієнт, що описує негативні для даної тенденції процеси, k – коефіцієнт пропорційності. Причому передбачається, що завжди $n_0 \le N$. Тоді замість першого рівняння маємо:

$$\begin{cases} \dfrac{dy(t)}{dt} = ky(t)(N - ry(t)), \\ y(t_0) = y_0. \end{cases}$$

Модель, що заснована на наведеному вище рівнянні, називається логістичною. Незважаючи на уявну простоту, подібне узагальнення закону Мальтуса аж ніяк не є примітивним. Навпроти, воно дозволяє явно включити в опис динаміки популяцій винятково важливий зворотний зв'язок, роль якого у оточуючому нас світі важко переоцінити. Логістичне рівняння, власне кажучи, варто вважати феноменологічним: ми не знаємо, як діють конкретні механізми, що знижують по мірі зростання $y(t)$ швидкість її зміни. І це, у даному випадку, серйозна перевага.

Існує два класи розв'язків логістичного рівняння, які, залежно від значень коефіцієнтів k_0 і n_0, описують зростання та убування залежності

$y(t)$. Їхня типова поведінка зображене на рис. 15. Як видно, логістична модель, на відміну від закону Мальтуса, описує досягнення системою деякого рівноважного стану.

Наведене вище логістичне рівняння має два рівноважних розв'язки: $y(t)=0$ і $y(t)=N$. З формальної точки зору перший з них хиткий, тому що при малих значеннях $y(t)$ його відхилення від нуля приводить до зростання. Однак у практичному плані це не зовсім так. Справа в тому, що реальні обсяги інформаційних потоків є дискретними множинами, і якщо в якийсь момент $y(t)$ приймає значення, менше за одиницю, то зрости воно вже не зможе. Тому у випадку опису того, що відбувається в реальності, розв'язок $y(t)=0$ також варто вважати рівноважним.

Друге ж рішення $y(t)=N$ є рівноважним у будь-якому сенсі. Дійсно, при $y(t)>N$ включаються механізми спаду залежності, а при $y(t)<N$ – відповідно зростання.

Розглянемо поведінку динаміки тематичного інформаційного потоку, обсяг якого визначається логістичним рівнянням. Необхідно підкреслити, що висновки, які будуть зроблені нижче, залишаються (з точністю до числових значень констант) справедливими також при будь-яких значеннях коефіцієнтів і навіть для широкого класу моделей з різними функціями обмеження експонентного зростання.

На рис. 2.1 зображена результуюча залежність обсягів інформаційного потоку від часу при різних початкових умовах. У точках A і B швидкість зміни кількості повідомлень спрямовується до нуля: це стаціонарні стани. Між A і B швидкість позитивна (кількість повідомлень зростає), а вище точки B - негативна (кількість повідомлень убуває).

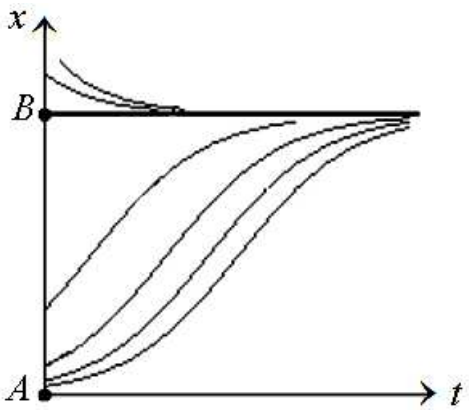


Рис. 2.1. Узагальнена логістична модель

Модель передбачає, що згодом встановлюється стаціонарний режим B, що виглядає цілком природно: більший інформаційний потік зменшується, менший - збільшується.

Логістична модель задовільно описує численні явища насичення. Поблизу A, коли обсяг інформаційного потоку малий, вона дуже близька до мальтузіанської моделі. Але при досить великих x спостерігається різка відмінність від мальтузіанського зростання: замість спрямування x до нескінченності кількості публікацій наближається до стаціонарного значення B.

Розглянемо, як логістична модель може застосовуватися під час аналізу інформаційних потоків, а саме визначення мінімальної початкової кількості c повідомлень (яку можна, наприклад, виділити для початку деякої рекламної кампанії). Нехай x - обсяг тематичного інформаційного потоку. На динаміку цієї величини здійснюється вплив інших тематик, які зменшують його розповсюдження, що описується таким чином:

$$\dot{x} = x - x^2 - c.$$

Обчислення показують, що поведінка системи різко змінюється при деякому критичному значенні c (рис. 2.2).

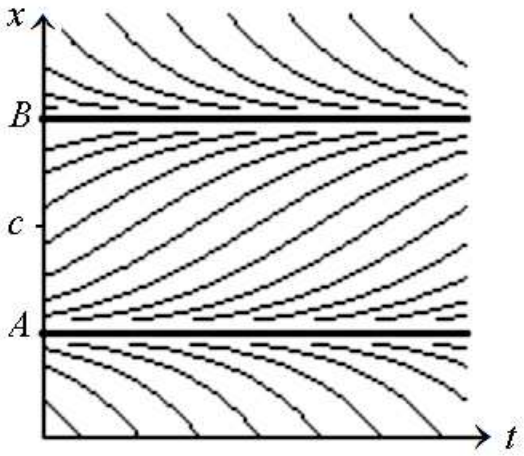


Рис. 2.2. Дві постійні точки логістичної моделі

Якщо величина c мала, то зміни (у порівнянні з відсутністю конкуренції, коли $c = 0$) полягають у наступному. Система має два рівноважних стани, A і B. Стан B стійкий: обсяг інформаційного потоку в цьому випадку трохи менша, ніж при безконкурентній ситуації. Цей обсяг відновлюється при малих відхиленнях від рівноважного значення B.

Стан A нестійкий: якщо внаслідок будь-яких причин обсяг інформаційного потоку впаде нижче рівня A, то надалі кількість тематичних повідомлень буде зведена нанівець за цілком скінченний час.

Очевидно, що при наявності сприятливих зовнішніх умов (при деякій щільності ресурсу) обсяг інформаційного потоку зростає вільно, що сприяє логістичному росту. У цьому випадку навіть більш складні моделі повинні давати результати, подібні наведеним. З другого боку це означає, що основні параметри для конкретизації загальної моделі, можуть визначатися в результаті аналізу спрощеної логістичної моделі.

Отже, логістична модель успішно описує досягнення тематичним інформаційним потоком деякого рівноважного стану.

Інформаційну динаміку в загальному випадку можемо представити як процес, обумовлений виникненням і зникненням окремих тематик, що відбуваються на тлі загальних тенденцій інформаційного простору.

Зафіксуємо певну тематику й припустимо, що в момент часу $t = 0$ існує n_0 фонових публікацій. Внаслідок того, що (у рамках прийнятої моделі)

актуальність тематики зберігається протягом проміжку часу λ, можна розглядати окремо дві часові області: $0 < t \le \lambda$ с $D > 0$ и $t > \lambda$ с $D = 0$ (у рамка даної моделі $D = const$ для кожної області - рівень актуальності теми) і, відповідно, функції $u(t)$ і $v(t)$, які є рішеннями для цих областей й "зшиваються" у точці λ:

$$y(t) = \begin{cases} u(t), \ 0 < t < \lambda, \\ v(t), \ t > \lambda, \\ u(t) = v(t), \ t = \lambda. \end{cases}$$

Першої області відповідає процес зростання кількості публікацій в умовах ненульової актуальності теми ($D > 0$) і, можливо, перехід до стану насичення.

Реакція медійних засобів ніколи не буває миттєвою: завжди існує певна затримка в часі. Цей аспект ураховується в моделі шляхом введення фактору запізнювання τ.

Відповідна динаміка описується рівнянням, що після перевизначення коефіцієнтів й їхнього нормування до *N*, для функції $u(t)$ можна представити у вигляді:

$$\frac{du(t-\tau)}{dt} = pu(t-\tau)(1-qu(t-\tau)) + Du(t-\tau),$$

$$u(0) = n_0.$$

Підкреслимо, що змістовно величина p визначає нормовану ймовірність появи публікації в одиницю часу незалежно від актуальності теми. Цей фактор відображає фонові механізми генерації інформації (типовим прикладом може бути механічний передрук матеріалів із престижних інформаційних джерел). Величина ж D характеризує безпосередній вплив актуальності даної теми. Параметр q характеризує зменшення швидкості росту кількості публікацій й є величиною, зворотною до асимптотичного значення залежності $u(t)$ при $D = 0$.

Для другої області, описуваною функцією $v(t)$, відповідно, маємо:

$$\frac{dv(t-\lambda)}{dt} = pv(t-\lambda)(1-qv(t-\lambda)).$$

При цьому повинне враховуватися умова рівності функцій $u(t)$ й $v(t)$ у момент $t = \lambda$:

$$v(\lambda) = u(\lambda).$$

Наведені вище нелінійні диференціальні рівняння є варіантами запису рівняння Бернуллі:

$$y' = ay^2 + by,$$

яке лінеарізується стандартною заміною $z = 1/y$:

$$z' + bz + a = 0.$$

Загальне рішення цього рівняння має вигляд:

$$z = \frac{1}{\mu(x)}[C - a\int \mu(x)dx]$$

с інтегруючим множником:

$$\mu(x) = e^{bx}.$$

Змінні C визначаються: для першої області з початкових умов, а для другий – з умови «зшивання». Шляхом нескладних перетворень знаходимо рішення для першої області:

$$u(t) = \frac{u_s}{1 + (\frac{u_s}{n_0} - 1)\exp[-(p+D)(t-\tau)]},$$

де u_s – асимптотичне значення u, величина якого визначає область насичення:

$$u_s = \frac{p+D}{pq}$$

Таким чином, ми бачимо, що модель правильно описує залежність, що має *S-подібну* (логістичну) форму, представлену на рис. 2.3.

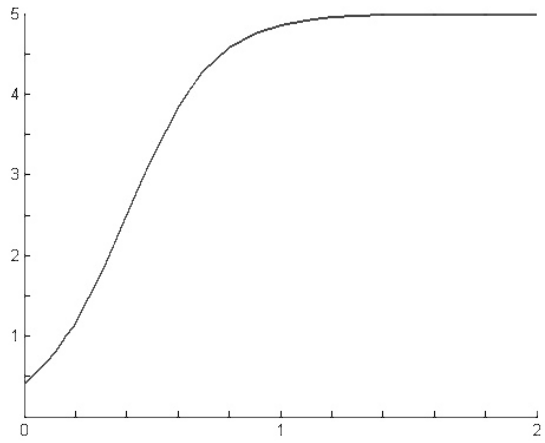


Рис. 2.3. Функція росту $u(t)$

Помітимо, що рішення не залежить від значення n_0, що свідчить про неістотність початкових умов для інформаційної динаміки. Яким би не була початкова кількість публікацій, насичення буде визначатися винятково параметрами, які характеризують фонову швидкість зростання кількості публікацій, кількісну міру актуальності й негативні для процесу фактори.

Крива, представлена на рис. 2.3 має точку перегину:

$$t_{\inf} = \frac{1}{p+D}\ln(\frac{u_s}{n_0}-1)+\tau.$$

Таким чином, для першої області маємо так звану *S*-подібну залежність, а при $t \sim t_{\inf}$ поведінка $u(t)$ наближається до лінійної й відповідає лінійній моделі.

Представимо тепер для зручності вираження для $u(t)$ трохи в іншому вигляді:

$$u(t) = \frac{u_s \exp[(p+D)t]}{\exp[(p+D)t]+(\frac{u_s}{n_0}-1)\exp[(p+D)\tau]},$$

звідки видно, що за умови

$$t < \frac{1}{p+D}\ln(\frac{u_s}{n_0}-1)+\tau = t_{\inf}$$

залежність $u(t)$ має експонентний характер, тобто для значень t, значно менших $t_{\inf}$, модель збігається з експонентною моделлю.

Для другої області, відповідно, маємо (рис. 2.4):

$$v(t) = \frac{v(\lambda)}{qv(\lambda) + (1 - qv(\lambda))\exp[-p(t-\lambda)]},$$

з огляду на умову «зшивки»:

$$v(\lambda) = u(\lambda).$$

Якщо залежність $u(t)$ встигає досягти насичення за проміжок часу $t < \lambda$, то наведене вище рівняння можна спростити, представивши його в такий спосіб:

$$v(t) = \frac{v_s(p+D)}{p + D(1 - \exp[-p(t-\lambda)])},$$

де $v_s = 1/q$ - асимптотичне значення залежності $v(t)$.

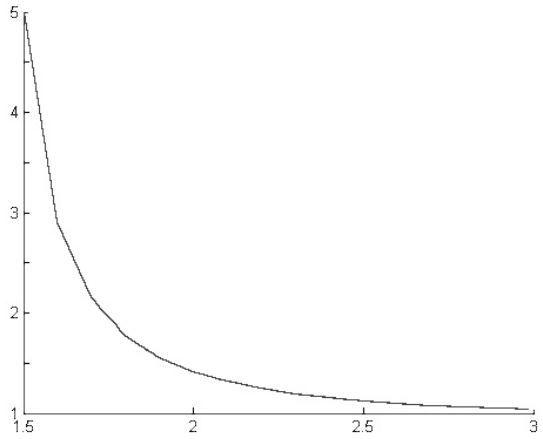


Рис. 2.4. Функція спаду $v(t)$

Як й очікувалося, величина v_s також не залежить ні від початкової умови, ні від умови “зшивання” з функцією $u(t)$ на границі областей. Як видно, отримана залежність має область насичення u_s (при $t \le \lambda$. і асимптотику v_s, що описує поступове зменшення числа публікацій до фонового рівня. А це означає, що вона, принаймні, на якісному рівні, узгоджується із загальною уявою про характер інформаційної динаміки, отриманими на основі дослідних даних. Крім того, на локальних ділянках

вона непогано апроксимується лінійною й експонентною моделями. Типова повна залежність $y(t)$ наведена на рис. 2.5.

У випадку інформаційних потоків, які асоціюються з конкретними темами, необхідно описувати динаміку кожного з таких потоків окремо, беручи до уваги те, що ріст одного з них автоматично приводить до зменшення інших і навпаки. Тому обмеження на кількість повідомлень по всіх тематиках поширюється й на сукупність всіх монотематических потоків.

У випадку вивчення загального інформаційного потоку спостерігається явище “перетікання” публікацій з одних, що гублять актуальність тематик, до інших. Дійсно, кожен інформаційний ресурс, веб-сайт, має певні потужності, які залежать як від технічних аспектів, так і від кон'юнктури предметної області, тобто кожен ресурс публікує більш-менш стандартну кількість повідомлень в одиницю часу.

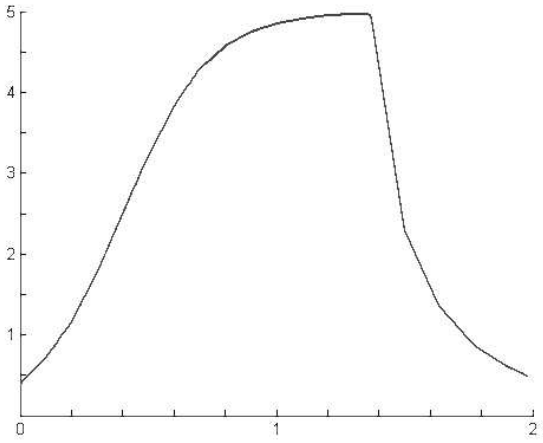


Рис. 2.5. Узагальнений графік динаміки тематичного інформаційного потоку

Загальна динаміка повинна описуватися системою рівнянь, кожне з яких відноситься до окремого монотематичного потоку. Підкреслимо, що загальні політематичні потоки є стаціонарними по кількості публікацій, динаміка ж в основному визначається «конкурентною боротьбою» окремих тематик.

Наведену вище систему рівнянь «конкурентної боротьби» в рамках узагальненої логістичної моделі можна представити в такому вигляді:

$$\frac{dy_i(t)}{dt} = \left(p_i + D_i(t, \lambda_i)\right) \cdot \left(y_i(t) - \sum_j r_{ij} \cdot y_i(t) y_j(t) \right).$$

У цих співвідношеннях коефіцієнти p_i та D_i мають той же зміст, що й раніше, а λ_i є точками, у яких відповідні D_i досягають максимальних значень.

Нижче пропонуються результати моделювання частотних характеристик деяких тематичних інформаційних потоків у рамках логістичної моделі. Безсумнівним достоїнством цієї моделі є те, що вона поєднує в собі простоту вихідних формулювань із гнучкістю в постановці завдань.

2.4. Взаємодія тематичних інформаційних потоків

Канонічне логістичне рівняння описує динаміку одиничної популяції, яка взаємодіє винятково з «навколишнім середовищем». У дійсності ж подібні ситуації виникають украй рідко, оскільки популяції (інформаційні потоки) активно взаємодіють між собою. У теорії популяційної динаміки розроблений класифікація різних форм такої взаємодії [7, 26]:

- нейтралізм (відсутність прямого впливу популяцій один на одного);
- конкуренція (взаємне придушення популяцій);
- аменсалізм (однобічне придушення однієї популяції);
- хижацтво (знищення особинами однієї популяції особин іншої);
- симбіоз (продуктивне співіснування популяцій).

Кожна з цих форм має варіанти, так що загальна картина взаємодії популяцій виглядає досить складною та різноманітною. При цьому варто також ураховувати, що взаємодія популяцій може бути не тільки прямою (наприклад, поїдання одним видом іншого), але й опосередкованим (наприклад, спільне споживання обмежених ресурсів).

У динаміку взаємодіючих популяцій виділяються дві категорії впливів, що відрізняються часовим характером:

- фазові (однократні);
- параметричні (постійні).

У рамках логістистичної моделі опис n взаємодіючих популяцій у загальному випадку здійснюється за допомогою системи рівнянь, записаної у такому вигляді:

$$\frac{dm_i(t)}{dt} = m_i(t)[p_i - \sum_{j=1}^{n} q_{ij} m_j(t)]$$

$$m_i(0) = m_{0i}$$

Тут тип описуваного процесу визначається величиною та знаком коефіцієнтів p_i та q_{ij}, причому варто мати на увазі, що в кожнім рівнянні діагональні члени m_i m_i відповідають внутрішньовидовому, а перехресні m_i m_j - міжвидовій взаємодії.

Важливим моментом є також та поведінка, яку мала б популяція при відсутності взаємодії. Наведена вище система рівнянь може описувати широкий спектр залежностей, однак її рішення (що відносяться до реальних процесів), відповідають одному з наступних режимів:

- стаціонарний;
- автоколивальний;
- квазістохастичний.

Як правило, ці режими повною мірою проявляють себе на досить великих проміжках часу. Разом з тим, перехідні процеси, які передують установленню певного режиму, винятково поліморфни й на превелику силу піддаються класифікації.

Наведений вище опис динаміки популяцій у рамках логістичної моделі спочатку було сформульовано для біологічних систем, однак на даний час поширено також на інші області досліджень, у тому числі й тематичні інформаційні потоки.

У зв'язку з великою можливою кількістю тематик, головним моментом цього дослідження варто вважати характер взаємодії тематичного потоку із зовнішньої стосовно нього середовищем. Біологічні популяції еволюціонують самі по собі, під дією власних іманентних законів, тоді як інформаційні потоки, як правило, породжуються саме змінами, що відбуваються в зовнішньому стосовно інформаційного простору середовищі. Тому ключову роль у даному випадку грають механізми реакції авторів публікацій на події реального життя.

У рамках даної роботи було проведене моделювання типових ситуацій з наступним порівнянням отриманих залежностей з наборами дослідних даних. Таким чином, у першому наближенні можна виявити загальні найважливіші закономірності.

Оскільки ті рівняння, з якими нам доводиться мати справа при рішенні даної проблеми, як правило, не допускають знаходження рішень в аналітичній формі, автори вдалися до чисельних методів.

Монотематична динаміка

Найбільш простий випадок часової залежності числа повідомлень, що надходять до інформаційного потоку у зв'язку з деякою подією. У цьому випадку крива динаміки інформаційного потоку виглядає просто: спочатку вона різко зростає, досягає насичення, а потім убуває, спрямовуючись при цьому до деякого значення, близького до нуля.

Отримані вище залежності в цілому збіглися з дослідними даними, що відносяться до повідомлень певної категорії. Однак були виявлені також інші види тематичних інформаційних потоків, для яких дана модель виявилася неприйнятною.

Крім того, така реалізація моделі є дуже спрощеною, тому необхідно побудувати її в більш адекватному варіанті. У даній роботі замість прямокутної сходинки для визначення змінної інтенсивності тематики використається гладка функція, а саме, така, що відповідає розподілу Гауса:

$$R(t) = ae^{-b(t-\tau)^2}$$

Параметр τ фіксує положення максимуму тимчасової залежності реакції мережі на подію, що пройшла.

При цьому виникає принаймні два способи інтерпретувати поведінки потоку:

- подія відбувається в початковий момент, але реакція на нього зростає поступово (свого роду інерція сприйняття);
- подія відбувається в момент *t*, і реакція на нього досить швидка, але вона очікувана, і тому обговорення її починається заздалегідь.

Обидва ці варіанта реалістичні та можуть зіставлятися з дослідними даними. Типова якісна залежність монотематичного інформаційного потоку від часу наведена на рис. 2.6.

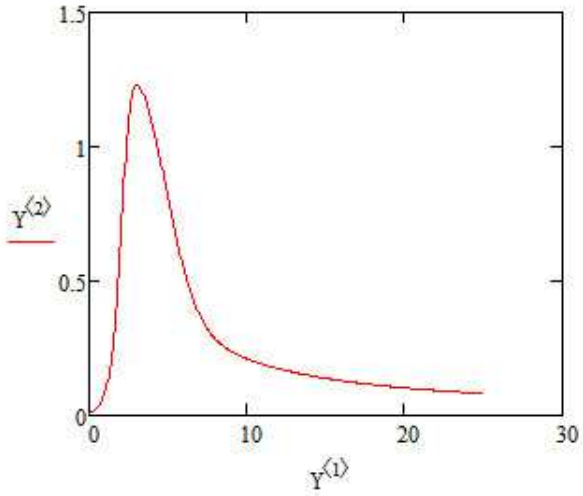


Рис. 2.6. Якісна залежність монотематичного інформаційного потоку (вісь 0Y) від часу (вісь 0X)

Ми бачимо, що зображена на рис. 2.6 крива істотно відрізняється від стандартної гаусіани. Отже, отриманий результат не є тривіальним: вигляд функції $R(t)$ лише частково визначає результуючу форму залежності, що цікавить нас.

При виборі функції $q(t)$, що описує ефект зміни числа публікацій, викликаний зміною ефективного обсягу доступних ресурсів, будемо

виходити з того, що він може проявлятися у двох варіантах, які відрізняються перевагою того або іншого типу зворотних зв'язків: самозбудження та самозагасання. У першому випадку тема вже після завершення подій, що її породили, стає усе більш і більше актуальною, у другому ж її актуальність постійно падає.

У загальному вигляді вираз для $q(t)$ виберемо в такий спосіб:

$$q(t) = (c_0 + ct)^h + d[1 + \sin(\omega t + \varphi)],$$

$$h = \pm 1.$$

Перший член у цьому виразі описує залежно від знака h зменшення (h = 1) або збільшення (h = -1) доступної області ресурсів, а другий – періодичні її зміни (одиниця у квадратних дужках забезпечує те, що функція $q(t)$ є завжди позитивною). Типові залежності, з урахуванням наведеного виразу, представлені на рис. 2.7 і рис. 2.8.

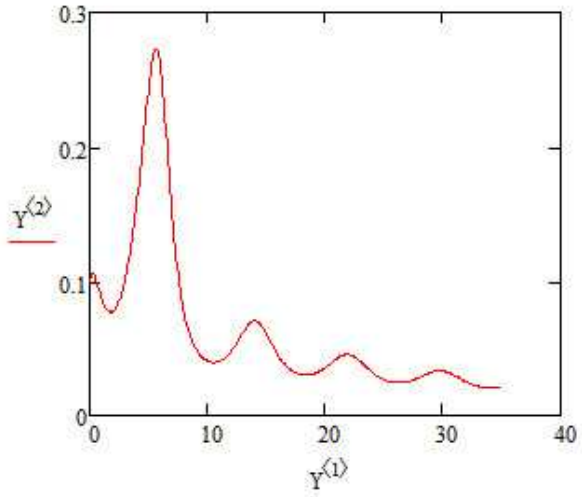


Рис. 2.7. Випадок h = 1

Наведені співвідношення виявляються досить гнучкими, щоб описати загальну поведінку часових залежностей обсягів тематичних інформаційних потоків. Наведені формули містять досить багато параметрів, однак у практичному плані, як правило, їхній набір виявляється надлишковим.

Природно, реальні залежності виглядають набагато складніше внаслідок впливу на них ряду випадкових факторів, які практично не підлягають

явному урахуванню. Приведемо як приклади декілька залежностей, отриманих при аналізі реальних тематичних інформаційних потоків.

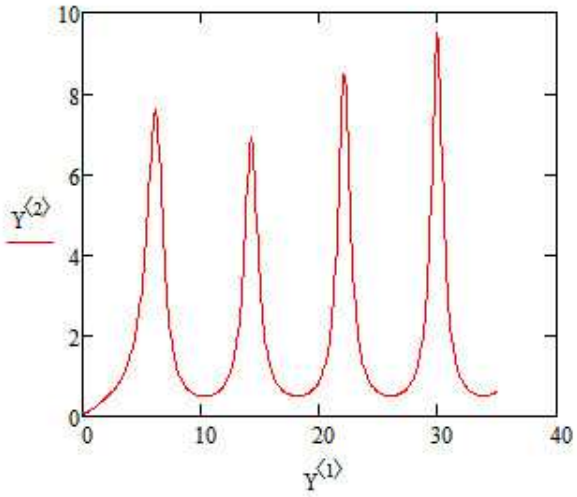


Рис. 2.8. Випадок h = -1

На рис. 2.9 наведено значення інформаційного потоку по темі «теракти» за листопад-грудень 2006 р. Перший пік на наведеному графіку відповідає публікаціям про очікування терактів у Великобританії, а другий - терактам у Багдаді. На рис. 2.10 наведені результати моделювання даного процесу відповідно до наведеної вище формули.

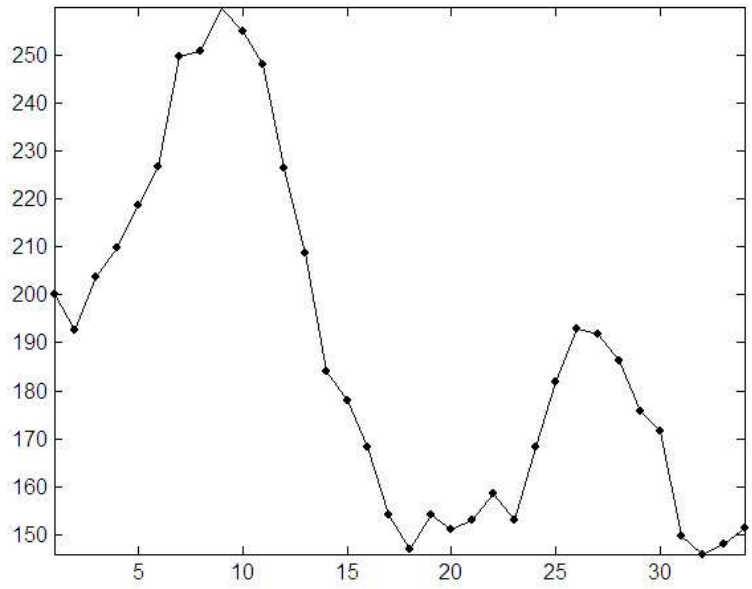


Рис. 2.9. Інформаційний потік за темою «теракти» (вісь X - номер дня, вісь Y - кількість документів)

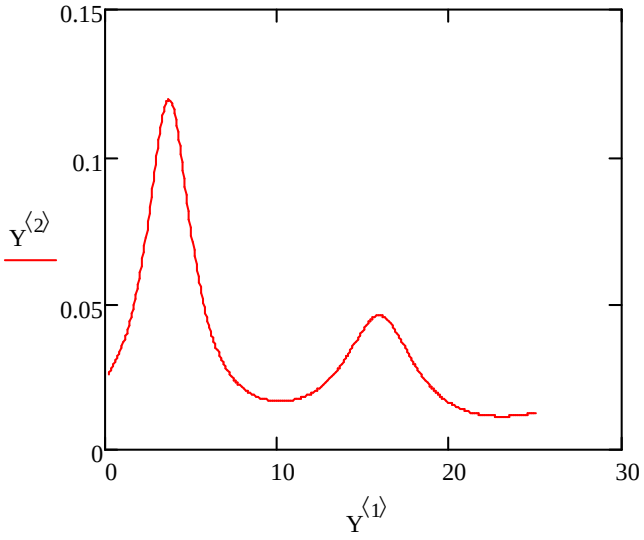


Рис. 2.10. Результат моделювання

На рис. 2.11 наведено значення інформаційного потоку по темі «отруєння О. Литвиненко» за листопад-грудень 2006 р. На рис. 2.12 наведені результати моделювання даного інформаційного потоку.

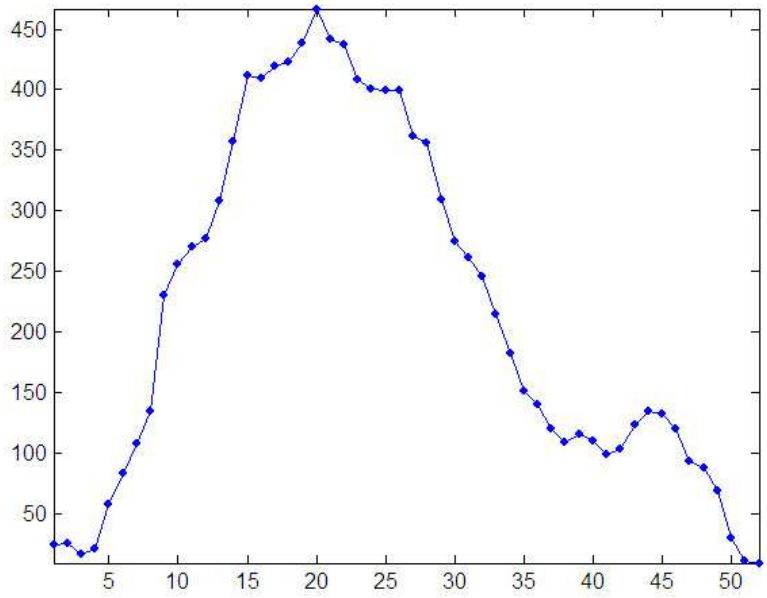


Рис. 2.11. Інформаційний потік за темою «отруєння О. Литвиненко»

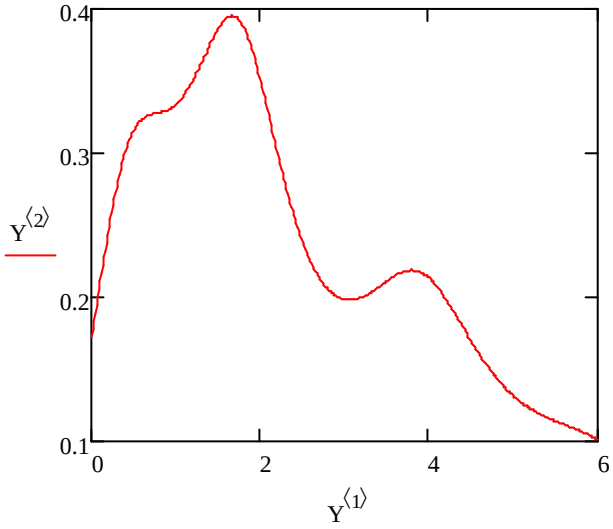


Рис. 2.12. Результат моделювання

На рис. 2.13 наведено значення інформаційного потоку за темою «Хізбалла» за жовтень-грудень 2006 р. На рис. 2.14 наведені результати моделювання даного інформаційного потоку. У цьому випадку звертає на себе увагу явна циклічність розглянутого ряду. При цьому циклічність, пов'язана із днями тижня була заздалегідь «згладжена».

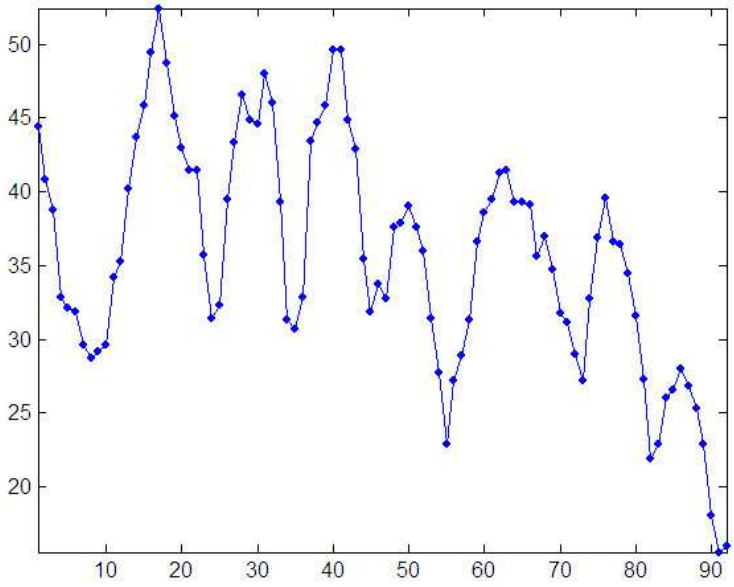


Рис. 2.13. Інформаційний потік за темою «Хізбалла»

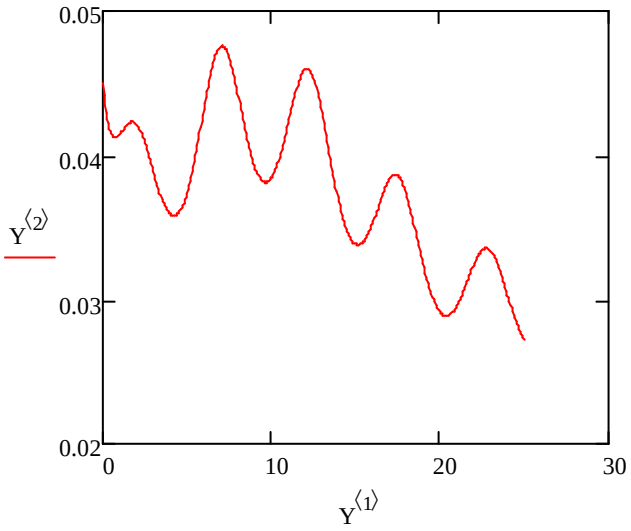


Рис. 2.14. Результат моделювання

Наведений модельний опис експериментальних даних на перший погляд може здатися занадто спрощеним, оскільки не становить великої складності змоделювати відповідну криву у рамках інших парадигм. Сильна ж сторона даної моделі полягає у тому, що при її простоті вона дозволяє без подальших істотних ускладнень описувати такі нетривіальні процеси, як динаміку взаємодіючих тематик.

Динаміка взаємодіючих тематик

Вивчення динаміки у випадку взаємодії тематик ускладнюється тією обставиною, що реальні інформаційні потоки містять безліч залежностей, у відношенні яких, загалом кажучи, важко сказати, які з них взаємодіють переважно між собою. Більше того, одержавши той або інший набір дослідних даних, зазвичай буває важко з повною визначеністю віднести його до взаємодіючих або невзаємодіючих тематик.

Основні типи взаємодій, що описуються системами логістичних рівнянь добре відомі й включають декілька характерних форм. Як приклад опишемо дві найцікавіші з них: конкуренцію й симбіоз.

Конкуренція

Випадку конкуренції відповідають позитивні значення обох перехресних коефіцієнтів q_{ij}. Це означає, що взаємодія тематик відбувається таким чином, що зростання числа публікацій за однією з них супроводжується скороченням числа публікацій за іншою.

На рис. 2.15 і 2.16 наведено два характерних випадки конкуренції тематик.

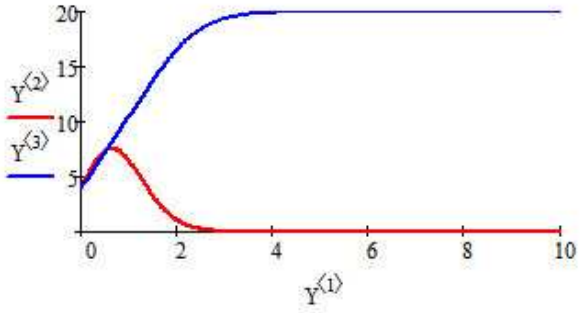


Рис. 2.15. Конкуренція, випадок 1

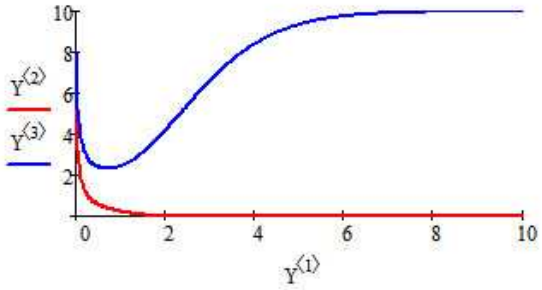


Рис. 2.16. Конкуренція, випадок 2

Цікавим випадком конкуренції є автоколивальний режим, у якому кількості публікацій здійснюють незатухаючі коливання. Він може виникати при нульових значеннях діагональних коефіцієнтів q_{ii}.

Приклад подібної залежності зображено на рис. 2.17.

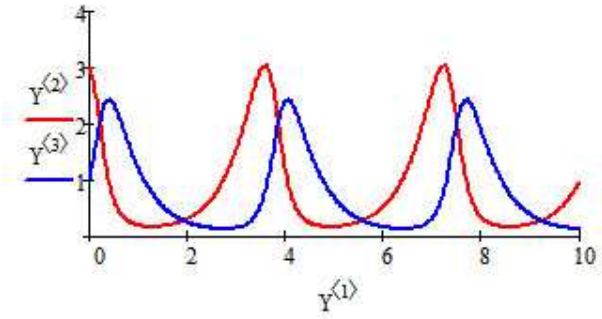


Рис. 2.17. Конкуренція, випадок 3 - автоколивальний режим

Симбіоз

Симбіоз виникає при негативних значеннях коефіцієнтів p_2 та q_{21}, тобто при умовах, коли тематичні потоки не тільки споживають певні ресурси, але й «підживлюють» один одного. Приклад наведений на рис. 2.18.

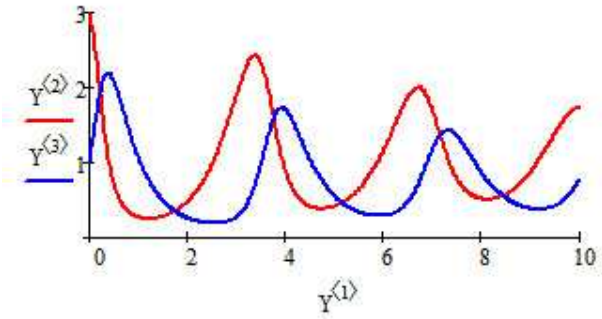


Рис. 2.18. Симбіоз

Щоб краще зрозуміти особливості узагальненої логістичної моделі, подивимося на те, як тематики впливають одна на одну. Для цього спочатку побудуємо залежності для двох окремих тематик, які еволюціонують кожна за законом, обумовленим функціями $p(t)$ і $q(t)$, а потім дослідимо їхню спільну динаміку, за умови, що відповідні закони залишилися незмінними. Результати наведені відповідно на рис. 2.19 та 2.20.

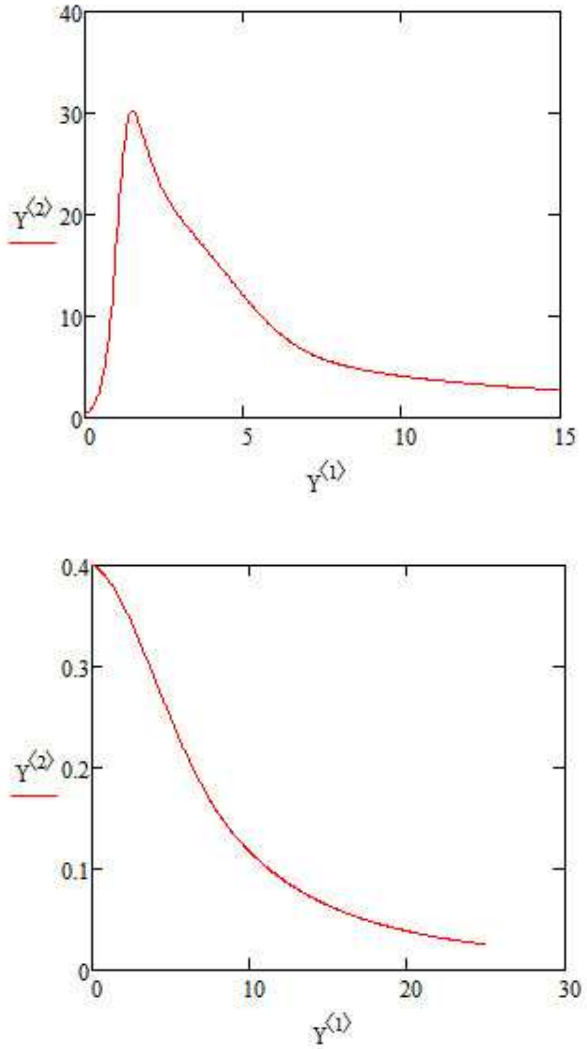


Рис. 2.19. Роздільна еволюція тематик

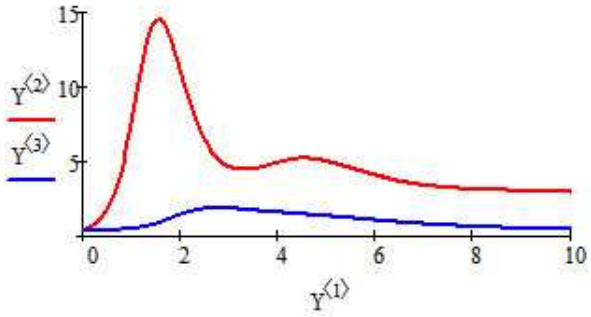


Рис. 2.20. Сумісна динаміка тематик

З рис. 2.20 видно, що взаємний вплив тематик носить не тільки кількісний, але і якісний характер. Поведінка кожної кривої тепер, дійсно, визначається не тільки їхніми власними функціями $p(t)$ і $q(t)$, але й характером впливу одна на одну.

Приведемо як приклади кілька взаємних залежностей реальних тематичних інформаційних потоків.

На рис. 2.21 наведено спільні значення інформаційних потоків за тематиками Nokia й Motorola за листопад-грудень 2006 р., а на рис. 2.22 - результат моделювання цього процесу. На рис. 2.23 наведені значення спільних тематичних потоків, обумовлених двома особистостями - Ахметовим і Богатирьовою у контексті Партії Регіонів на вказаний час. Відповідно, на рис. 2.24. наведені результати моделювання.

Як ми бачимо, обидва наведених випадки відбивають стан конкуренції інформаційних потоків.

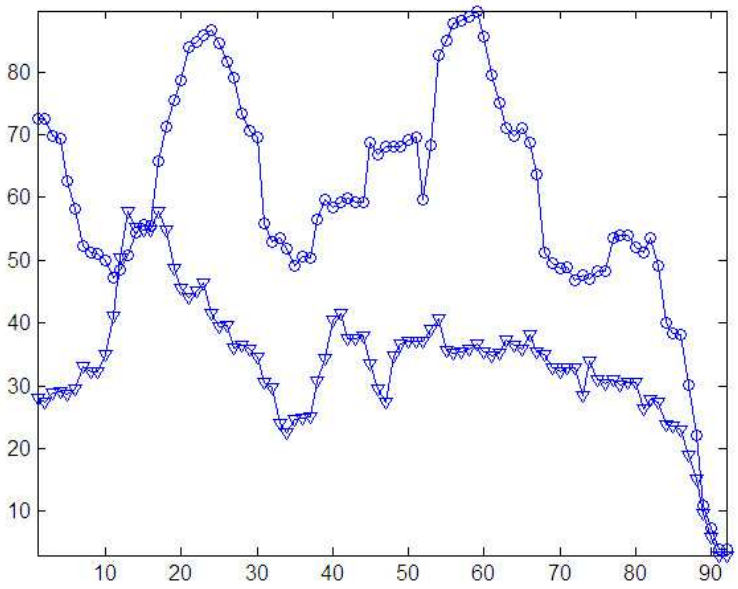


Рис. 2.21. Потоки за темами Nokia (▽) і Motorola (o)

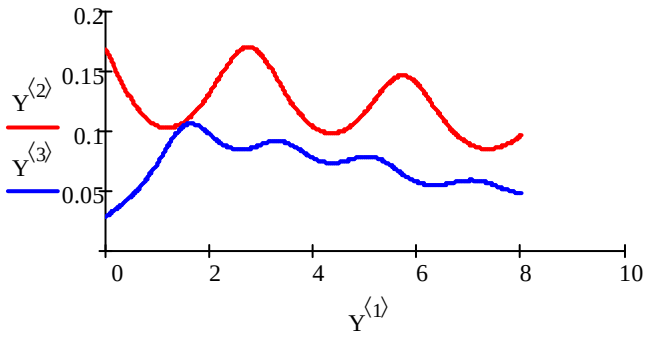


Рис. 2.22. Результати моделювання

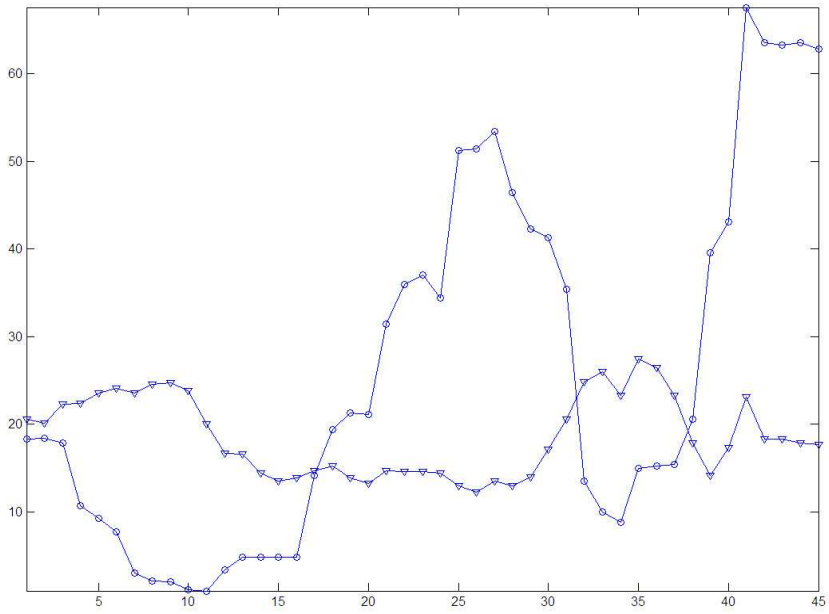


Рис. 2.23. Потоки по персонам – Ахметов (▽) и Богатирьова (о)

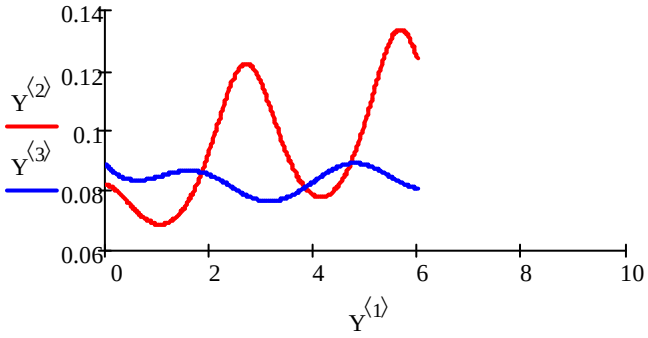


Рис. 2.24. Результати моделювання

2.5. Емерджентний підхід до моделювання

У складних системах (а сучасні інформаційні мережі, без сумніву, є такими), серед багатьох інших характеристик, найбільш чітко проявляється цілісність, тобто наявність таких властивостей, які не притаманні жодному елементу (у випадку, який розглядається, - документу), що складають систему, узятому окремо. Ця властивість, яку називають «емерджентністю», є результатом виникнення між елементами системи особливих синергетичних зв'язків. Під терміном «емерджентність», уперше введеному Ф. Льюісом розуміється те, що у системах ціле є найчастіше більшим, ніж сума частин [68], тобто на кожному рівні складності виникають нові, часто непередбачені якості, які не властиві складовим частинам.

Так, наприклад, якщо в якості системи розглядати годинник - прилад, що показує поточний час, то жодна з його деталей час показувати не зможе. Вона не може показувати навіть «частину часу». Властивість показувати час з'являється у всіх деталей разом, причому після того, як вони будуть певним чином зібрані в єдиний комплекс та, тим самим, вступлять один з одним у певну взаємодію.

Емерджентність інформаційної системи не дає можливості обмежитися вивченням її елементів і зв'язків між ними, а припускає цілісний аналіз всієї системи. До кінця XX-го століття при аналізі складних, у тому числі й соціальних систем, в основному використовувався редукціоністський підхід, що мав на меті пояснити множину властивостей складних систем властивостями їхніх елементів – «атомів» або «молекул». Внаслідок розвитку системного аналізу, появи науки про складність, технологічному прориву в обчислювальних можливостях ситуація різко змінилася. Зараз набули розвитку такі напрямки, як теорії хаосу, складних мереж, нелінійних систем і систем, що самоорганізуються. Виявилося, що багато властивостей складних систем не можуть бути виведені із заздалегідь визначеного набору динамічних рівнянь. Навпаки, рівняння можуть бути отримані тільки в результаті чисельного моделювання.

Разом з тим, очевидно, що неможливо розробити та застосовувати на практиці деяку універсальну методику моделювання інформаційних систем. Це в першу чергу пов'язане зі слабкою формалізацією багатьох понять і факторів, насамперед суб'єктивних. У кожному окремому випадку доводиться довіряти інформованості та інтуїції аналітиків, які професійно займаються питаннями аналізу інформаційних процесів. Іноді їм вдається точно прогнозувати окремі закономірності, параметри, які чітко проявляються на рівні суспільної практики.

З об'єктивними факторами справа полягає інакше. Вони цілком піддаються аналізу на статистичному рівні й допускають кількісні оцінки, які можуть використовуватися для побудови обґрунтованих прогнозів. Сучасні

методи прикладної статистики, аналізу часових рядів включають великий арсенал детально розроблених та апробованих методів. Однак статистика дозволяє описувати лише формальні аспекти явищ, які вивчаються, залишаючи за бортом аспекти змістовні. Тому існує необхідність розширення набору інструментальних засобів, що використовуються при аналізі та моделюванні інформаційних потоків. Одним з найбільш перспективних напрямків у цьому плані є математичне моделювання. Сьогодні математичне моделювання широко застосовується в багатьох галузях науки і техніки, разом з тим, моделювання інформаційних систем залишається відкритою проблемою.

Стосовно інформаційних потоків перспективним є моделювання, обумовлене деякими реалістичними правилами поведінки окремих елементів системи (документами, тематиками), що уточнюються деякими параметрами, які змінюються при моделюванні. У цьому випадку велику цінність отримує також і зворотна задача - за реальною поведінкою деякої залежності оцінити величину параметрів моделі. Знання загальної поведінки стійких рішень дозволяє прогнозувати розвиток інформаційних потоків навіть у тому випадку, коли не існує точного уявлення про конкретні механізми, що визначають їхню динаміку, причому такого роду прогнози можуть виявитися більш точними, ніж отримані традиційними експертними методами. Якщо ж рішення виявляються нестійкими, то із цього також може бути отримана важлива інформація щодо системи, яка дозволяє в окремих випадках прогнозувати, в який бік може бути спрямована динаміка окремих інформаційних потоків.

Спроби моделювання інформаційних потоків здійснювалися вже давно, але вони гальмувалися обчислювальними труднощами, особливо у випадку необхідності опису динаміки систем зі зворотними зв'язками. На цей час є досить багато можливостей для ефективної комп'ютерної обробки даних, що дозволяє, з одного боку, підготовляти набори вхідних параметрів на підставі аналізу результатів статистичних досліджень, а з іншого боку - вирішувати

формалізовані задачі з достатнім ступенем точності та у припустимий час. Все це дає підстави думати, що найближчим часом математичне моделювання стане основним інструментальним засобом аналізу та керування інформаційними потоками.

Одним із застосувань концепції емерджентності до моделювання на даний час є багатоагентне моделювання. Ідея багатоагентного (*Agent-based*) моделювання виникла в середині XX-століття. Відповідно до неї агент - це деяка абстрактна сутність, який притаманна активність автономна поведінка, можливості приймати рішення відповідно до деякого набору правил, взаємодіяти з оточенням та іншими агентами, а також еволюціонувати. Багатоагентні моделі широко застосовуються для аналізу децентралізованих систем, закономірності динаміки функціонування яких не вивчені в достатній мірі. Ці моделі використовуються з метою отримання уяви щодо загальної поведінки складних систем, виявлення правил функціонування систем з урахуванням припущень про індивідуальну поведінку її окремих компонентів (агентів). Мета багатоагентного моделювання може бути сформульована як створення комп'ютерних мікросвітів, у яких агенти взаємодіють, реагуючи на умови зі свого оточення та здійснюючи зміни.

Відповідно до визначення К. Лангтона [50] моделювання складних адаптивних систем часто ґрунтується на наступних принципах:

- модель складається з популяції простих агентів;
- не існує єдиного агента (центру), що направляє інших агентів;
- кожен агент докладно розглядає способи, якими здійснюється проста реакція на локальні зміни в оточенні, включаючи контакти з іншими агентами;
- не існує єдиного правила в системі, яке б описувало глобальну поведінку.

Відповідно до цих принципів будь-яка поведінка на рівні вищім за індивідуальний, є емерджентною, породженою взаємодіями локальних

агентів. Тобто прості правила можуть викликати складну поведінку та структури.

Багатоагентні моделі, на відміну від динамічних моделей є децентралізованими. При цьому складна глобальна поведінка системи є результатом діяльності великої кількості агентів, кожний з яких функціонує за простими правилами, оточений іншими агентами і взаємодіє з ними та з середовищем. Багатоагентні моделі дозволяють досліджувати досить широке коло проблем, для яких суворі аналітичні методи виявляються неефективними.

Останнім часом у зв'язку з бурхливим розвитком комп'ютерних технологій важливим і перспективним, з погляду застосування на практиці математичного моделювання, є клас так званих імітаційних моделей. Така модель являє собою алгоритм, за допомогою якого комп'ютер генерує набори даних, що описують задані характеристики реальної системи, яка представляє інтерес. При цьому операції, що виконуються машиною, не мають ніякого відношення до природи та властивостей системи, яка досліджується. Відзначимо, що сам по собі факт з'ясування можливості застосування імітаційного моделювання є чималим досягненням сучасної науки. Дійсно виявляється, що структура реального процесу у значній мірі не залежить від його природи і, так сказати, матеріальної основи. Числа, що одержуються в результаті маніпулювання іншими числами за певними абстрактними правилами, можуть у точності відповідати числам, що описують конкретні процеси, які відбуваються в реальному світі.

Зрозуміло, при розробці імітаційної моделі приймаються в розрахунок властивості явища, яке досліджується, але на рівні не внутрішніх механізмів, які або не відомі, або занадто складні для явного використання, а загальних характеристик протікання відповідних процесів.

У плані практичного застосування імітаційні моделі зручні тим, що дозволяють проводити так звані машинні експерименти, метою яких є вивчення зміни поведінки об'єкта дослідження залежно від змін внутрішніх

параметрів або (і) зовнішніх умов. При цьому імітаційне моделювання (при наявності задовільних моделей) дозволяє одержати дані на цілком прийнятному рівні точності.

Побудова імітаційних моделей - це досить складне завдання, яке вимагає крім знання предметної області, ще й високого професіоналізму у сфері програмування. Однак у випадку успіху результати найчастіше окуплюють витрати.

При цьому ближче всього до моделювання інформаційних потоків виявляється напрямок індивідуум-орієнтованого моделювання (англ. *Individual-based Modelling*), яке, у свою чергу, є компонентою багатоагентного моделювання. Нижче наведено опис моделі, побудованої на основі клітинних автоматів, найбільш доступного інструмента в рамках індивідуум-орієнтованого моделювання. «Клітинні» моделі, через свою простоту можуть сприйматися як абстрактні іграшки, що дають лише якісні результати, які лише віддалено нагадують реальність. Однак, як зауважують багато дослідників, при правильному завданні правил функціонування клітинних автоматів, вони не рідко дають більш реалістичні результати, ніж інші моделі.

2.6. Моделі на базі теорії клітинних автоматів

Звернемося ще до одного напрямку у вивченні процесів, пов'язаних з інформаційними потоками - до дифузії інформації. Нагадаємо, що в природничих науках під дифузією розуміють взаємне проникнення друг у друга дотичних речовин, викликане, наприклад, тепловим рухом їхніх часток.

Модель базується на тому припущенні, що інформація також у певному сенсі складається з «часток» - документів (повідомлень). Множину процесів, близьких до динаміки інформаційних потоків, можна моделювати досить точно, якщо чітко параметризувати та встановити їхні граничні параметри.

Процеси дифузії інформації, як і процеси дифузії у фізиці, досить точно моделюються за допомогою методів клітинних автоматів. Концепція клітинних автоматів була вперше запропонована більше півстоліття тому назад Дж. фон Нейманом (J. Von Neumann) [29] і розвинена С. Вольфрамом (S. Wolfram) у фундаментальній монографії “Новий вид науки” [64, 64].

Клітинні автомати є корисними дискретними моделями для дослідження динамічних систем [31]. Дискретність моделі, а точніше, можливість представити модель у дискретній формі, може вважатися важливою перевагою, оскільки відкриває широкі можливості використання комп'ютерних технологій. Клітинні автомати в цьому сенсі займають особливе місце, оскільки їхня дискретність поєднується з іншими перевагами.

Головним достоїнством клітинних автоматів є їхня абсолютна сумісність із алгоритмічними методами рішення задач. Скінчений набір формальних правил, заданий на обмеженій множині елементів (кліток), допускає точну реалізацію у вигляді алгоритмів. Однак звідси ж випливає й головний недолік клітинних автоматів: обчислювальні труднощі, які виникають при розрахунках відповідних масштабів. Адже на кожній ітерації необхідно сканувати весь набір кліток і для кожної з них виконувати необхідні операції. Коли кліток і ітерацій дійсно багато потрібні значні ресурси, у тому числі обчислювальні та часові.

Система клітинних автоматів являє собою дискретну динамічну систему, сукупність однакових кліток, однаковим образом з'єднаних між собою. Всі клітки утворять мережу (сітку) клітинних автоматів. Стан кожної клітки визначаються станом кліток, що входять у її локальний окіл (найближчих сусідів). Околом кінцевого автомата з номером j називається множина його найближчих сусідів. Стан j-го клітинного автомата в момент часу $t+1$, таким чином, визначається в такий спосіб:

$$y_j(t+1) = F\left(y_j,\ O(j),\ t\right),$$

де F – деяке правило, яких можна виразити, наприклад, мовою булевої алгебри. У багатьох задачах вважається, що сам елемент відноситься до своїх найближчих сусідів, тобто $y_j \in O(j)$, у цьому випадку формула спрощується: $y_j(t+1) = F\left(O(j),\ t\right)$. Клітинні автомати у традиційному розумінні задовольняють таким правилам:

- зміна значень всіх кліток відбувається одночасно (одиниця виміру - такт);
- мережа клітинних автоматів однорідна, тобто правила зміни станів для всіх кліток однакові;
- на клітку можуть вплинути лише клітки з її локальної околиці;
- множина станів клітки кінцева.

Теоретично клітинні автомати можуть мати будь-яку розмірність, однак найчастіше розглядають одномірні й двовимірні системи клітинних автоматів.

Модель дифузії інформації, що будемо розглядати надалі, є двовимірною, тому подальший формалізм стосується цього випадку. У двовимірному клітинному автоматі сітка реалізуються двовимірним масивом. Тому в цьому випадку зручно перейти до двох індексів, що цілком коректно для двовимірних кінцевих сіток.

У випадку двовимірних сіток, елементами яких є квадрати, найближчими сусідами, що входять в окіл елемента $y_{i,j}$, можна вважати або тільки елементи, розташовані вниз і вліво-вправо від нього (так називана околиця фон Неймана: $y_{i-1,j}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j}$), або додані до них ще й діагональні елементи - окіл Г. Мура: $y_{i-1,j-1}$, $y_{i-1,j}$, $y_{i-1,j+1}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j-1}$, $y_{i+1,j}$, $y_{i+1,j+1}$).

У моделі Мура кожна клітка має вісім сусідів. Для усунення крайових ефектів сітка топологічно «згортається в тор» (рис. 2.25), тобто перший рядок вважається продовженням останнього, а останній - попередником

першого. Те ж саме відноситься й до стовпців. Це дозволяє визначати загальне співвідношення значення клітки на кроці $t+1$ в порівнянні із кроком t:

$$y_{i,j}(t+1) = F\left(y_{i-1,j-1}(t), y_{i-1,j}(t), y_{i-1,j+1}(t), y_{i,j-1}(t), y_{i,j}(t), y_{i,j+1}(t), y_{i+1,j-1}(t), y_{i+1,j}(t), y_{i+1,j+1}(t)\right).$$

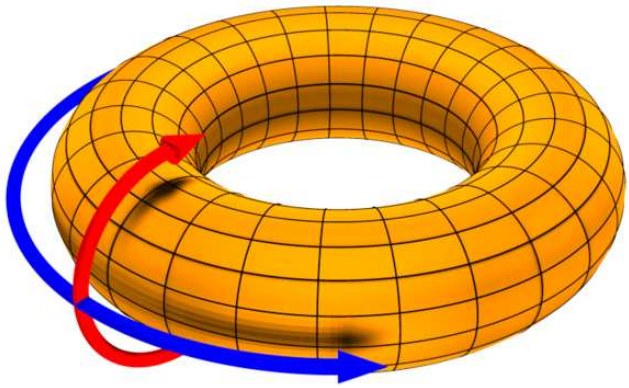


Рис. 2.25. Згортання площини у тор. Джерело: wikimedia.org

С. Вольфрам, класифікуючи різні клітинні автомати, виділив ті, динаміка яких істотно залежить від початкового стану. Підбираючи різні початкові стани, можна одержувати різноманітні конфігурації та типи поведінки. Саме до таких систем відноситься класичний приклад - гра "Життя", винайдена Дж. Конвеєм (J. Conway) і відома широкому загалу читачів завдяки публікації в книзі М. Гарднера (M. Gardner) [6].

Клітинні автомати з успіхом застосовуються при моделюванні процесів поширення інновацій [41], застосовуються при моделюванні електоральних процесів, у цьому випадку передбачається, що виборчі преференції людини визначаються установками її найближчого оточення [35]. В одній з моделей передбачається, що індивід приймає рішення в момент $t+1$ за республіканців або демократів відповідно до правила простої більшості. У цій моделі враховувалися погляди індивіда та чотирьох його найближчих сусідів у момент t (околиця фон Неймана). Модель досліджувалася на великому часовому відрізку - до 20 000 тактів. Виявилося, що партійна боротьба приводить до дуже складних конфігурацій, які істотно залежать від вихідного розподілу.

Як спрощену модель дифузії інформації спочатку розглянемо визнану модель поширення інновацій [41]. Подібна модель функціонує за такими правилами: кожен індивід, що здатний прийняти інновацію, відповідає одній квадратній клітці на двовимірній площині. Кожна клітка може перебувати у двох станах: 1 - новинка прийнята; 0 - новинка не прийнята. Передбачається, що автомат, сприйнявши інновацію один раз, запам'ятовує її назавжди (стан 1, що не може бути зміненим). Автомат схвалює рішення відносно прийнятті новинки, орієнтуючись на думку восьми найближчих сусідів, тобто якщо в околі даної клітки (використається окіл Мура) є m прихильників новинки, p - імовірність її прийняття (генерується в ході роботи моделі) і якщо $pm > R$, (R - фіксоване граничне значення), те клітка приймає інновацію (значення 1). На думку авторів цієї моделі, клітинне моделювання дозволяє будувати значно більше реалістичні моделі ринку інновацій, чим традиційні підходи.

Разом з тим динаміці поширення інформації властиві деякі додаткові властивості, які були враховані в представленій нижче моделі. У моделі дифузії інформації, поряд з тими ж умовами, які відносяться до клітинного простору, околу Мура та імовірнісне правило прийняття новини, додатково до в умовам дифузії інновацій передбачалося, що клітка може бути в одному із трьох станів: 1 - «свіжа новина» (клітка офарблюється в чорні кольори); 2 - новина, що застаріла, але збережена у вигляді відомостей (сіра клітка); 3 - клітка не має інформації, переданої новинним повідомленням (клітка біла, інформація не дійшла або вже забута). У моделі прийняті такі правила поширення повідомлень:

- спочатку все поле складається з білих кліток за винятком однієї, чорної, котра першої «прийняла» новину (рис. 2.26а);
- біла клітка може перефарбовуватися тільки в чорні кольори або залишатися білої (вона може одержувати новину або залишатися «у невіданні»);

- біла клітка перефарбовується, якщо виконується умова, аналогічне моделі дифузії інновацій: $pm > 1$ (ця умова трохи модифікується для $m \le 2$: $1.5 \cdot pm > 1$);
- якщо клітка чорна, а навколо її винятково чорні й сірі, то вона перефарбовується в сірі кольори (новина застаріває, але зберігається як відомості);
- якщо клітка сіра, а навколо її винятково сірі й чорні, то вона перефарбовується в білі кольори (відбувається старіння новини при її загальновідомості).

Описана система клітинних автоматів цілком реалістично відбиває процес поширення повідомлень серед окремих інформаційних джерел. Було реалізовано наведений вище алгоритм на полі розміром 40 х 40 (розміри були обрані винятково з метою наочності). З'ясувалося, що стан системи клітинних автоматів повністю стабілізується за обмежену кількість тактів, тобто процес еволюції виявився збіжним. Приклад роботи моделі наведений на рис. 54.

Численні експерименти з даним клітинним автоматом показують, що період його збіжності становить від 80 до 150 кроків.

Типові залежності кількості кліток, які перебувають у різних станах залежно від кроку ітерації наведені на рис. 2.26. При аналізі наведених графіків варто звернути увагу на такі особливості: 1 - сумарна кількість кліток, які перебувають у всіх трьох станах на кожному кроці ітерації постійна й дорівнює кількості кліток, 2 - при стабілізації клітинних автоматів співвідношення сірих, білих і чорних кліток приблизно становить: 3:1:0; існує точка перетину всіх трьох кривих.

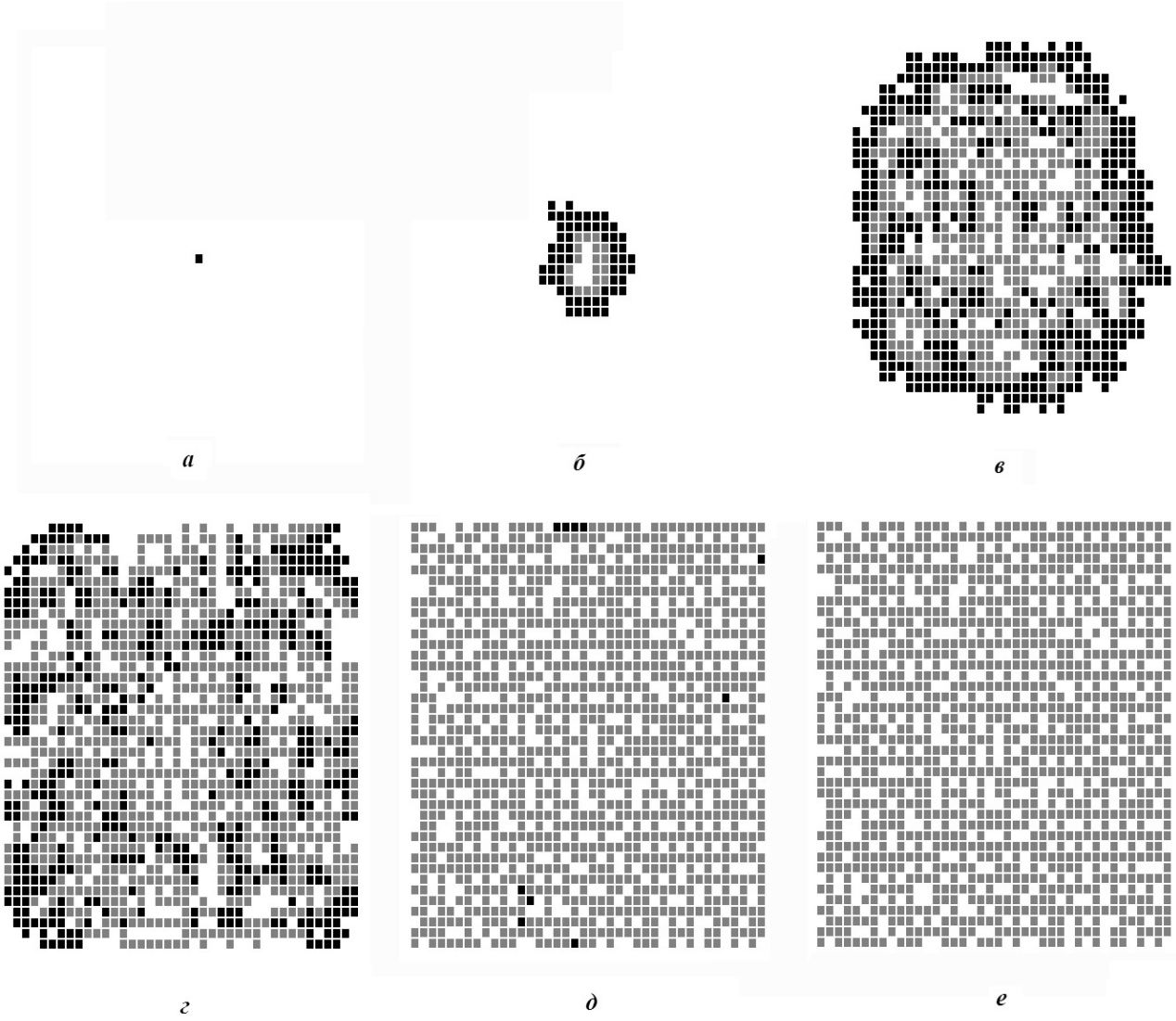


Рис. 2.26. Процес еволюції системи клітинних автоматів «дифузії новин»: а - вихідний стан; б-д - проміжні стани; е - кінцевий стан

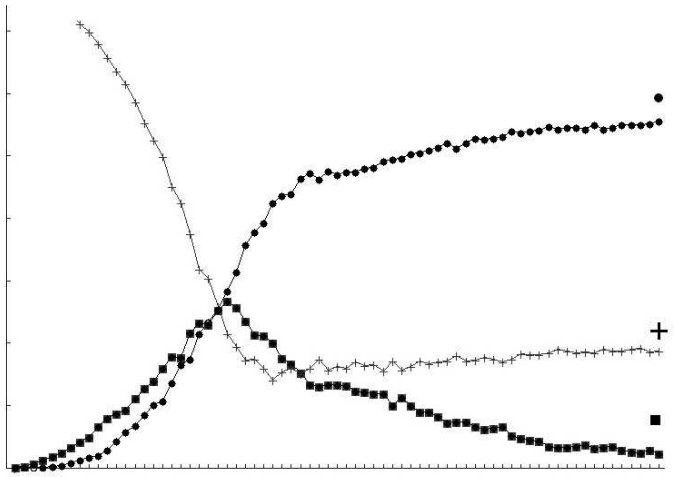


Рис. 2.27. Кількість кліток кожного із кольорів залежно від кроку еволюції: білі клітки - (+); сірі клітки - (•); чорні клітки – (■)

Детальний аналіз отриманих залежностей дозволив провести аналогії даної моделі «дифузії інформації» з деякими аналітичними міркуваннями [35]. Результати моделювання дають підстави припустити, що еволюція сірих кліток описується деякою безперервною функцією:

$$x_g = f\left(t, \tau_g, \gamma_g\right),$$

де t - час (крок еволюції), τ_g - зсув за часом, що забезпечує одержання необхідного фрагмента аналітичної функції, γ_g - параметр крутості даної функції.

Відповідно, динаміка білих кліток x_w (кількість кліток у момент t) може моделюватися «перевернеою» функцією x_g з аналогічними параметрами:

$$x_w = 1 - f\left(t, \tau_w, \gamma_w\right).$$

Оскільки, як було сказано вище, завжди виконується умова балансу, тобто загальна кількість кліток у будь-який момент часу завжди постійно, та умова нормування можна записати в такий спосіб:

$$x_g + x_w + x_b = 1,$$

де x_w - кількість чорних кліток у момент часу t.

Таким чином, одержуємо:

$$x_b = 1 - x_g - x_w = f\left(t, \tau_w, \gamma_w\right) - f\left(t, \tau_g, \gamma_g\right).$$

Вигляд представленої на рис. 2.25 залежності дозволяє припустити, що як функція $f\left(t, \tau, \gamma\right)$ може бути обране наступне вираження (логістична функція):

$$f\left(t, \tau, \gamma\right) = \frac{C}{1 + e^{\gamma(t-\tau)}},$$

де C - деяка константа, що нормує.

На рис. 2.28 наведені графіки залежності x_g, x_w, x_b від кроку еволюції системи клітинних автоматів, отримані в результаті аналітичного моделювання.

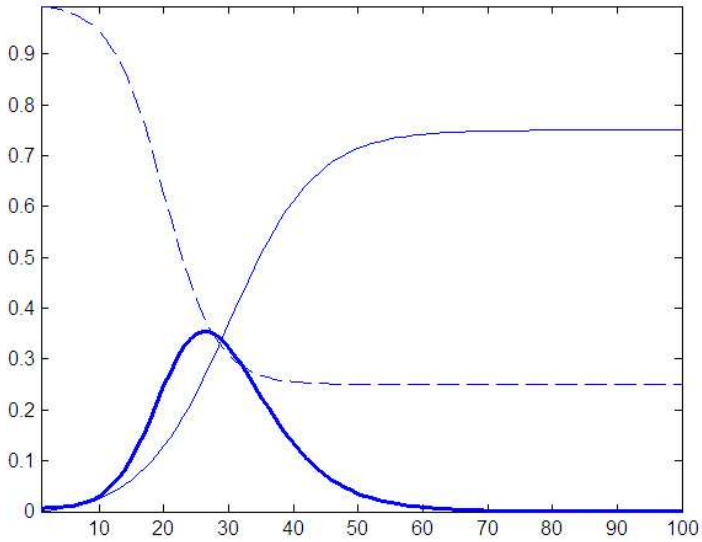


Рис. 2.28. Безперервні залежності, отримані в результаті аналітичного моделювання, залежно від кроку еволюції: суцільна лінія – сірі (x_g); пунктирна лінія – білі (x_w); суцільна жирна лінія – чорні (x_b)

Слід зазначити, що залежність дифузії новин, отримана в результаті моделювання, добре погоджується з «життєвою» поведінкою тематичних інформаційних потоків на інтернет-джерелах (веб-сайтах), а на локальних часових проміжках - із традиційними моделями.

3. МЕТОДИ АНАЛІЗУ ІНФОРМАЦІЙНИХ ПОТОКІВ

3.1. Часові ряди з параметрів інформаційних потоків

Однією з підстав успішного моделювання та аналізу інформаційних потоків є врахування оперативно отриманих за допомогою спеціального програмного інструментарію масивів параметрів цих потоків, які дозволяють більш глибоко зрозуміти природу предметної області.

Розглянемо можливості засобів аналізу часових рядів на прикладі дослідження інформаційних потоків веб-публікацій, зібраних з мережі Інтернет системою контент-моніторингу InfoStream [8]. Ця система забезпечує доступ до унікального ретроспективного фонду, що перевищує 80 млн. записів за 10 років та підтримку аналітичної роботи в режимі реального часу, у тому числі побудову сюжетних ланцюжків, дайджестів, діаграм появи у часі та таблиць взаємозв'язків понять.

Тематика досліджуваного інформаційного потоку може визначатися запитами визначалася запитом до системи InfoStream, у розглянутому нижче прикладі запиту, який відноситься до розвитку кризових явищ в Україні впродовж 2008 року [34]:

(парламентсь~криз)|(політичн~криз)|(фінансов~криз)|(економічн~криз)

В цьому досліджені аналізувався інформаційний потік, який формувався з понад тисячі українських мережних інформаційних ресурсів, серед яких лідерами за кількістю релевантних запиту публікацій були такі авторитетні джерела, як Укрінформ, УНІАН, РБК-Україна, Радіо Свобода, Кореспондент.net, Главред тощо. Ретроспективний період дослідження становив весь 2008 рік, тобто 366 днів. За цей період системою InfoStream було охоплено понад 12 млн. документів. У результаті пошуку за вищенаведеним запитом було знайдено 57245 релевантних документів. На основі обробки цих даних було отримано відповідну картину експериментальних даних - часовий ряд за заданий період.

На рис. 3.1 наведено графік кількості відповідних тематичних публікацій за днями 2008 року.

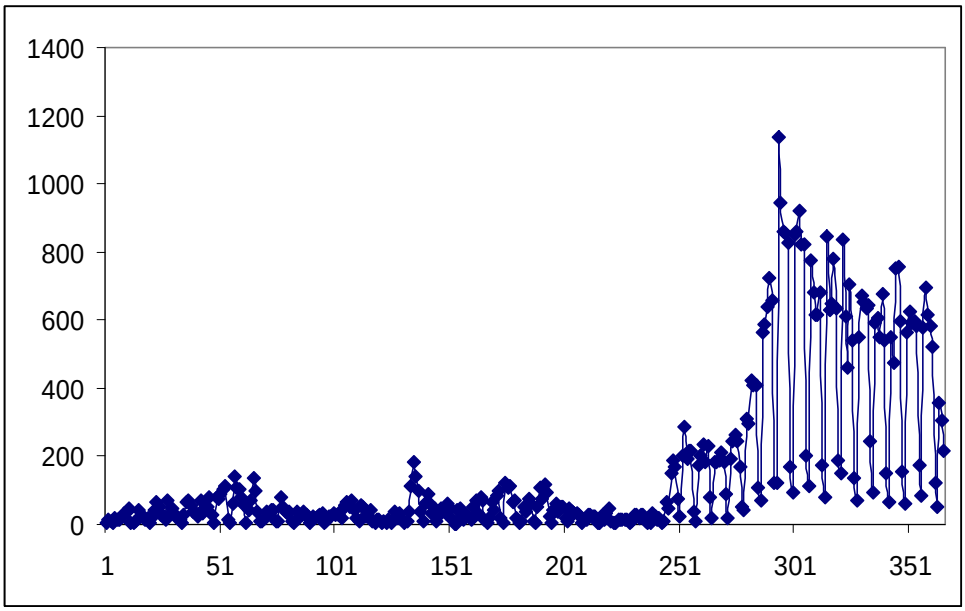


Рис. 3.1. Динаміка кількості публікацій за першим запитом за днями 2008 року (разом 57245 публікацій)

Наведений графік враховує тижневі коливання (у вихідні дні, наприклад, в мережі публікується значно менше документів, ніж у будні).

На наведеному графіку можна бачити, що приблизно в районі 250-го дня року загальна кількість повідомлень щодо кризової проблематики почала різко збільшуватися (посилилася парламентська криза).

Існує декілька підходів до видалення періодичної складової в досліджуваному числовому ряду. Найпростіший з них - метод зваженого «ковзного середнього». У основі цього методу згладжування лежить принцип, який полягає в тому, що розкид середнього з N членів часового ряду характеризується величиною дисперсії, рівною S^2 / N, де S^2 – дисперсія вихідного ряду.

Пророднo, поведінку значень наведеного вище числового ряду найкраще виконувати на проміжках, що визначаютья невеликими вікнами спостережень (у даному випадку вікно спостереження вибиралося рівним 7 – числу днів тижня), і обчислювати значення нового ряду «згладжених» величин, визначуваних таким чином:

$$S_t = \frac{1}{7}\sum_{i=t-3}^{t+3} X_i.$$

Очевидно, змінюючи i від $t-3$ до $t+3$, відбувається своєрідне «ковзання» по осі часу, відповідно, застосований метод називається методом «ковзного середнього».

Слід відмітити, що ділення на ширину вікна спостереження, дозволяє при необхідності переходити до змінних вікон спостереження, що особливо актуально в граничних точках. На рис. 3.2 приведений графік «згладженого» ряду, відповідного, даному початковому.

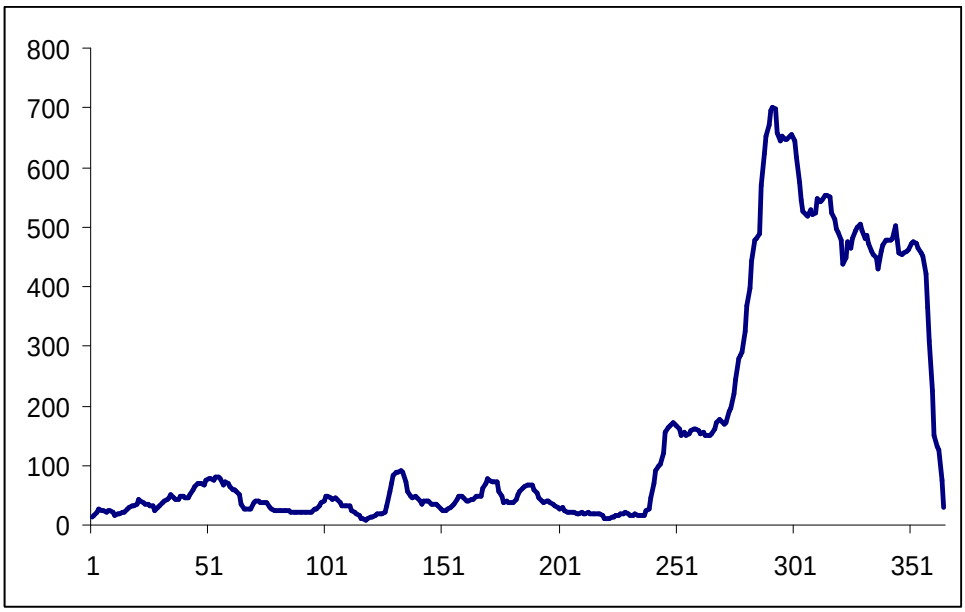


Рис. 3.2. «Згладжений» ряд, який відповідає вихідному

Різновидом ковзного середнього, яке ми розглянемо, є експонентне ковзне середнє (*exponential moving average, EMA*). Його можна розуміти як зважене ковзне середнє, у якого вага зменшуються експонентно з віддаленістю значення ряду за часом, від поточного. Такий розподіл дозволяє зосередитися при аналізі на поточних даних і не пропустити важливі сигнали. Треба відмітити, що крива експонентного ковзного середнього зазвичай краще апроксимує графік.

Математична формула для розрахунку експонентного ковзного середнього є рекурсивною і при значенні коефіцієнта згладжування рівного α має вигляд:

$$E(i) = \alpha X(i) + (1-\alpha)E(i-1),$$

де $X(i)$ – значення ряду спостережень в точці i, $E(i-1)$ – значення експонентного ковзного середнього, розрахованого для попередньої точки ряду, α – регулюючий коефіцієнт. Початкове значення $E(1)$ приймається рівним $X(1)$. Чим більше значення регулюючого коефіцієнта α, тим краще крива експоненціального ковзаючого середнього апроксимує графік, оскільки більше значення надається поточним значенням. Справедливо і зворотне твердження, що для маленьких значень регулюючого коефіцієнта α більше значення надається минулим періодам. Залежно від характеру початкової залежності використовуються різні значення коефіцієнта. На практиці часто використовується значення *2/3*.

Криві експонентного ковзного середнього трактуються так само, як і криві простого ковзаючого середнього. В процесі аналізу слід знати, яке використовується значення регулюючого коефіцієнта. Криві експоненціального ковзного середнього швидше реагують на зміну значень початкового ряду при більшому значенні такого коефіцієнта, оскільки надають більші ваги поточному періоду. На рис. 3.3 наведено діаграму відхилень відповідних значень експонентного ковзного середнього від значень вихідного ряду у залежності від значень регулюючого коефіцієнта α. Більш світлі значення на діаграмі відповідають більшим відхиленням.

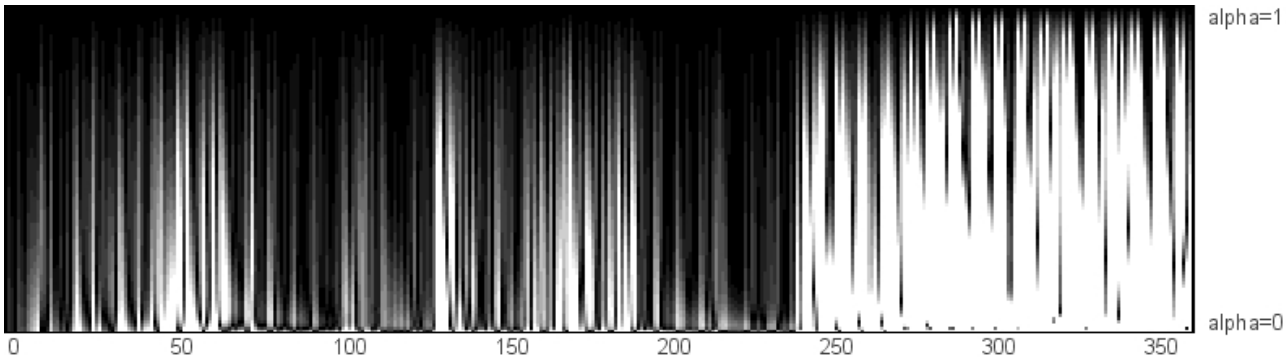


Рис. 3.3. Діаграма відхилень значень експонентного ковзного середнього від значень вихідного ряду

Наведена діаграма, наприклад, дозволяє виявляти монотонні участки ряду, який досліджується, періоди різких коливань.

Криві експонентного ковзного середнього часто використовуються при короткочасному аналізі, оскільки вони дозволяють відловлювати швидкі зміни. Для порівняння, криві простого ковзаючого середнього, навпаки, використовуються в довгостроковому аналізі, оскільки добре показують довгострокові тенденції.

3.2. Самоподібність інформаційних потоків

«Самоподібність являє собою поняття, яке поєднує фрактали, хаос та ступеневі закони. Самоподібність або інваріантність відносно змін масштабу або розміру являє собою відмітну рису багатьох законів природи та незліченних явищ в оточуючому нас світі. Самоподібність є у дійсності однією з вирішальних симетрій, яка формує наш всесвіт та оказує вплив на наші спроби її зрозуміти» [38]

Явище, що має властивість самоподібності, виглядає однаково або однаково себе поводить при його розгляді з різним ступенем «збільшення» або у різному масштабі. Масштабуючої величиною може бути простір (довжина, ширина) або час. У даному розгляді розглядаються, зокрема, часові ряди, які демонструють властивість самоподібності.

Властивості самоподібності фрагментів інформаційного простору наочно демонструє, наприклад, новий інтерфейс, представлений на веб-сайті служби News Is Free (http://newsisfree.com) у режимі бета-тестування. На цьому сайті відображається стан інформаційного простору у вигляді посилань на джерела й окремі повідомлення. При цьому враховується два основних параметри відображення – ранг популярності й оперативність інформації. Укрупнене представлення окремих джерел і/або документів – найбільш популярних і актуальних – наведено на рис. 3.4. Коли граничний ранг популярності й оперативності інформації підвищується, її дроблення

вже не дозволяє без особливих зусиль ідентифікувати окремі документи (Рис. 3.5).

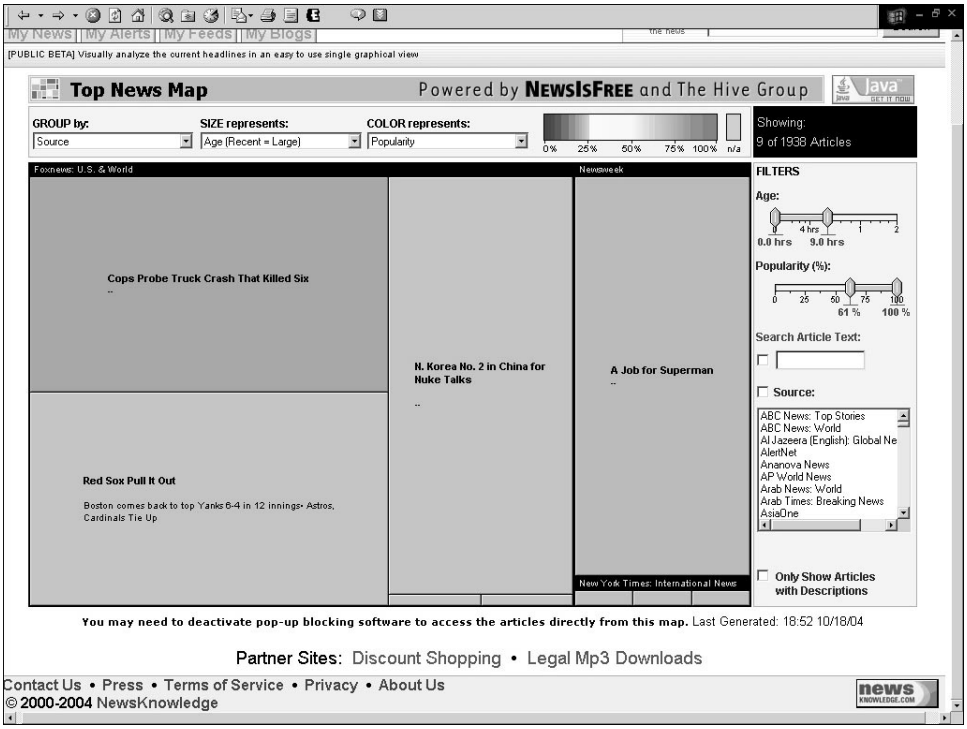


Рис. 3.4. Кластер актуальних документів з популярних видань

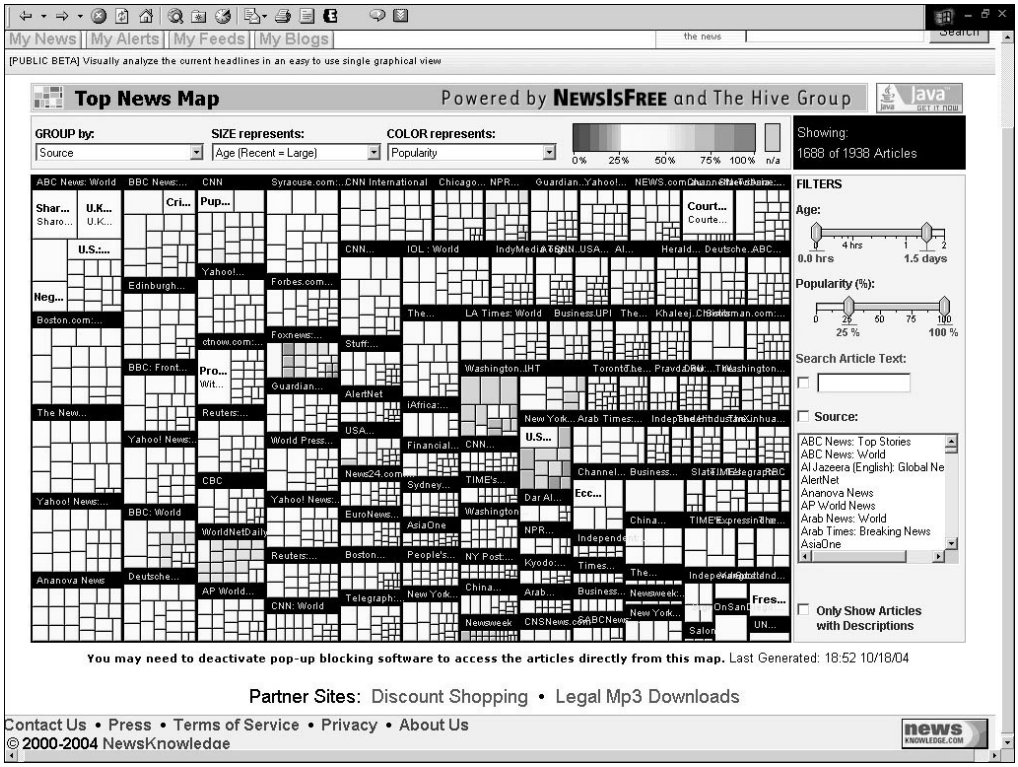


Рис. 3.5. Кластер неоперативних документів із джерел низької популярності

Чітке визначення самоподобного стохастичного процесу засновується на прямому масштабуванні неперировної змінної часу. Стохастичний процес $X(t)$ є стохастично самоподібним з параметром H $(0,5 \le H \le 1)$, якщо для будь-якого дійсного значення $a > 0$ процес $a^{-H} X(at)$ має ті ж самі статистичні характеристики, що й сам процес $X(t)$. Це твердження можна виразити трьома наступними умовами:

– середнє:

$$E\left[X(t)\right] = \frac{E\left[X(at)\right]}{a^H};$$

– дисперсія:

$$\sigma\left[X(t)\right] = \frac{\sigma\left[X(at)\right]}{a^{2H}};$$

– автокореляція (див п. 3.3):

$$K\left[X(t), X(s)\right] = \frac{K\left[aX(t), aX(s)\right]}{a^{2H}}.$$

Параметр H, який має назву параметра Херста (*Hurst parametr*) або параметром сомоподібності (*self-similarity parametr*) , являє собою ключову міру самоподібності. Точніше, H являє собою міру стійкості статистичного явища, або міру дії довгострокової залежності статистичного процесую Значення $H = 0,5$ вказує на відсутність довгострокової залежності. Чим ближче значення H до 1, тим вище ступень стійкості довгострокової залежності.

Розглянемо для прикладу процес броуновкого руху $B(t)$ та доведемо його самоподібністьз параметром $H = 0,5$ у відповідності з наведеним вище визначенням. Розглянемо три умови самоподібності:

– за визначенням, $E\left[B(t)\right]$=0. Тому $E\left[B(t)\right] = E\left[B(at)\right] / a^{0,5}$,

що задовольняє першій вимозі;

– відомо, що дисперсія $\sigma\left[B(t)\right]$ дорівнює t, тому $\sigma\left[B(at)\right] = at = a\sigma\left[B(t)\right]$,

що задовольняє другій вимозі;

– загальновідомо, що автокореляція $K\left[B(t),B(s)\right]=\min\left[t,s\right]$. Звідсі:

$$K\left[B(at),B(as)\right]=\min\left[at,as\right]=a\min\left[t,s\right]=aK\left[B(t),B(s)\right],$$

що задовольняє третій вимозі.

Далі розглянемо випадок стохастичного процесу, визначеного у дискретних точках часу, так що стохастичний процес $X(t)$ визначається як $\left\{x_t,\ t=0,1,2,...\right\}$. Для таких процесів визначаються $m-$агреговані часові серії $\left\{x_k^{(m)},\ k=0,1,2,...\right\}$, підсумовуючи вихідні серії за неперекриваючими сусідніми блоками розміру m. Це може бути виражено таким чином:

$$x_k^{(m)}=\frac{1}{m}\sum_{i=km-m+1}^{km}x_i\ .$$

Агреговані часові серії можна розглядати як метод стиску часової шкали. При цьому $x^{(1)}$ може вважатися максимальним збільшенням або найвищою розподільною здатністю для цієї часової серії. Процес $x^{(5)}$, наприклад, являє собою той самий процес, зменшений у три рази. Якщо статистичні характеристики процеса зберігаються при стисканні, то можа вважати що йдеться про самоподібний процес.

Таким чином, можна запропонувати функціональне визначення самоподібності, а саме: процес x називається у точності самоподобним (*exactly self-similar*) з параметром β $(0<\beta<1)$, якщо для всіх $m=1,2,...$ виконується:

– для дисперсії:

$$\sigma\left[x^{(m)}\right]=\frac{\sigma\left[x\right]}{m^{\beta}};$$

– автокореляція (див п. 3.3):

$$K\left[x^{(m)},k\right]=K\left[x,k\right].$$

Можна показати, що параметр β пов'язується з визначеним раніш параметром Херста як $H=1-(\beta/2)$. Для стаціонарного ергодичного процесу

$\beta = 1$, а середня дисперсія за часом прямує до нуля зі швидкістю $1/m$. Для самоподібного процесу середня дисперсія часу затухає повільніше.

Наведене вище визначення дозволяє реалізувати найпростіший алгоритм визначення того, чи є дана часова серія самоподібною.

Якщо прологарифмувати наведену вище формулу для дисперсії, отримуємо:

$$\log\left(\sigma\left[x^{(m)}\right]\right) = \log\left(\sigma[x]\right) - \beta \log m.$$

Оскільки $\log\left(\sigma[x]\right)$ є монотоною константою, яка не залежіть від m, то графік залежності $\log\left(\sigma\left[x^{(m)}\right]\right)$ від m в логарифмічному масштабі буде являти пряму лінію з нахилом, який дорівнює $-\beta$. Графік можливо побудувати (звичайно, для фактичних даних слід використовувати вибіркову дисперсію замість теоретичної), якщо згенерувати агрегований процес на різних рівнях агрегації m, а після цього обчислити дисперсію. Зазвичай часові ряди, що утворюються з обсягів тематичних інформаційних потоків, лягають на пряму лінію з негативним нахилом. У цих випадках зазвичай визначають значення параметра H.

Іншою концепцією, пов'язаною з самоподібністю, є повільно затухаючі розподіли, або розподіли з «тяжкими хвостами» (heavy-tailed distributions). Повільно затухаючі розподіли можуть використовуватися для представлення щільності ймовірностей, які описують, наприклад, обсяги даних в інформаційних потоках. Відомо, що розподіл випадкової змінної X повільно затухає, якщо:

$$1 - F(x) = \Pr[X > x] \sim \frac{1}{x^{\alpha}} \text{ при } x \to \infty, \ 0 < \alpha.$$

У цілому, випадкова змінна с повільно затухаючим розподілом має нескінченну дисперсію та, можливо, нескінченним середнім значенням. Випадкова змінна з повільно затухаючим розподілом може приймати дуже великі значення з ймовірністю, якою неможливо знехтувати.

Самим простим повільно затухаючим розподілом є розподіл Парето з параметрами k та α $(k,\ \alpha<0)$ та наступними статистичними показниками:

$$f(x)=F(x)=0 \quad (x\le k);$$

$$f(x)=\frac{\alpha}{k}\left(\frac{k}{x}\right)^{\alpha+1};$$

$$F(x)=1-\left(\frac{k}{x}\right)^{\alpha} \quad (x>k;\ \ \alpha>0);$$

$$E[x]=\frac{\alpha}{\alpha-1}k \quad (\alpha>1).$$

Самоподібність інформаційного простору виражається, насамперед у тому, що при майже обвальному рості цього простору в останні десятиліття, гіперболічні частотні і рангові розподіли, одержувані в таких змістовних розрізах, як, наприклад, джерела й автори документів, практично не змінюють своєї форми. Закономірності, відкриті такими вченими, як Зіпф, Брэдфорд, Лоткі та інші, повною мірою свідчать про самоподібність інформаційного простору. З іншого боку, самоподібність (скейлінг) можна розглядати і як наслідок загальних структурних закономірностей інформаційного простору.

Як показано в [16, 58], для послідовності повідомлень тематичних інформаційних потоків у відповідності до скейлінгового принципу, кількість повідомлень, резонансів на події реального миру пропорційна деякому ступеню кількості джерел інформації (кластерів) та ітераційно триває протягом певного часу. Так само, як і у традиційних наукових комунікаціях, множина повідомлень в Інтернет з однієї тематики в часі являє собою динамічну кластерну систему, що виникає в результаті ітераційних процесів. Цей процес породжується републікаціями, прямим або спільним цитуванням, різними публікаціями - відбиттями тих самих подій реального світу, прямими посиланнями тощо.

Якщо розглядати інформаційні потоки як ряди публікацій протягом часу, то найбільш цікавим у рамках даного дослідження виявляється наявність

таких властивостей, як самоподібність (масштабна інваріантість, скейлинг), стійкі взаємні кореляції. Аналіз самоподібності інформаційних масивів може розглядатися як технологія, призначена для здійснення аналітичних досліджень із елементами прогнозування, придатна до екстраполяції отриманих залежностей.

3.3. Кореляційний аналіз

Одним з перших поширених методів сучасного аналізу часових рядів вимірювань є кореляційний аналіз. Важливими властивостями автокореляційної функції є можливість виявлення гармонічних складових, а також самоподібності вихідного процесу. Зупинимося більш детально на формалізмі кореляційного аналізу.

Якщо позначити через X_t член ряду кількості публікацій (наприклад, кількості електронних повідомлень, що надійшли у день t, $t = 1, \ ..., \ N$), то функція автокореляції для цього ряду X визначається як:

$$F(k) = \frac{1}{N-k} \sum_{t=1}^{N-k} (X_{t+k} - m)(X_t - m),$$

де m – середнє значення ряду X, яке надалі, не обмежуючи спільності, можна важати рівним нулю. Коефіцієнти кореляції для рядів вимірів X довжиною N розраховуються за формулою:

$$R(k) = \frac{F(k)}{\sigma^2},$$

де $F(k)$ – функція автокореляції; σ^2 - дисперсія.

Відомо, що функція автокореляції має ту властивість, що якщо прихована періодична складова існує, то її значення асимптотично наближається до квадрату середнього значення вихідного ряду. Крім того, функція автокореляції періодичного ряду також є періодичною, містить основну частоту та гармоніки.

Якщо вихідний ряд періодичний, тобто може бути представленій як:

$$X_t = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega t + \theta_n),$$

то його функція автокореляції буде дорівнювати:

$$F(k) = \frac{a_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty} a_n^2 \cos n\omega k.$$

Тобто ряд вимірів X є сумішшю деякої змістовної складової N та синусоїдального сигналу S, отже функція автокореляції ряду X містить явно виражену періодичну складову [51], частоту і гармоніки, про те без фазових кутів θ_n.

Розглянемо числовий ряд X, який є сумою деякої змістовної складової N і синусоїдальною сигналу S:

$$X_t = N_t + S_t.$$

Знайдемо функцію автокореляції для цього ряду (значення приведені до середнього $m = 0$):

$$F(k) = \tfrac{1}{N-k}\sum_{t=1}^{N-k} X_{k+t} X_t =$$

$$= \tfrac{1}{N-k}\sum_{t=1}^{N-k} (N_{k+t} + S_{k+t})(N_t + S_t) =$$

$$= \tfrac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t} N_t + \tfrac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t} S_t + \tfrac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t} S_t + \tfrac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t} N_t.$$

Очевидно, перший доданок - це функція неперіодична, яка асимптотично спрямовується до нуля. Оскільки взаємна кореляція між N та S відсутня, то третій і четвертий доданок також прямують до нуля. Таким чином, найзначніший ненульовий внесок робить другий доданок – автокореляція сигналу S. Звідсі функція автокореляції ряду X залишається періодичною.

Для експериментального підтвердження розглянутої гіпотези була сгенерована послідовність, яка за своєю природою нагадує реальний інформаційний потік. Передбачалося, що щоденна кількість повідомлень в мережі росте за експонентним законом (з дуже невеликим значенням експонентного ступеня), і на цю кількість накладаються коливання, пов'язані

з тижневою циклічністю інформаційних джерел. Також приймається до уваги елемент випадковості, виражений відповідними відхиленнями.

Для отримання відповідного часового ряду були розглянуті значення функції:

$$y = ae^{0.001x} + \sin(\pi x / 7 + a),$$

яка реалізує просту модель інформаційного потоку – експонента відповідає за зростання кількості публікацій в часі (загальна тенденція), синус – за тижневу періодичність, параметр a – за випадкові відхилення. Кількість публікацій не може бути від'ємним числом.

На рис. 3.6 зображений графік моделі (вісь абсцис – змінна x – день, вісь ординат – змінна y – кількість публікацій).

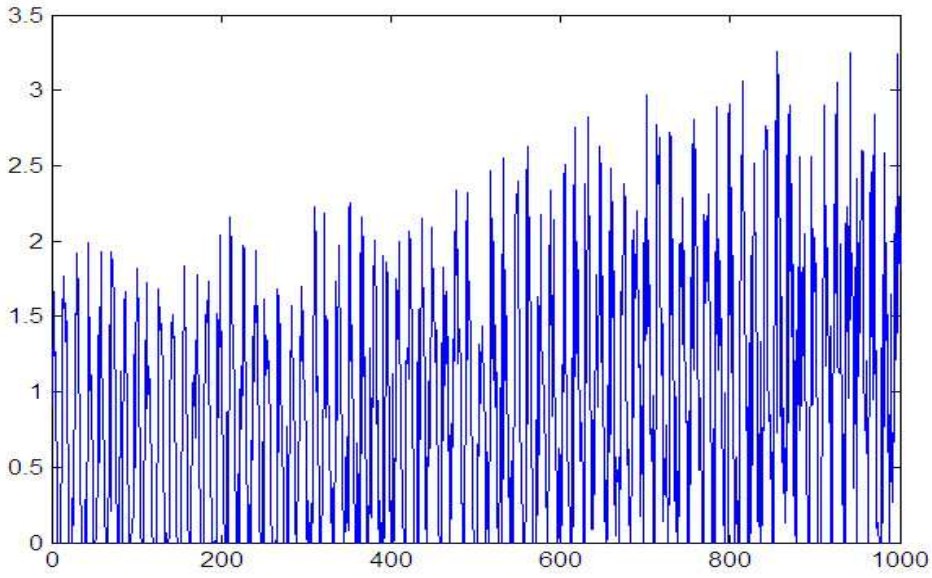


Рис. 3.6. Модель потоку з експонентним зростанням

На рис. 3.7 приведений графік значень коефіцієнтів кореляції (вісь абсцис – змінна *k*, вісь ординат – коефіцієнт кореляції *R(k)*.

Графічне представлення коефіцієнта автокореляції для ряду спостережень, що відповідає динаміці розглянутого віще тематічного інформаційного потоку свідчить про незмінність кореляційних властівостей за днями тіжня (рис. 3.8), а тренд – про можливу самоподібність вихідного часового ряду.

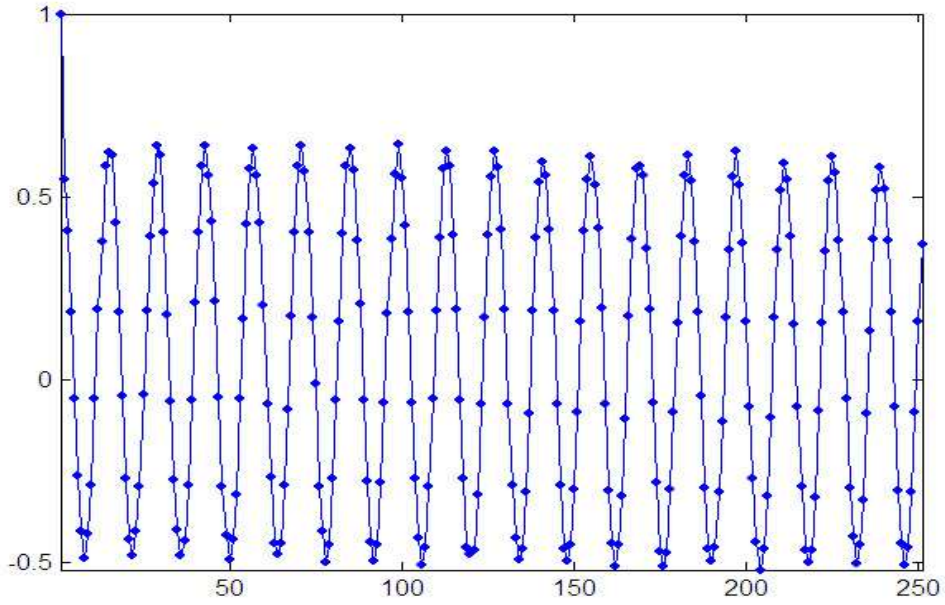


Рис. 3.7. Значення коефіцієнтів кореляції моделі

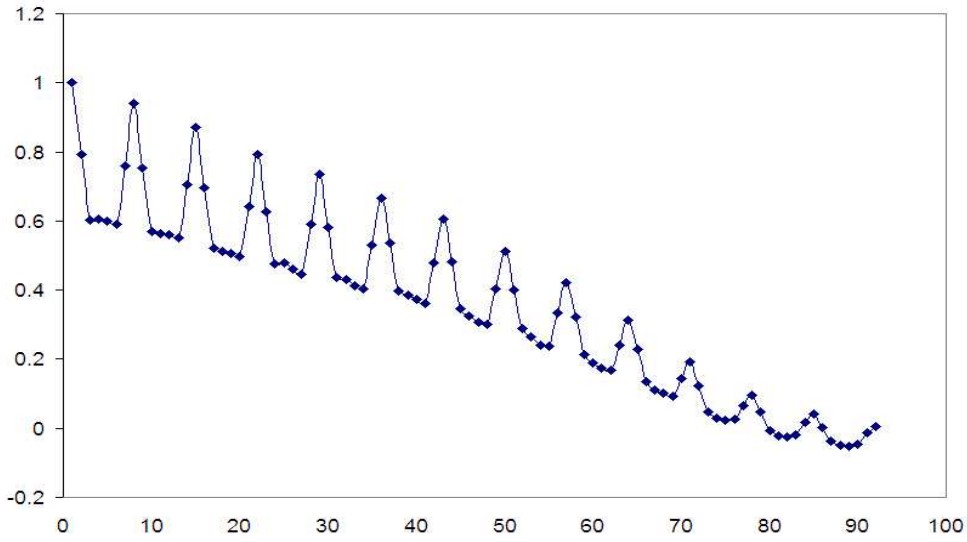


Рис. 3.8. Коефіцієнти кореляції ряду спостережень $R(k)$ (вісь ординат) залежно від k (вісь абсцис)

Кореляційна функція «згладженого» ряду не містить явно виражених гармонік і підтверджує припущення про те, що основна періодична складова даного ряду відповідає 7-денному (тижневому) циклу. Разом з тим, коефіцієнти кореляції ряду спостережень, усередненого за тижнями, апроксимуються гіперболічною функцією, що свідчить про довгострокову залежність початкового ряду. Взаємна залежність членів згладженого ряду

без урахування циклічної складової також підтверджується порівнянням з «перемішаним» рядом.

3.3. Дисперсійний аналіз

Метод DFA (Detrended fluctuation analysus) найчастіше використається для виявлення статистичної самоподібності рядів вимірів [56, 30]. Цей метод є варіантом дисперсійного аналізу одномірних випадкових блукань та дозволяє досліджувати ефекти тривалих кореляцій у рядах, що досліджуються. У рамках алгоритму DFA аналізується середньоквадратична помилка лінійної апроксимації в залежності від розміру апроксимаційного околу (вікна спостереження). Нехай є ряд вимірів x_t, $t \in 1,...,N$. Позначимо середнє значення цього ряду вимірів: $\langle x \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. З вихідного ряду будується ряд накопичення:

$$X_t = \sum_{k=1}^{t} \left(x_k - \langle x \rangle \right).$$

Після цього ряд X_t розділяється на часові вікна довжиною L, будується лінійна апроксимація ($L_{j,L}$) за значеннями $X_{k,j,L}$ усередині кожного вікна (у свою чергу, $X_{j,L}$ - підмножина X_t, $j = 1,...,J$, $J = N / L$ - кількість вікон спостереження) і розраховується відхилення точок ряду накопичення від лінійної апроксимації:

$$E(j,L) = \sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(X_{k,j,L} - L_{k,j,L} \right)^2} = \sqrt{\frac{1}{L}\sum_{k=1}^{L} |\Delta_{k,j,L}|^2},$$

де $L_{k,j,L}$ - значення локальної лінійної апроксимації в крапці $t = (j-1)L + k$.

Тут $|\Delta_{k,j,L}|$ - абсолютне відхилення елемента $X_{k,j,L}$ від локальної лінійної апроксимації.

Далі обчислюється середнє значення:

$$F(L) = \frac{1}{J}\sum_{j=1}^{J} E(j,L),$$

після чого, у випадку $F(L) \propto L^{\alpha}$, де α деяка константа, робляться висновки щодо наявності статистичної самоподібності та характер поведінки досліджуваного ряду вимірів.

Цей метод було застосовано до ряду значень кількості публікацій, отриманих за представленим вище запитом. На рис. 3.9 представлена залежність середньоквадратичної помилки апроксимації від довжини околу апроксимації в подвійному логарифмічному масштабі.

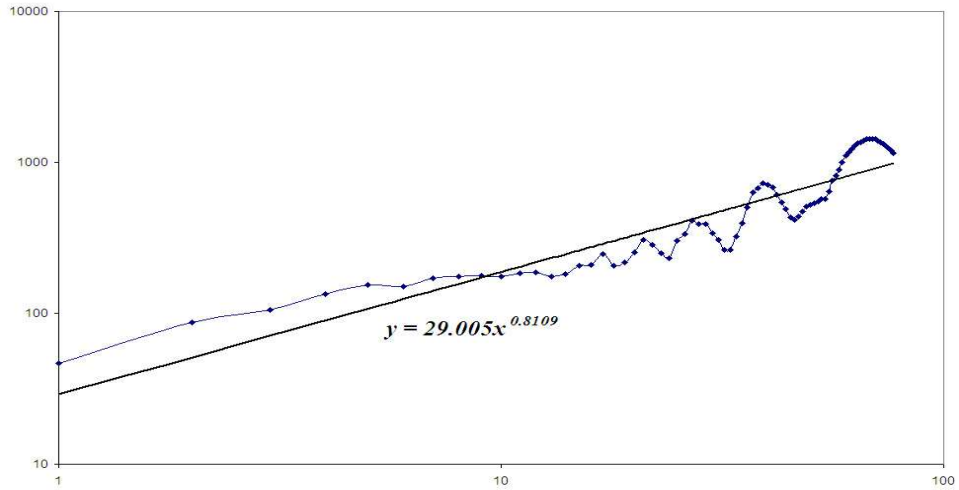


Рис. 3.9. Залежність середньоквадратичної помилки лінійної апроксимації D від довжини вікна спостереження k

Близькість залежності $D(k)$ до лінійної ще раз підтверджує наявність локального скейлінгу часового ряду впродовж другого півріччя 2008 року.

Для вивчення поведінки часових рядів прийнято використовувати ще один показник – індекс розкиду дисперсії (IDC), так званий чинник Фано (U. Fano). Ця величина визначається як відношення дисперсії кількості подій (у нашому випадку – кількості публікацій) на заданому вікні спостережень k до відповідного математичного очікування:

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

Для собіподібних процесів виконується співвідношення:

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

де C та H – константи.

На рис. 3.10 приведений графік значень $F(k)$ у логарифмічному масштабі, при цьому $C \approx 6.8$ та $H \approx 0.65$.

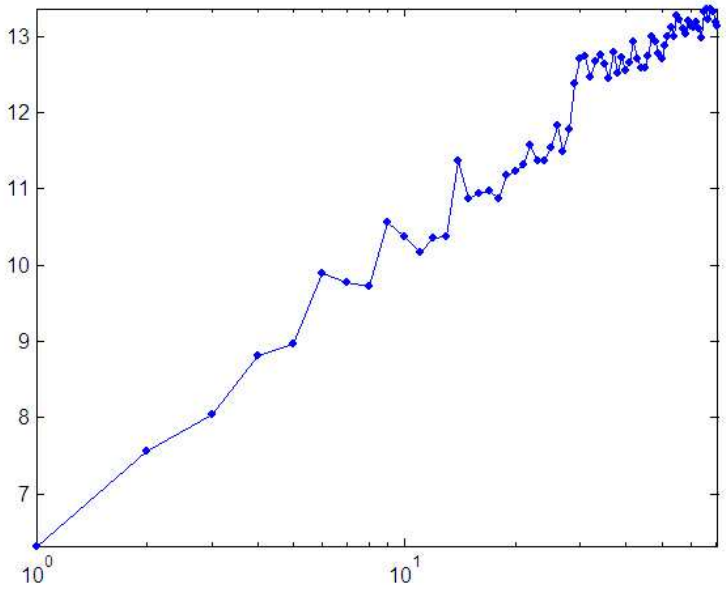


Рис. 3.10. Залежність чинника Фано від ширини вікна спостережень

Задачі виявлення та візуалізації трендів, виявлення гармонійних складових, трендів, локальних особливостей часових рядів, фільтрації шуму сьогодні вирішуються методами фрактального, вейвлет- і Фур'є-аналізу. З метою візуалізації та аналізу часових рядів, пов’язаних з публікаціями в інформаційному просторі мережі Інтернет, розроблено ΔL-метод дисперсійного аналізу, призначений для аналізу та візуалізації стану часових рядів інтенсивності публікацій за визначеною тематикою [22].

Як і у методі DFA, розглянемо поведінку відхилення точок ряду накопичення від лінійної апроксимації (але у цьому разі абсолютне значення) $|\Delta_{k,j,L}|$. Побудова відповідних діаграм значень $|\Delta_{k,j,L}|$, що залежать фактично від двох параметрів - L і $t = (j-1)L + k$ названо ΔL-методом візуалізації. Така візуалізація у вигляді «рельєфної» діаграми являє собою певний інтерес для вивчення особливостей процесів, що відповідають вихідним рядам вимірів.

ΔL-метод виявляється досить ефективним для виявлення гармонійних складових досліджуваного ряду. На рис. 3.11 показана ΔL-діаграма ряду, що

відповідає синусоїді ($y(i) = \sin(i\pi/7),\ \ i = 1,...,366$). Застосування ΔL-методу до ряду, складеному з кількості публікацій, зібраних системою InfoStream з Інтернет без урахування тематичного розподілу, має явно виражену гармонійну складову (загальна кількість публікацій залежить від дня тижня), що можна бачити на рис. 3.12. Крім того, на цій діаграмі помітні відхилення від загальної динаміки об'ємів публікацій у святкові дні.

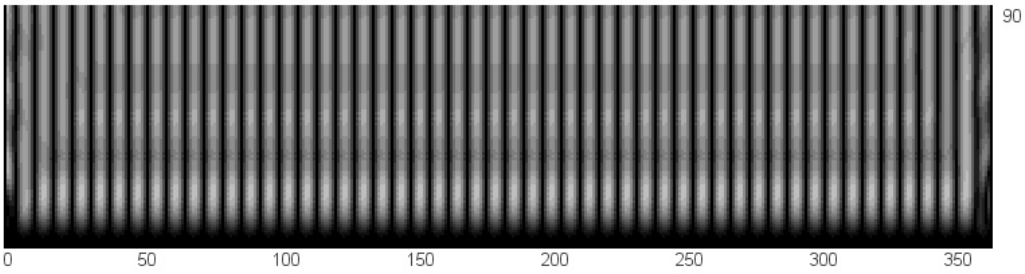


Рис. 3.11. ΔL-діаграма синусоїди

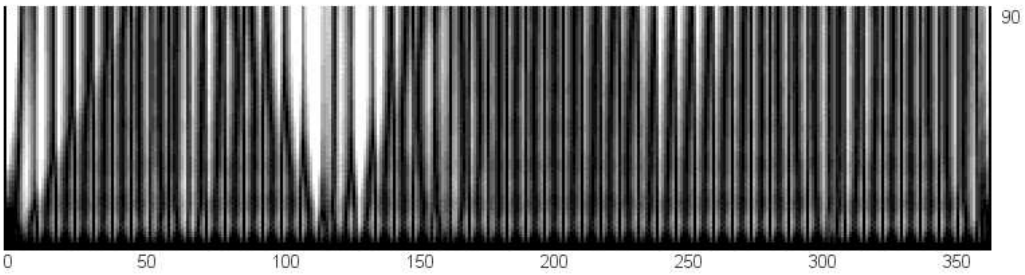


Рис. 3.12. ΔL-діаграма ряду з кількості публікацій, що збиралися щодоби системою InfoStream у 2008 році

«Рельєфні діаграми», одержувані в результаті запропонованого ΔL-методу (більш світлі тони відповідають більшим значенням $|\Delta_{k,j,L}|$), нагадують скейлограми, одержані в результаті безперервних вейвлет-перетворень. Варто звернути увагу на те, що темні смуги в центрі багатьох областей світлого зафарбування свідчать про «стабілізацію» більших значень розглянутого ряду на високому рівні.

ΔL-метод застосовується для реальних часових рядів, наприклад тих, що відбивають інтенсивність публікацій даної тематики в Інтернеті. На рис. 3.13 наведено ΔL-діаграму для розглянутого вище часового ряду з кількості публікацій повідомлень за добу по вибраній тематиці в мережі Інтернет протягом року.

На рис. 3.14 наведено ΔL-діаграма готівкового курсу долара у гривнях протягом 2008 року. Ще наочніше, ніж у випадку застосування вейвлет-аналізу можна переконатися у тому, що найзначніші відхилення на діаграмі у цьому випадку настають з деякою часовою затримкою у порівняння з діаграмою публікаціями щодо кризової тематики.

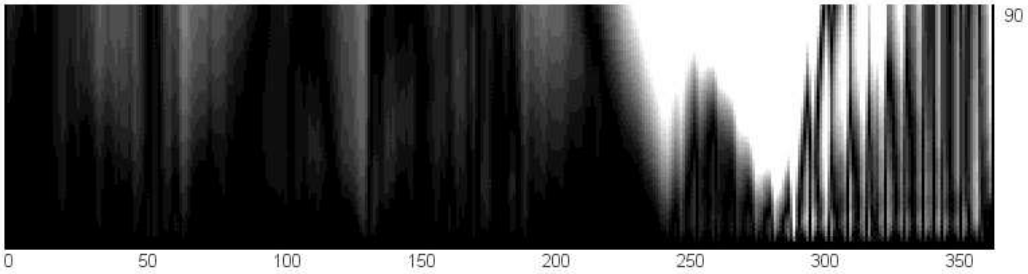


Рис. 3.13. ΔL-діаграма часового ряду інтенсивності тематичних публікацій (вісь абсцис - дні року, вісь ординат - величина вікна вимірів)

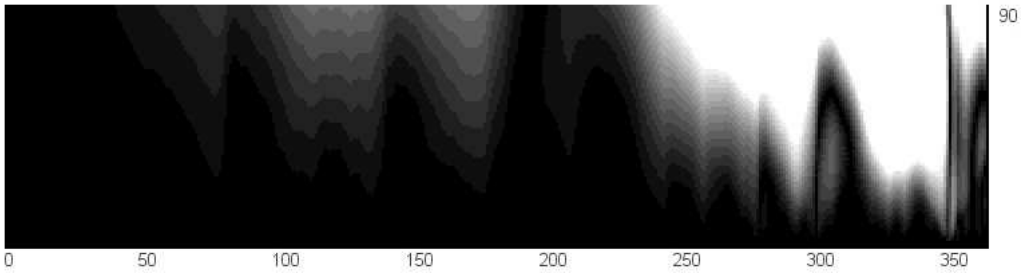


Рис. 3.14. ΔL-діаграма часового ряду курсів готівкових курсів долара у гривнях (вісь абсцис - дні року, вісь ординат - величина вікна вимірів)

Запропонований метод візуалізації абсолютних відхилень ΔL, як і метод вейвлет-перетворень, дозволяє (і як показано на прикладі – не гірше) виявляти одиничні й нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу. Слід зазначити, що метод вейвлет-перетворень може застосовуватися з використанням різноманітних вейвлетів. У випадку застосування ΔL-методу не потрібно вирішувати складну задачу вибору й обґрунтування застосування відповідного вейвлету; на відміну від методів фрактального аналізу запропонований підхід не вимагає значних об'ємів точок ряду вимірів. Цей метод досить простий у програмній реалізації й базується на такому потужній теоретичній основі як DFA, виявився досить ефективним при аналізі часових рядів.

3.4. Фрактальний аналіз

Для дослідження часових рядів обсягів повідомлень у тематичних інформаційних потоках сьогодні усе ширше використовується теорія фракталів, традиційна область застосування якої - фрактальна геометрія, обробка зображень та інше. Разом з тим часові ряди, породжувані тематичними інформаційними потоками, також мають фрактальні властивості і можуть розглядатися як стохастичні фрактали. Цей підхід розширює область застосування теорії фракталів на інформаційні потоки, динаміка яких описується засобами теорії випадкових процесів.

З іншого боку, теорія фракталів розглядається як підхід до статистичного дослідження, що дозволяє одержувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їхньої внутрішньої структури і зв'язків.

На думку С.А. Іванова, всі основні закони наукової комунікації, такі як закони Парето, Лотки, Бредфорда, Зіпфа, можуть бути узагальнені саме в рамках теорії стохастичних фракталів. Теорія фракталів [33] широко застосовується як підхід до дослідження рядів вимірів, який дозволяє одержувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їхньої внутрішньої структури.

Термін фрактал утворений від латинського слова fractus – дробовий, що складається з фрагментів. Він був запропонований Бенуа Мандельбротом [28] у 1975 році для позначення нерегулярних самоподібних математичних структур. Популярна сьогодні фрактальна геометрія одержала свою назву лише в 1977 році завдяки книзі Мандельброта «The Fractal Geometry of Nature». У його роботах були використані наукові результати багатьох дослідників, що працювали в цій самій галузі (насамперед, Пуанкаре, Кантора, Хаусдорфа).

Основне визначення фракталу, надане Мандельбротом, звучить так: "Фракталом зветься структура, що містить частини, які у якомусь розумінні подібні цілому". Однак самоподібність – це лише необхідна, але не достатня властивість фракталів. Головна особливість фракталів полягає у тому, що їхня розмірність не відповідає звичним геометричним уявам. Використовується спеціальне поняття фрактальної розмірності, введене Хаусдорфом та Безиковичем, визначене наступним чином. Нехай є множина G у просторі $\mathbb{R}^n$. Разіб'ємо простір $\mathbb{R}^n$ на n-вимірні куби з довжиною ребра δ та позначимо кількість кубів, необхідних для покриття множини G, через $N(\delta)$. Тоді величина розмірності Хаусдорфа-Безиковича D повинна задовольняти такій вимозі:

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

Це визначення можна спростити, зробивши його більш прийнятним для практичного застосування. Як бачимо, при $\delta \approx 0$, воно еквівалентне:

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

На даний час інформаційний простір у цілому, через його обсяги і динаміку змін, прийнято розглядати як стохастичний. Сьогодні в моделюванні інформаційного простору все частіше використовується фрактальний підхід, що базується на властивості його самоподібності, тобто збереження внутрішньої структури множин при змінах їх розмірів або масштабів при їх розгляді ззовні.

Застосування теорії фракталів при аналізі інформаційного простору дозволяє з загальної позиції глянути на емпіричні закони, що складають теоретичні основи інформатики. В інформаційному просторі виникають, ростуть і формуються кластери документів, що відбивають сучасні процеси комунікації [15].

Сьогодні фрактальні особливості WWW уже досить широко використовуються при вирішенні таких задач, як оптимізація механізмів

сканування, аналіз і прогноз розвитку інформаційних ресурсів, побудова нових Web-сервісів. Дослідження фрактальних властивостей інформаційних потоків покажемо на прикладі аналізу електронних ЗМІ до проблематики інтеграції України з НАТО. Для проведення необхідних досліджень автором було застосовано бази даних сервера інформаційної підтримки прийняття рішень, встановленого у Національному центрі з питань євроатлантичної інтеграції України [27].

При цьому на сервері реалізовано повномасштабне інформаційне сховище, що враховує особливості євроатлантичної проблематики, накопичує та надійно зберігає інформацію для використання в аналітичній роботі. Комплекс контент-моніторингу виконує основну "чорнову" роботу зі збору інформації з мережі Інтернет та забезпечує створення та постійне поповнення документального сховища оперативними повідомленнями, які застосовувалися, у тому числі і для проведення статистичних досліджень. Типова задача комплексу контент-моніторингу - це побудова діаграм динаміки появи понять у часі. На рис. 3.15, наприклад, проілюстровано, як за допомогою цієї діаграми відслідковується поява україномовних повідомлень щодо євроатлантичної інтеграції України впродовж заданого періоду.

У той час, як для традиційних засобів наукової комунікації підходи до статистичних досліджень інформаційних масивів з погляду теорії фракталів були вперше досліджені Ван Рааном [60], що аналізував масиви статей і зв'язки, утворені цитуванням, інформаційні потоки повідомлень із Інтернет до останнього часу не асоціювалися із фракталами. Це пов'язано із проблемами ідентифікації інформаційних потоків як фрактальних множин, а також із труднощами знаходження основ для побудови кластерів - повідомлень у політематичних потоках, що породжують багаторазове цитування.

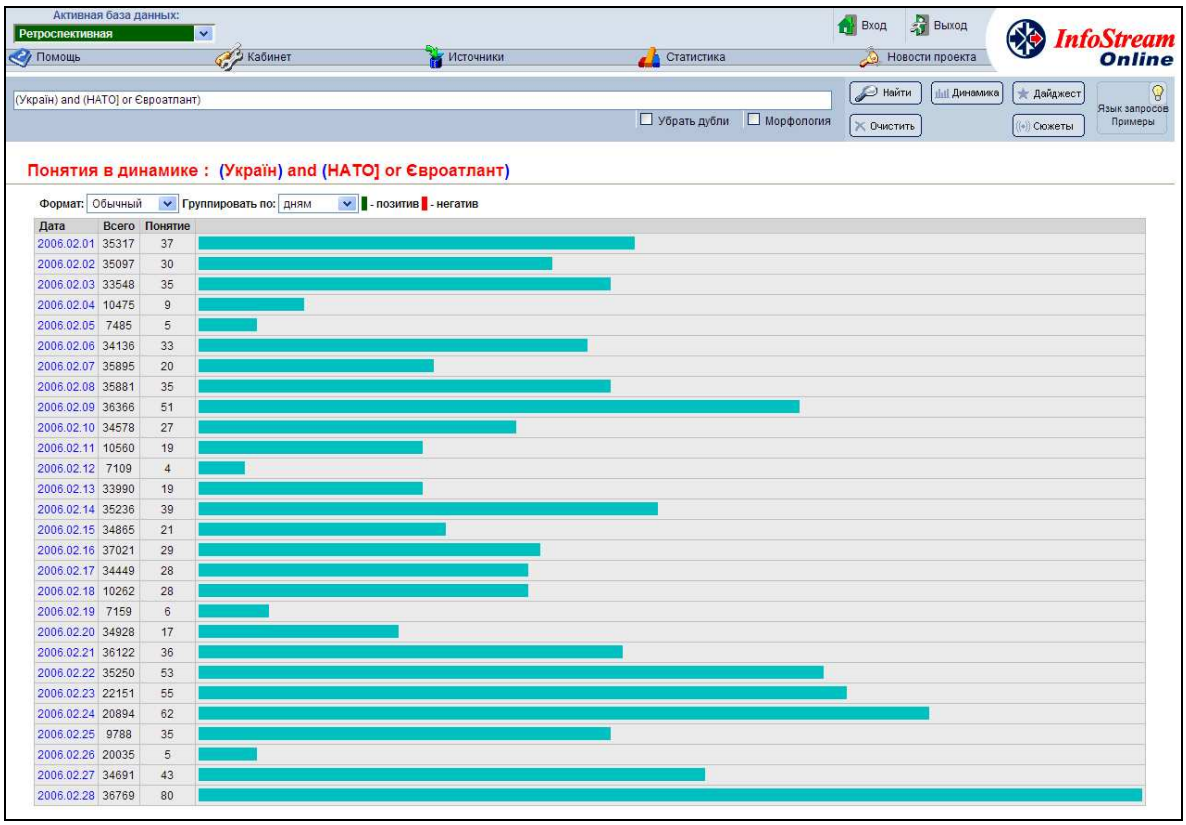


Рис. 3.15. Динаміка появи у часі поняття, визначеного запитом: «(Україн) and (НАТО] or Євроатлант)»

Фрактальна розмірність [15] у кластерній системі, що відповідає тематичним інформаційним потокам, показує ступінь заповнення інформаційного простору повідомлень впродовж певного часу:

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

де $N_{публ}$ – розмір кластерної системи (загальне число електронних публікацій в інформаційному потоці); N_k – число кластерів (тематик або джерел), ρ - фрактальна розмірність інформаційного масиву як кластерного утворення; ε - коефіцієнт масштабування. У наведеному співвідношенні між кількістю повідомлень і кластерів проявляється властивість збереження внутрішньої структури безлічі при зміні масштабів його зовнішнього розгляду.

Найважливішою характеристикою рядів, що мають хаотичну поведінку, є, як відомо, фрактальна розмірність, яка у багатьох випадках може обчислюватися в результаті так званого R/S-аналізу. Загально кажучи, обчислюється не сама фрактальна розмірність, а показних Херста, який

зв’язаний з нею простим співвідношенням. R/S-аналіз базується на аналізі нормованого розкиду - відносини розкиду R значень досліджуваного ряду до середньоквадратичного відхилення S.

З’ясуємо, як обчислюється значення розмаху R. Для тимчасового ряду $F(n)$, $n=1, ..., N$, обчислюється середнє значення:

$$\langle F \rangle_N = \frac{1}{N}\sum_{n=1}^{N} F(n),$$

ряд накопичених значень:

$$X(n,N) = \sum_{i=1}^{n}\left(F(i) - \langle F \rangle_N\right),$$

після чого обчислюється безпосередньо розмах:

$$R(N) = \max_{i\le n\le N} X(n,N) - \min_{i\le n\le N} X(n,N).$$

Г.Е. Херст експериментально виявив, що для часових рядів, що мають властивісті самоподібності, справедливо:

$$R/S = \left(\frac{N}{2}\right)^{H},$$

де H — показник Херста, який для досить широкого класу рядів пов'язаний з хаусдорфовою (фрактальною) розмірністю D постою формулою: $D+H=2$.

Головна умова, при якій показник Херста пов'язаний з фрактальною розмірністю відповідно до наведеної формули, визначена Е. Федером таким чином: «... розглядають клітки, розміри яких малі в порівнянні як із тривалістю процесу, так і з діапазоном зміни функції; тому співвідношення справедливе, коли структура кривої, що описує фрактальну функцію, досліджується з високою розподільною здатністю, тобто в локальній межі». Ще однією важливою умовою є самоафінність функції. Не вдаючись у подробиці, помітимо, що, наприклад, для інформаційних потоків ця властивість інтерпретується як самоподібність, яка виникає в результаті процесів їхнього формування. Можна відзначити, що зазначеними

властивостями відповідають не всі інформаційні потоки, а лише ті, які характеризуються достатньою потужністю та ітеративністю при формуванні.

На рис. 3.16 представлено співвідношення R/S для ряду кількості публікацій за днями 2008 року, що відповідає першому запиту. Очевидно, характер нормованого розмаху різко змінюється в районі 250 дня року, приблизно тоді, коли пролунали перші серйозні заяви на вищому рівні щодо фінансово-економічної кризи. Тобто маємо фактично два різних ряди – з 1 по 250 та з 251 по 366. Як можна бачити, крива нормованого розмаху для другого ряду (рис. 3.17) задовільно апроксимується прямою у подвійному логарифмічному масштабі. Нахил цієї прямої відповідає показнику Херста.

Чисельні значення H характеризують різні типи корельованої динаміки (персистентності). При H = 0,5 спостерігається некорельована поведінка значень ряду, а значення $0{,}5 < H < 1$ відповідають ступеню автокореляції ряду. Як можна бачити, значення Херста для досліджуваного інформаційного потоку відповідає величині ~0,89, що підтверджує припущення щодо самоподібності та ітеративності процесів в інформаційному просторі.

На основі розглянутого прикладу розглянута висока персистентность процесу, що, зокрема, свідчить про загальну тенденцію збільшення публікації з вибраної тематики. Подібний аналіз інформаційних масивів може розглядатися як технологія для здійснення прогнозування.

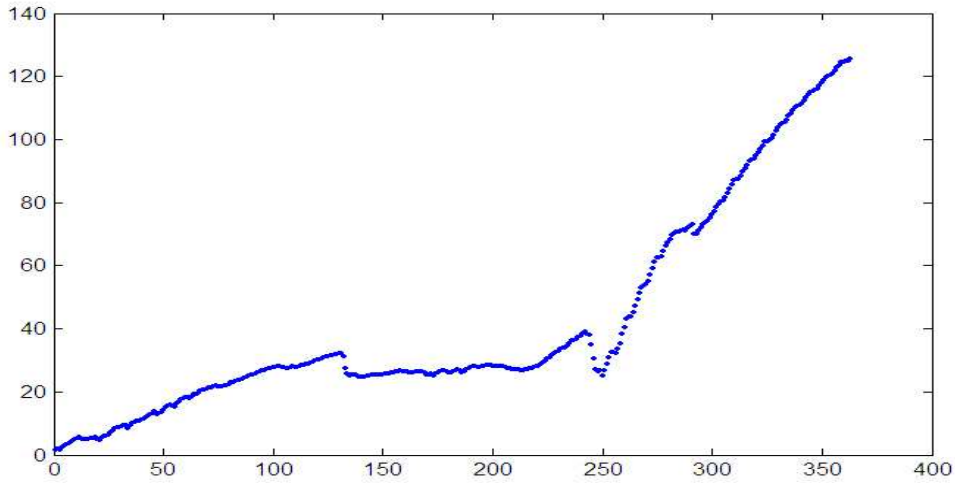


Рис. 3.16. Показник нормованого розкиду для всього періоду спостережень ряду, сформованому за першим запитом

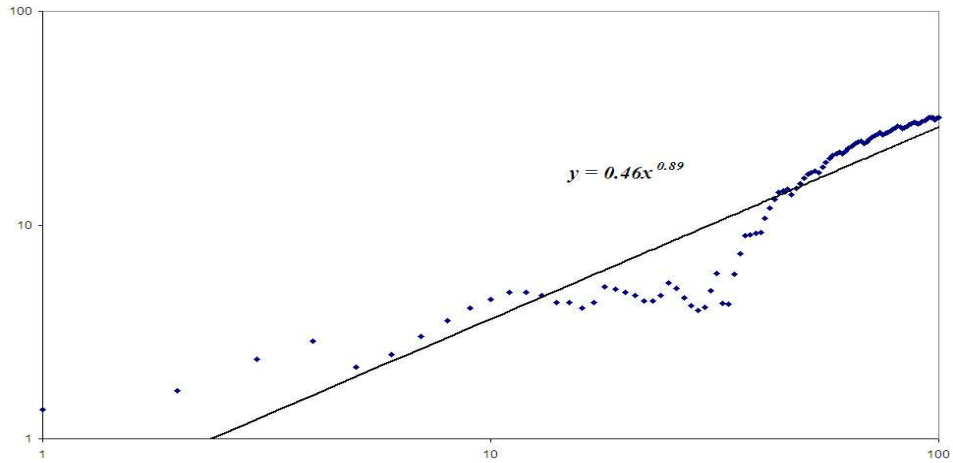


Рис. 3.17. Показник нормованого розкиду в логарифмічній шкалі за останні 120 днів року

3.6. Вейвлет-аналіз

До кола сучасних інструментальних засобів оцінки рядів спостережень відноситься також вейвлет-аналіз [2]. Він особливо ефективний у тих випадках, коли крім загальних спектральних характеристик потрібно виявляти локальні в часі особливості поведінки процесу, який досліджується. Основою вейвлет-аналізу є вейвлет-перетворення, яке є особливим типом лінійного перетворення, базисні функції якого (вейвлети) мають специфічні властивості. Аналіз даних з використанням вейвлет-перетворень є зручним, надійним і потужним інструментом дослідження часових рядів і дозволяє представити результати у наочному вигляді, зручному інтерпретації.

Вейвлетом (малою хвилею) називається деяка функція, зосереджена в невеликій околиці деякої точки та різко убутна до нуля в міру видалення від її як у часовий, так і в частотній області. Існують найрізноманітніші вейвлети, що мають різні властивості. Разом з тим, усі вейвлети мають вигляд коротких хвильових пакетів з нульовим інтегральним значенням, локалізованих на часовій осі, які є інваріантними до зсуву і до масштабування.

До будь-якому вейвлету можна застосувати дві операції:

- зрушення, тобто переміщення області його локалізації в часі;

- масштабування (розтягання або стиск).

Головна ідея вейвлет-претворення полягає в тому, що нестаціонарний часовий ряд розподіляється на окремі проміжки (так званні «вікна спостереження»), і на кожному з них виконується обчислення скалярного добутку (величини, що показує ступінь близькості двох закономірностей) досліджуваних даних з різними зрушеннями деякого вейвлета на різних масштабах. Вейвлет-перетворення генерує набір коефіцієнтів, за допомогою яких представляється початковий ряд. Вони є функціями двох змінних: часу і частоти, і тому утворюють поверхню у трьохвимірному просторі. Ці коефіцієнти, що показують, наскільки поведінка процесу в даній точці аналогічно вейвлету на даному масштабі. Чим ближче вид аналізованої залежності в околиці даної точки до виду вейвлета, тим більшу абсолютну величину має відповідний коефіцієнт. Негативні коефіцієнти показують, що залежність схожа на "дзеркальне відбиття" вейвлета. Використання цих операцій, з урахуванням властивості локальності вейвлета в частотно-часовій області, дозволяє аналізувати дані на різних масштабах і точно визначати положення їхніх характерних рис у часі.

Технологія використання вейвлетів дозволяє виявляти одиничні та нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу, зокрема, обсягів тематичних публікацій в Інтернет. При цьому можуть виявлятися моменти виникнення циклів, а також моментів, коли за періодами регулярної динаміки настають хаотичні коливання.

Часовий ряд, що розглядається, може апроксимуватися кривою, що, у свою чергу, може бути представлена у вигляді суми гармонійних коливань різної частоти й амплітуди. При цьому коливання, що мають низьку частоту, відповідають за повільні, плавні, великомасштабні зміни значень вихідного ряду, а високочастотні – за короткі, дрібномасштабні зміни. Ніж сильніше змінюється описувана даною закономірністю величина на даному масштабі, тим більшу амплітуду мають складові на відповідній частоті. Таким чином, досліджуваний часовий ряд можна розглядати в частотно-часовій області -

тобто про дослідження закономірності, що описує процес у залежності як від часу, так і від частоти.

Неперервне вейвлет-перетворення для функції $f(t)$ будується за допомогою неперервних масштабних перетворень і переносів вибраного вейвлета $\psi(t)$ з довільними значеннями масштабного коефіцієнта a та параметра зсуву b:

$$W(a,b) = (f(t), \psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi^* \left(\frac{t-b}{a} \right) dt .$$

Отримані коефіцієнти представляються у графічному вигляді картою коефіцієнтів перетворення, або скейлограмою. На скейлограмі по одній осі відкладаються зрушення вейвлета (вісь часу), а по іншій – масштаби (вісь масштабів), після чого точки схеми, що вийшла, офарбовуються залежно від величини відповідних коефіцієнтів (чим більше коефіцієнт, тим яскравіше кольори зображення). На скейлограмі видні всі характерні риси вихідного ряду: масштаб та інтенсивність періодичних змін, напрямок і величина трендів, наявність, розташування та тривалість локальних особливостей.

Наприклад, відомо, що комбінація декількох різних коливань може мати настільки складну форму, що виявити їх візуально не представляється можливим. Періодичні зміни, що відбуваються для значень коефіцієнтів вейвлет-перетворення на деякій неперервній множині частот виглядають як ланцюжок "пагорбів", що мають вершини, розташовані в точках (по осі часу), у яких ці зміни досягають найбільших значень.

Іншим важливим показником є виражена тенденція динаміки часового ряду (тренд) поза залежністю від періодичних коливань. Наявність тренда може бути неочевидною при простому розгляді часового ряду, наприклад, якщо тренд поєднується з періодичними коливаннями. Тренд відбивається на скейлограмі як плавна зміна яскравості уздовж осі часу одночасно на всіх масштабах. Якщо тренд наростаючий, то яскравість буде збільшуватися, якщо убутний - зменшуватися. Ще одним важливим фактором, якому