УДК 681.3

В.М. Фурашев, Д.В. Ланде

Практичні засади прогнозу можливих загроз та ризиків шляхом аналізу взаєзв'язку подій з інформаційним простором

Науково-дослідний центр правової інформатики
Академії правових наук України

Проаналізовано взаємозв'язок між подіями, пов'язаними із кризовими явищами в Україні, та їх відбиттями на веб-ресурсах мережі Інтернет. Подано механізми вейвлет- і фрактального аналізу тчасових рядів кількості тематичних публікацій. Показано ефективність і практичну дієвість розглядуваниxx механізмів для аналізу тенденцій і прогнозу соціальних явищ.
Ключові слова: інформаційна безпека, моніторинг ресурсів Інтернет, вейвлет-аналіз, фрактальний аналіз, прогнозування

Найбільш ефективним засобом забезпечення національної або особистої безпеки, відстоювання національних або особистих інтересів є передбачення та уникнення, будь-якими засобами, реальних або потенційних подій і явищ, які можуть бути реальними або потенційними загрозами.

Саме тому людство знаходиться у постійному пошуку механізмів передбачення, в першу чергу, "загрозливих" подій і явищ і, по-друге, їх усунення або нейтралізації. Цілком зрозуміло, що кожен етап розвитку суспільства потребує своїх механізмів, але всі ці механізми базуються на збиранні, аналізі та синтезі відповідної інформації.

У сучасному світі, враховуючи стан і тенденції розвитку інформаційного середовища, особливо у розвинутих країнах, з`являються нові можливості та попити у цій сфері.

Серед загроз національним інтересам і національній безпеці в Законі України "Про основи національної безпеки України" [1] (стаття 7) серед потенційних загроз у інформаційної сфері окремо зазначається: *"…намагання маніпулювати суспільною свідомістю, зокрема, шляхом поширення недостовірної, неповної або упередженої інформації*."

Сучасний інформаційний простір являє собою унікальну можливість отримання будь-якої інформації з любого питання, але при наявності відповідного інструментарію, застосування якого дозволяє аналізувати взаємозв'язок можливих подій або подій, які вже відбуваються, з інформаційною активністю визначеного кола джерел інформації.

Цю взаємозв'язаність можна проілюструвати на конкретних прикладах. Дослідження авторів проводилися на наборі документальних корпусів, що містять повідомлення онлайнових ЗМІ різних обсягів, зібраних з мережі Інтернет системою InfoStream [2], яка забезпечує інтеграцію та моніторинг мережних інформаційних ресурсів. За допомогою цієї системи виконується автоматизований збір інформації з веб-сайтів у режимі реального часу, її структурування, групування за семантичними ознаками, а також ефективний тематичний вибірний розподіл і надання доступу до інформації у пошукових режимах. За допомогою системи InfoStream охоплюються новини з тисяч вітчизняних і закордонних Web-сайтів, здійснюється їх обробка та узагальнення. 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 визначалася наведеними нижче запитами до системи InfoStream щодо розвитку кризових явищ в країні. Розглядалися такі запити інформаційно-пошуковою мовою системи InfoStream:

1. Запит (широкий, загальнотематичний):
(парламентсь~криз)|(політичн~криз)|(фінансов~криз)|(економічн~криз)
2. Запит (економічно-фінансова криза):
(фінансов~криз)|(економічн~криз)
3. Запит (заяви та виступи Прем'єр-міністра України щодо фінансово-економічної кризи):
((фінансов~криз)|(економічн~криз))&((виступ~тимошенк)|(заяв~тимошенк)|(виступ~юлі)|(заяв~юлі))&тимошенк
4. Запит (заяви та виступи Президента України щодо фінансово-економічної кризи):
((((фінансов~криз)|(економічн~криз))&
((виступ~ющенк)|(заяв~ющенк)|(виступ~віктор~ющенк)|(заяв~віктор~ющенк))

Досліджувалися інформаційні потоки, що надходили з понад тисячі українських мережних інформаційних ресурсів, серед яких лідерами за кількістю релевантних запитам 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 мережних документів. У результаті пошуку за найбільш широким, першим, запитом, який враховує всі основні аспекти кризових явищ, було знайдено 57245 релевантних документів. На основі обробки цих даних були отримані повні картини експериментальних даних - часові ряди за заданий період. На рис. 1 наведено графік кількості публікацій за першим запитом за днями 2008 року.

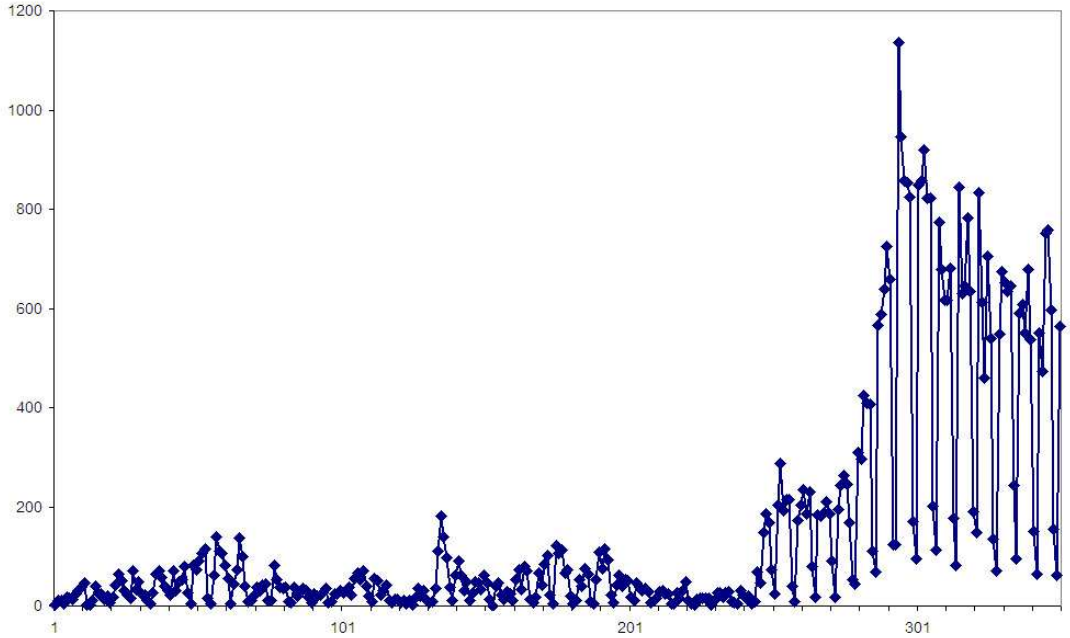

Рис. 1. Динаміка кількості публікацій за першим запитом за днями 2008 року (разом 57245 публікацій)

Представлений графік враховує тижневі коливання (у вихідні дні, наприклад, в мережі публікується значно менше документів, ніж у будні). Для більш наглядного відображення тенденцій подібні графіки згладжуються методом «ковзного середнього» з вікном спостереження у 7 діб. На рис. 2 наведено графіки, що відповідають першому та другому запитам. Зокрема, можна бачити, що приблизно в районі 250-го дня року загальна кількість повідомлень щодо кризової проблематики почала різко збільшуватися (посилилася парламентська криза), а лише з деяким запізненням, але значно потужніше розвився потік даних щодо другого запиту. На рис. 3 представлені згладжені графіки, які відповідають третьому та четвертому запитам.

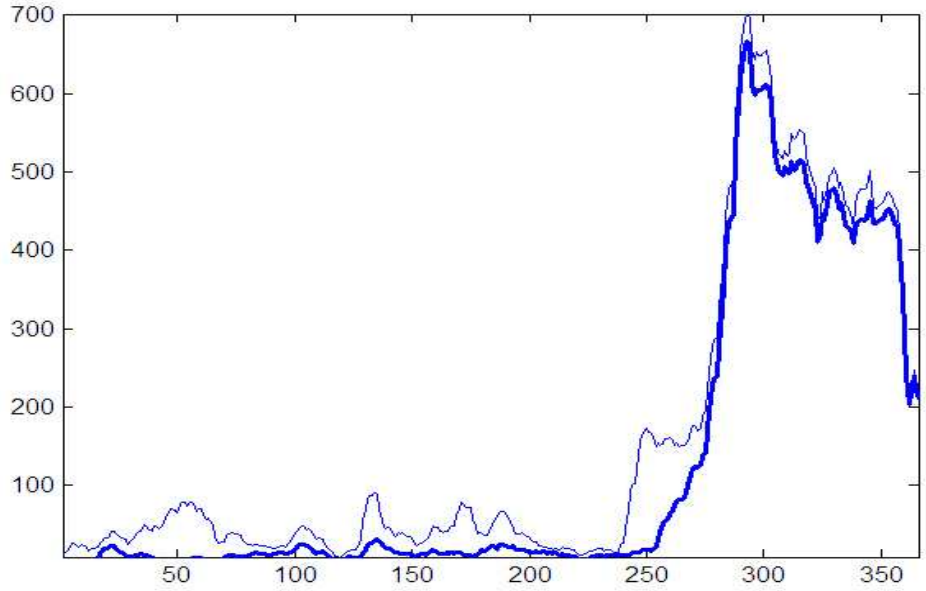

Рис. 2. Згладжений графік кількості публікацій за першим (тонка лінія) та другим (жирна лінія) запитом за днями 2008 року

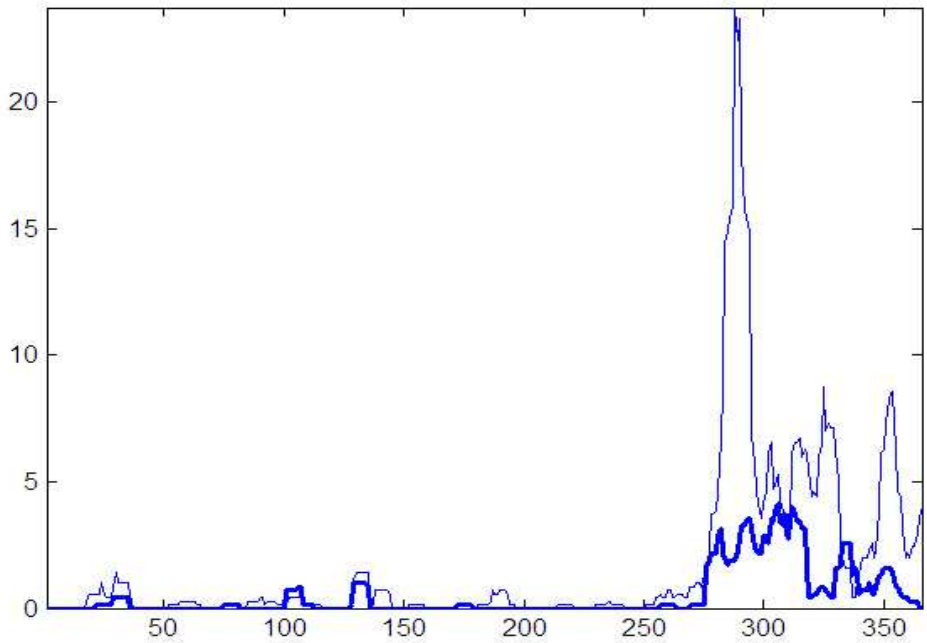

Рис. 3. Згладжений графік кількості публікацій за третім (595 публікацій, тонка лінія) та четвертим запитом (188 публікацій, жирна лінія) за днями 2008 року

Для третього запиту, на графіку якого піки виражені найбільш явно, можна надати такі змістовні пояснення:

1 – 24.01.2008 – коментарі до різких публікацій у російській пресі щодо заяви Ю.Тимошенко з приводу ціни на транзит газу

2 – 14.05.2008 – заява Ю.Тимошенко щодо можливості розриву союзу із Президентом

3 – 11.7.2008 – реакція на виступ Ю.Тимошенко під час розгляду проекту резолюції про недовіру уряду

4 – 10.10.2008 – Тимошенко спростовує плітки щодо її відставки

5 – 16.10.2008 - заява Ю.Тимошенко щодо формального подання В. Пінзеником змін до бюджету з урахуванням коштів на перевибори

6 – 20.10.2008 – переговори з МФО щодо надання фінансової допомоги Україні

7 – 13.11.2008 – заяви щодо можливості отримання фінансової підтримки з боку Світового банку

8 – 19.12.2008 – коментарі щодо різкої відповіді правління НБУ на заяву прем'єр-міністра

Характерно, що пікові значення на інших графіках найчастіше збігаються з наведеними датами, найбільше це характерно для другого запиту (*«(фінансов~криз)|(економічн~криз)»*).

Завдання вивчення статистичних властивостей мережних документальних масивів [3-5] є багатоплановим, припускає активне використання сучасних методів, що дозволяють більш глибоко зрозуміти специфіку предметної області. У цьому плані дуже перспективними представляються методи теорії детермінованого хаосу [6, 7], застосування теорії фракталів при аналізі інформаційного простору.

Теорія фракталів широко застосовується як підхід до статистичного дослідження, що дозволяє одержувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їхньої внутрішньої структури. 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.

Найважливішою характеристикою рядів, що мають хаотичну поведінку, є, як відомо, показник Херста [6], який визначається в результаті *R/S*-аналізу [4], що базується на аналізі нормованого розкиду - відносини розкиду значень досліджуваного R ряду до середньоквадратичного відхилення S.

У випадках, коли співвідношення R/S має сталий тренд можна говорити про співвідношення:

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

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

На рис. 4 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. 5) 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 досліджуваних інформаційних потоків відповідає величині ~0,89, що підтверджує припущення щодо самоподібності та ітеративності процесів в інформаційному просторі.

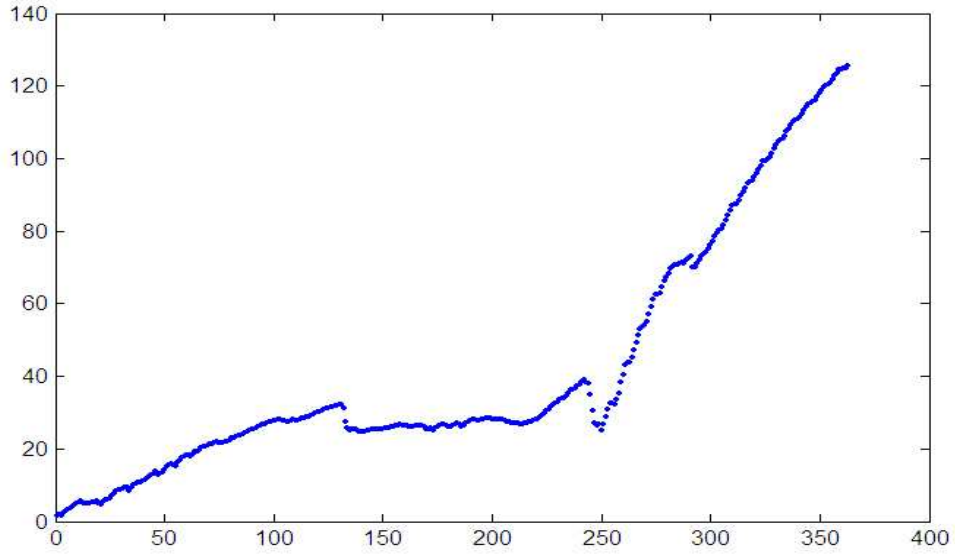

Рис. 4. Показник нормованого розкиду для всього періоду спостережень ряду, сформованому за першим запитом

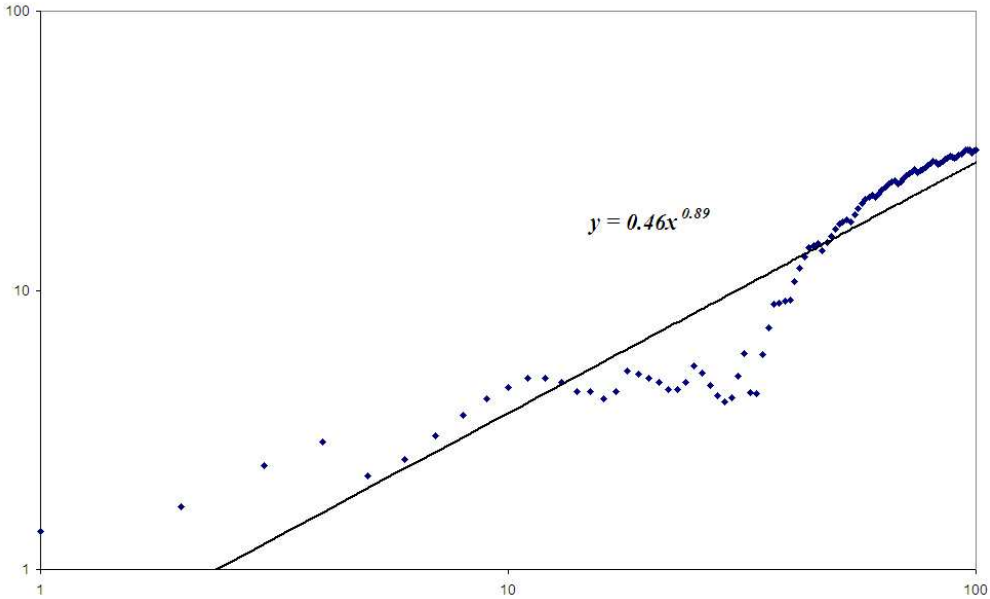

Рис. 5. Показник нормованого розкиду в логарифмічній шкалі за останні 120 днів року

Для часових рядів, що розглядалися, застосовувався ще один з підходів до виявлення самоподібності - метод DFA (Detrended Fluctuation Analysis) [8-10]. 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 апроксимаційної ділянки (вікна спостереження). Цей метод був застосований до ряду значень кількості публікацій, отриманих за представленими вище запитами. У методі DFA для різних ділянок ряду спостережень однакової довжини *k* досліджуваного ряду будується лінійна апроксимація, для якої потім обчислюється середньоквадратична помилка *D*(*k*). На рис. 6 представлена залежність середньоквадратичної помилки апроксимації від довжини ділянок апроксимації в подвійному логарифмічному масштабі.

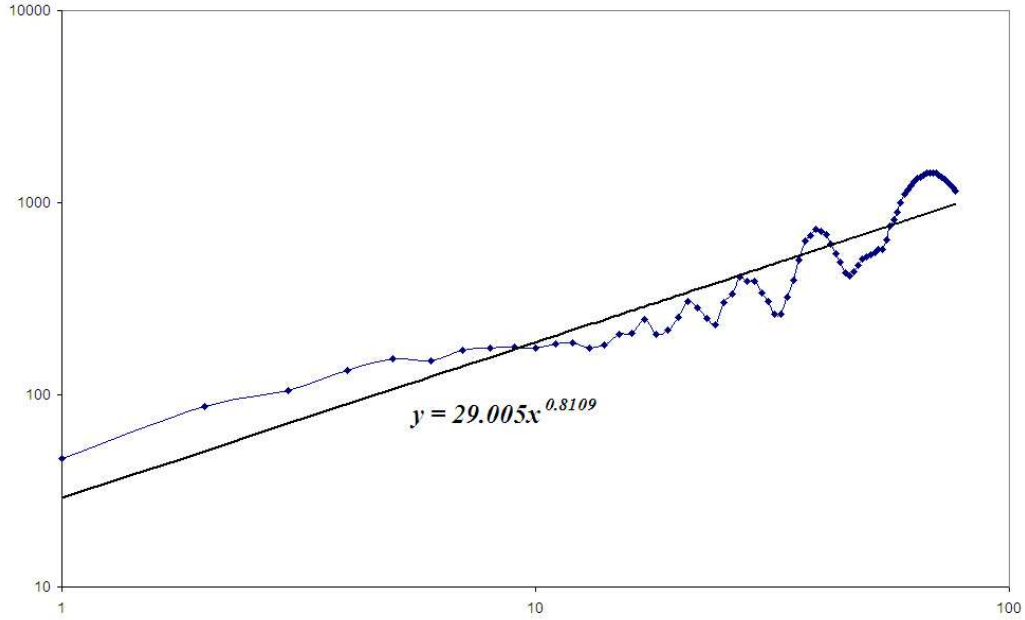

Рис. 6. Залежність середньоквадратичної помилки лінійної апроксимації D від довжини вікна спостереження k

Близькість залежності *D*(*k*) до лінійного ще раз підтверджує наявність локального скейлингу у другому півріччі 2008 року.

До кола найпоширеніших інструментальних засобів математичного моделювання та оцінки рядів спостережень відноситься також вейвлет-аналіз [11, 12]. Він особливо ефективний в тих випадках, коли крім загальних спектральних характеристик потрібно виявляти локальні в часі особливості поведінки досліджуваного процесу. Аналіз даних з використанням вейвлет-перетворень є зручним, надійним і потужним інструментом дослідження часових рядів і дозволяє представити результати у наочному вигляді, зручному інтерпретації.

Основою вейвлет-аналізу є вейвлет-перетворення, яке є особливим типом лінійного перетворення, базисні функції якого (вейвлети) мають специфічні властивості. *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 й у 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-якого 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 (т. з. «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 поводження процесу в даній крапці 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 [13, 14].

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, яке буде розглядатися далі, для функції $f(t)$ будується за допомогою неперервних масштабних перетворень і переносів 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 можна представити в графічному вигляді, якщо по одній осі відкласти зрушення вейвлета (вісь часу), а по іншій – масштаби (вісь масштабів), і офарбувати точки схеми, що вийшла, залежно від величини відповідних коефіцієнтів (чим більше коефіцієнт, тим яскравіше кольори зображення), називають картою коефіцієнтів перетворення, або скейлограмою. На скейлограмі видні всі характерні риси вихідного ряду: масштаб та інтенсивність

періодичних змін, напрямок і величина трендів, наявність, розташування та тривалість локальних особливостей.

Отримані скейлограми інтерпретуються таким чином [15,16]. У першу чергу аналізується так звана «енергія» (сума квадратів значень коефіцієнтів перетворення на кожному масштабі). Масштаби, що мають низькі значення енергії, практично не містять корисної інформації й можуть бути виключені з розгляду.

Одним з найважливіших аналітичних показників часових рядів є періодичність, тобто повторюваність через певні проміжки часу. При цьому 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 неочевидним при простому розгляді часового ряду, наприклад, якщо незначний тренд сполучається 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, який необхідно враховувати при аналізі тимчасових рядів, є локальні особливості, тобто можливі різкі, стрибкоподібні зміни характеристик вихідного ряду. Локальні особливості представлені на скейлограмі вейвлет-перетворення як лінії різкого перепаду яскравості, що виходять із крапки, що відповідає часу виникнення стрибка. Локальні особливості можуть мати як випадковий, так і систематичний характер, при цьому "маскувати" періодичні залежності або короткостроковий тренд. Аналіз локальних особливостей дозволяє відновити інформацію щодо динаміки вихідного процесу та у деяких випадках прогнозувати подібні ситуації.

Таким чином, кожний з основних факторів динаміки має свій, характерний відбиток на скейлограмі, при цьому вся аналітична інформація представляється в наочному й зручному для вивчення виді. Завдяки наочності подання результатів у вигляді скейлограми, іноді досить одного погляду, щоб побачити на ній найбільш яскраві фактори. На рис. 7 наведена скейлограма - результат неперервного вейвлет-аналізу (вейвлет Гауса) часового ряду, що відповідає процесу, який досліджується.

Наведений приклад показує, що вейвлет-аналіз дозволяє виявляти не тільки очевидні аномалії в досліджуваному ряді, але й критичні значення, які приховані за відносно невеликими абсолютними значеннями елементів ряду. Наприклад, на скелетоні на більшості частот відмічено не тільки 250-й день, але й неявні екстремуми (25-й та 130-й дні).

Основне спостереження: публікацій з приводу заяв лідерів країни щодо кризових явищ у сотні разів менше загальної кількості публікацій за цією темою. Разом з тим, виступи і заяви лідерів країни є камінцями збудження лавин таких публікацій, які впливають на суспільну думку та, у кінцевому рахунку, на інформаційну безпеку держави.

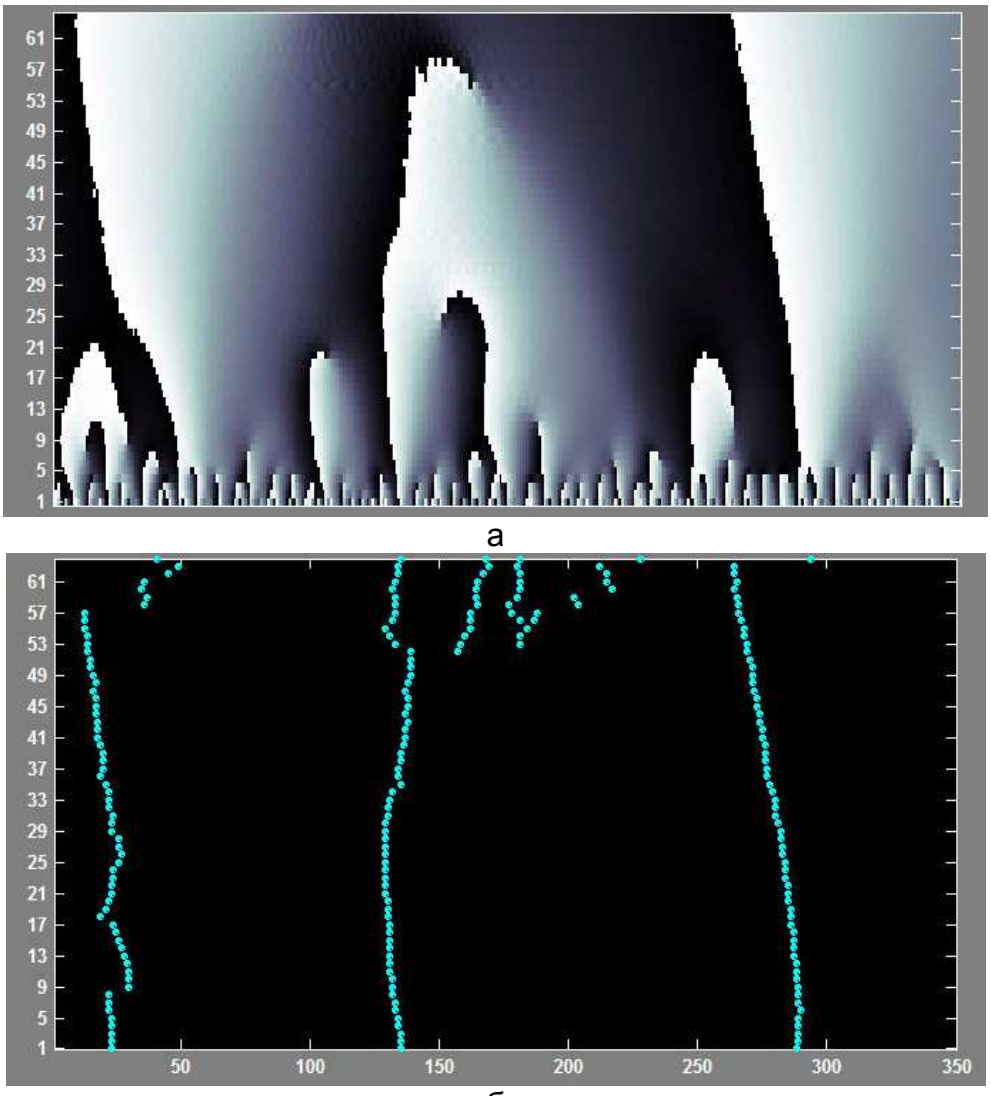

Рис. 7. Результат вейвлет-аналізу (неперервне вейвлет-перетворення):
а – вейвлет-скейлограма;
б – лінії локальних максимумів (скелетон)

Републікації, цитування, гіпертекстові та контекстні посилання тощо породжують самоподібність, наявність високого рівня статистичної кореляції в інформаційних потоках на тривалих часових інтервалах. Зокрема, на розглянутому прикладі висока персистентность процесу свідчить про загальну тенденцію високого рівня відображення у ЗМІ інформації щодо визначеної тематики.

Дані спостереження підтверджуються розвитком розглянутих питань, а також положеннями та практичними висновками робіт [17-19], що означає практичну дієвість запропонованих механізмів.

Враховуючи тенденції розвитку сучасного світу, національного суспільства та, особливо те, що країна “входить” у передвиборчий процес, застосування та подальший розвиток запропонованих підходів до прогнозної оцінки явищ та подій допоможе, у визначених умовах, запобігти або взагалі уникнути їх негативного впливу на окремі сторони суспільного життя.

Список література

[1] Закон України “Про основи національної безпеки України”. - Відомості Верховної Ради (ВВР), 2003, N 39, ст. 351.
[2] Григорьев А.Н., Ландэ Д.В., Бороденков С.А. та ін. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: науч.-метод. пособие. - К.: ООО «Старт-98», 2007. — 40 с.
[3]. Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1. — 2005. — Вып. 11. — С. 21–33.
[4] Ландэ Д.В. Фрактальные свойства тематических информацийнных потоков из Интернет // Реєстрація, зберігання і обробка даних. — 2006. — Т. 8, № 2. — С. 93–99.
[5] Иванов С.А. Стохастические фракталы в Информатике // Научно-техническая информация. Сер. 2. — 2002. — № 8. — С. 7–18.
[6] Федер Е. Фракталы. — М.: Мир, 1991. — 254 с.
[7] Van Raan A.F.J. Fractal Geometry of Information Space as Represented by Cocitation Clustering// Scientometrics. —1991. — Vol. 20, N 3. — P. 439–449.
[8] Peng C.-K., Havlin S., Stanley H.E., Goldberger A.L. Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series // CHAOS. — 1995. — Vol. 5. - P. 82.
[9] Stanley H.E., Amaral L.A.N., Goldberger A.L., Havlin S., Ivanov P.Ch., Peng C.-K. Statistical Physics and Physiology: Monofractal and Multifractal Approaches // Physica A. — 1999. - Vol. 270. — P. 309.
[10] Павлов А.Н., Сосновцева О.В., Зиганшин А.Р. Мультифрактальный анализ хаотической динамики взаимодействующих систем // Изв. ВУЗов. Прикладная нелинейная динамика. — 2003. — Т. 11, №. 2. — С. 39–54.
[11] Чуи К. Введение в вэйвлеты. - М.: Мир, 2001.
[12] Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. -1996. - Т. 166. - № 11. –С. 1145-1170.
[13] Давыдов А.А. Вейвлет-анализ социальных процес сов // Социолог. исслед. – 2003. - №11. - С. 97-103.
[14] Давыдов. А А. Системная социология. – М.: КомКнига, 2006. - 192 с.
[15] Buckheit J., Donoho D. Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. -27 p.
[16] Киселев А. Непрерывное вейвлет-преобразование в анализе бизнес-информации. URL: http://www.basegroup.ru/library/cleaning/wavelet_for_bussines/
[17] О.В. Литвиненко. Інформаційні впливи та операції - К.: НІСД, 2003. – 240 с. (Сер. «Національна безпека»; Вип. 6)
[18] В.М. Фурашев. До питання механізмів спотворення народного волевиявлення під час організації референдуму // Наук. журнал “Правова інформатика”. – К.: АПрН, НДЦПІ, 2009 – 1(13). - С. 54-64.
[19] В.М. Фурашев. Нормативно-правові засади системної інформатизації інформаційно-аналітичного забезпечення здійснення процедур виборчих і референдумних процесів - К.: Парламентське видавництво, 2006. – 144 с.

Рецензент: д.т.н., професор, заслужений діяч науки і техніки України І.Ф. Бінько

Поступила в редакцию 28.05.09.

Практические основы прогноза возможных угроз и рисков путем анализа взаимосвязи событий с информационным пространством

Проанализована взаимосвязь между событиями, связанными с кризисными явлениями в Украине, и их отражениями на веб-ресурсах сети Интернет. Представлены механизмы вейвлет- и фрактального анализа временных рядов количества тематических публикаций. Показаны эффективность и практическая действенность рассматриваемых механизмов для анализа тенденций и прогноза социальных явлений.

Ключевые слова: информационная безопасность, мониторинг ресурсов Интернет, вейвлет-анализ, фрактальный анализ, прогноз

Practical fundamentals of forecast of possible threats and risks by means of the analysis of interrelation of events with information space

In job the interrelation between the events connected to the crisis phenomena in Ukraine, and their reflections on web-resources is analyzed. Mechanisms wavelet- and fractal-analysis of time series of thematic publications quantity are submitted. Efficiency and practical effectiveness of considered mechanisms for the analysis of tendencies and the forecast of the social phenomena is shown.

Keywords: information safety, monitoring of Internet-resources, wavelet-analysis, fractal-analysis, the forecast