УДК 340.13+681.3+519.8

ЛАНДЕ Д.В., *доктор технічних наук, старший науковий співробітник*

ЗАХИСТ ПЕРСОНАЛЬНИХ ДАНИХ В УКРАЇНІ У ДЗЕРКАЛІ ВЕБ-ПРОСТОРУ

Анотація. *Досліджується взаємозв’язок між подіями, пов’язаними з проблематикою захисту персональних даних в Україні, та їх відбиттям на веб-ресурсах мережі Інтернет. Представлено механізми дисперсійного, вейвлет- і фрактального аналізу часових рядів обсягів тематичних публікацій. Наведено методологічні принципи дослідження, показано дієвість розглянутих механізмів для аналізу тенденцій соціальних явищ.*

Ключові слова: *захист персональних даних, Інтернет, веб-середовище, статистичний аналіз, ретроспективний аналіз.*

Аннотация. *Исследуется взаимосвязь между событиями, связанными с проблематикой защиты персональных данных в Украине, и их отражением на веб-ресурсах сети Интернет. Представлены механизмы дисперсионного, вейвлет- и фрактального анализа часовых рядов объемов тематических публикаций. Приведены методологические принципы исследования, показана действенность рассмотренных механизмов для анализа тенденций социальных явлений.*

Ключевые слова: *защита персональных данных, Интернет, веб-среда, статистический анализ, ретроспективный анализ.*

Summary. *Intercommunication is explored between events, related to issues of protection of the personal data in Ukraine, and their reflection on Internet web-resourses. The mechanisms of dispersional, wawelet- and fractal analysis of rows of sentinels of volumes of thematic publications are presented. Methodological principles of research are resulted, effectiveness of the considered mechanisms is shown for the analysis of tendencies of the social phenomena.*

Keywords: *protection of the personal data, Internet, web-environment, statistical analysis, retrospective analysis.*

Постановка проблеми. Системні дослідження такого багатоаспектного процесу, як захист персональних даних становлять не тільки теоретичний, а й суто практичний інтерес. Ці дослідження набувають особливої актуальності у нашій країні саме в даний час, коли Кабінет Міністрів України пропонує внести зміни до Закону “Про захист персональних даних” [1] та усунути прогалини, які були виявлені у ході застосування закону з 1 січня 2011 року. Відповідний законопроект, який розробило Міністерство юстиції, був ухвалений 10 травня 2012 року на засіданні Уряду. Цим законопроектом Уряд пропонуватиме Верховній Раді конкретизувати механізми захисту персональних даних, які у чинному законі виписані у більш загальний спосіб. Крім того, з урахуванням виявлених проблем, пропонується спростити державну реєстрацію баз персональних даних, а також розширити права суб’єкта персональних даних. Зокрема, пропонується скасувати державну реєстрацію баз персональних даних, ведення яких пов’язане із забезпеченням та реалізацією трудових відносин, звільнити від обов’язку реєстрації баз персональних даних громадські, релігійні організації, а також політичні партії, розширити права суб’єктів персональних даних, надавши їм право відкликати згоду на обробку своїх даних, вносити застереження стосовно обмеження права на обробку своїх персональних даних при наданні згоди тощо.

При ретроспективному аналізі будь-якого процесу або явища інтерес становлять певні характеристики їх розвитку, а саме [2, 3]:

- кількісна динаміка, притаманна процесу або явищу, наприклад, кількість подій в одиницю часу або кількість цільових повідомлень;
- визначення критичних, порогових точок, що відповідають кількісній динаміці процесу або явища;
- визначення проявів процесу явища у критичних точках, наприклад, виявлення основних сюжетів публікацій у ЗМІ;
- після виявлення основних проявів процесу або явища у критичних точках ці прояви ранжируються, досліджується динаміка розвитку окремих визначених проявів до та після визначення критичних точок;
- здійснюється статистичний, дисперсійний, вейвлет-, фрактальний аналіз загальної динаміки та динаміки окремих проявів, на основі яких робляться спроби прогнозування розвитку процесу або явища та окремих його проявів.

Інформаційний простір мережі Інтернет дозволяє аналізувати взаємозв’язок можливих подій або подій, які вже відбуваються, з інформаційною активністю визначеного кола джерел інформації. Дослідження у рамках цієї роботи проводилися на базі веб-повідомлень, зібраних з мережі Інтернет системою InfoStream [4], яка забезпечує інтеграцію інформаційних ресурсів. За допомогою цієї системи охоплюються новини з тисяч вітчизняних і закордонних веб-сайтів. Система забезпечує доступ до унікального ретроспективного фонду, що перевищує 100 млн. записів за більш як 10 років, та підтримку аналітичної роботи в режимі реального часу, у тому числі виявлення сюжетних ланцюжків та діаграм появи у часі.

Метою статті є аналіз інформаційного потоку щодо тематики захисту персональних даних.

Виклад основних положень. Тематика досліджуваного інформаційного потоку визначалася запитом до системи InfoStream в інформаційному просторі країни:

((захист~персональн~даних)|(защит~персональн~данн)) & country.UA

Документи, релевантні наведеному запиту, можуть бути представлені двома мовами (українською та російською), містити словосполучення типу “*Захист персональн даних*” або “*Защита персональных данных*”. Джерелами інформації мають бути веб-ресурси з нашої країни.

Досліджувалися інформаційні потоки, що надходили з понад тисячі українських мережних інформаційних ресурсів, серед яких лідерами за кількістю релевантних запитів публікацій були такі авторитетні джерела, як Finance.ua, “Украина-Сегодня”, “РБК-Украина”, “Українські національні новини”, ForUm, “Укррудпром”, “Українська правда” тощо. Ретроспективний період дослідження становив весь 2010 та весь 2011 роки, а також перші 4 місяці 2012 року, тобто 851 добу.

У результаті пошуку за наведеним запитом було знайдено 12052 веб-публікацій. На основі обробки цих даних були отримані повні картини експериментальних даних – часові ряди за заданий період. На Рис. 1 наведено графік кількості публікацій за запитом за вказаний період.

Представлений графік враховує тижневі коливання (у вихідні дні в мережі публікується значно менше документів, ніж у будні). Для більш наглядного відображення тенденцій подібні графіки згладжуються методом “ковзного середнього” з вікном спостереження, кратним 7 добам. На Рис. 2 наведено згладжений графік, що відповідає наведеній вище динаміці. Зокрема, можна бачити, що приблизно на 700-й день досліджуваного періоду загальна кількість повідомлень щодо захисту персональних даних різко збільшилася.

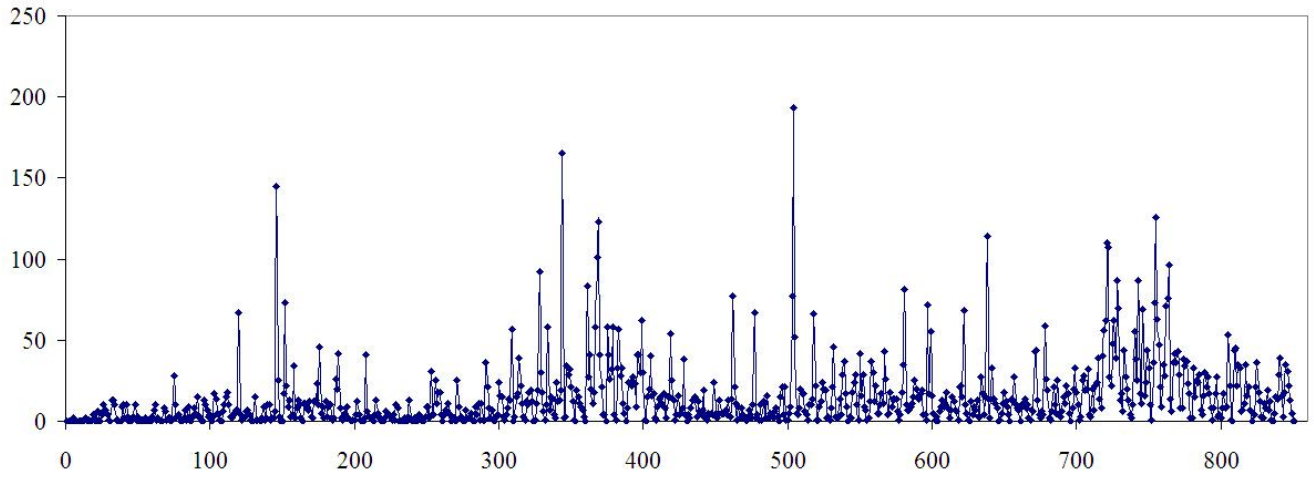


Рис. 1. Динаміка кількості публікацій за цільовим запитом за вказаний період (12052 веб-публікацій)

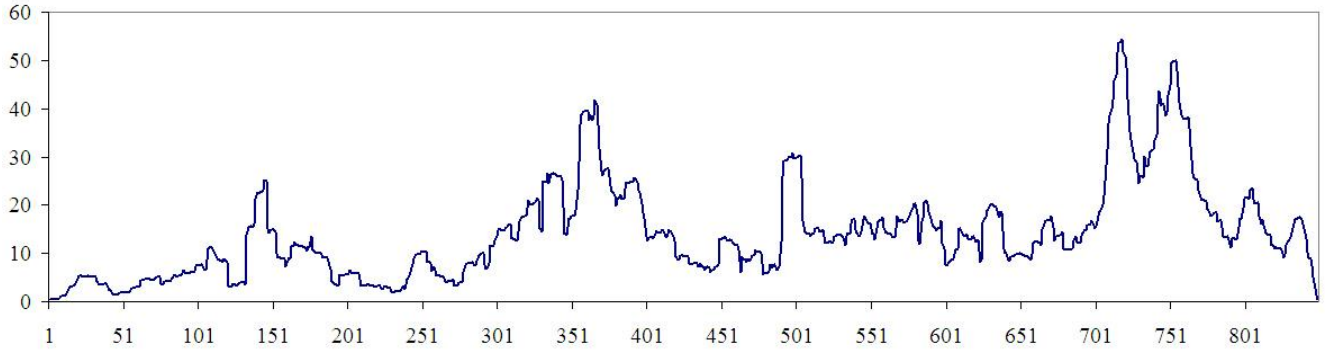


Рис. 2. Згладжений графік кількості публікацій за період досліджень

У Таблиці 1 наведено виявлені системою InfoStream основні сюжетні ланцюжки, які стосуються деяких пікових значень інтенсивності публікацій впродовж зазначеного періоду.

Таблиця 1. Основні сюжети щодо визначеної тематики

№	Дата	Заголовок повідомлення	Коментар
1	2010.05.26	Уряд затвердив план інтеграції до Євросоюзу	Для отримання дорожньої карти безвізового режиму Кабмін обіцяє забезпечити прийняття парламентом ряду законопроектів про захист персональних даних.
2	2010.06.01	Верховна Рада України ухвалила Закон “Про захист персональних даних”	Законом встановлюється, що володільцем чи розпорядником бази персональних даних можуть бути підприємства, установи і організації усіх форм власності, органи державної влади чи органи місцевого самоврядування, фізичні особи-підприємці, які обробляють персональні дані відповідно до закону.
3	2010.12.10	Указом Президента України утворено та реорганізовано 17 державних служб України	Указом Президента України № 1085/2010 “Про оптимізацію системи центральних органів виконавчої влади” постановлено утворити Державну службу України з питань захисту персональних даних.

№	Дата	Заголовок повідомлення	Коментар
4	2011.05.19	Україна до кінця року перейде на біометричні паспорти	Андрій Клюєв: “план лібералізації візового режиму також передбачає створення системи захисту персональних даних європейського рівня”.
5	2011.09.30	Податкова служба не має наміру збирати паспортні дані громадян в банках та обмінних пунктах.	Віталій Захарченко: “постанова Нацбанку щодо ідентифікації осіб, які обмінюють готівкову валюту, аж ніяк не відміняє законів, що регламентують банківську таємницю та захист персональних даних”.
6	2011.12.22	Клюєв: “Україна виконала левову частку роботи для скасування візового режиму з ЄС”	Доповнення до угоди про спрощення візового режиму між Україною та Європейським Союзом можуть бути підписані ще до кінця зими.
7	2012.01.25	Мін’юст зареєстрував положення про Єдиний державний реєстр осіб, які вчинили корупційні правопорушення	Відомості з Реєстру підпадатимуть під дію Закону “Про захист персональних даних”. Так, відомості з Реєстру надаватимуться виключно на запит державних органів, органів місцевого самоврядування з метою проведення спеціальної перевірки відомостей про осіб, які претендують на зайняття посад, пов’язаних з виконанням функцій держави або місцевого самоврядування
8	2012.03.15	Хорошковський: впровадження біометричних документів - технічне завдання	Під час засідання окрему увагу було приділено завершенню законодавчої роботи у сферах управління міграцією та удосконалення захисту персональних даних.

До вихідного часового ряду може застосовуватися ще один різновид згладжування –експоненціальне ковзне середнє. Крива експоненціального ковзного середнього іноді краще апроксимує графік, ніж крива зваженого ковзного середнього, але вона залежить від вибору коефіцієнта згладжування α. Математична формула для розрахунку експоненціального ковзного середнього є рекурсивною і при значенні коефіцієнта згладжування α має вигляд:

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

де: $X(i)$ – значення ряду спостережень в точці, $E(i-1)$ – значення експоненціального ковзного середнього, розрахованого для попередньої точки ряду, α – регулюючий коефіцієнт. Початкового значення $E(1)$ набуває рівним $X(1)$. Чим більше значення регулюючого коефіцієнта, тим краще крива експоненціального ковзного середнього апроксимує графік, оскільки більшу вагу мають поточні значення (Рис. 3).

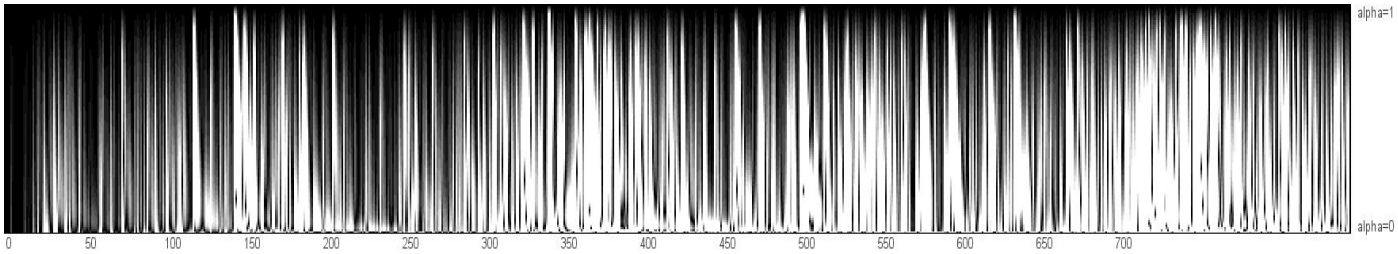


Рис. 3. Діаграми експоненціального згладжування при різних значеннях коефіцієнта α (світлі кольори відповідають більшим значенням)

Криві експоненціального ковзного середнього трактуються так само, як і криві звичайного (зваженого) ковзного середнього. В процесі аналізу слід знати, чому дорівнює значення регулюючого коефіцієнта.

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

Вивчення статистичних властивостей мережних документальних масивів [5 – 6] є багатоплановим, припускає активне використання сучасних методів, що дозволяють більш глибоко зрозуміти специфіку предметної області. У цьому плані перспективними представляються методи теорії детермінованого хаосу [6 – 7], застосування теорії фракталів при аналізі інформаційного простору. Найбільш цікавим у рамках даного дослідження виявляється наявність таких властивостей, як самоподібність (масштабна інваріантість, скейлинг), стійкі взаємні кореляції.

Важливою характеристикою рядів, що мають хаотичну поведінку, є показник Херста, який визначається в результаті так званого *R/S*-аналізу [7], що базується на аналізі нормованого розкиду – відносини розкиду *R* значень досліджуваного ряду до середньоквадратичного відхилення S. У разі якщо співвідношення *R/S* має сталий тренд, можна говорити про співвідношення:

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

де: *H* – показник Херста.

На Рис. 4 представлено співвідношення *R/S* для досліджуваного ряду з результатів наведеного вище запиту. Як можна бачити, крива нормованого розмаху (рис. 6) досі 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 ступеням автокореляції ряду.

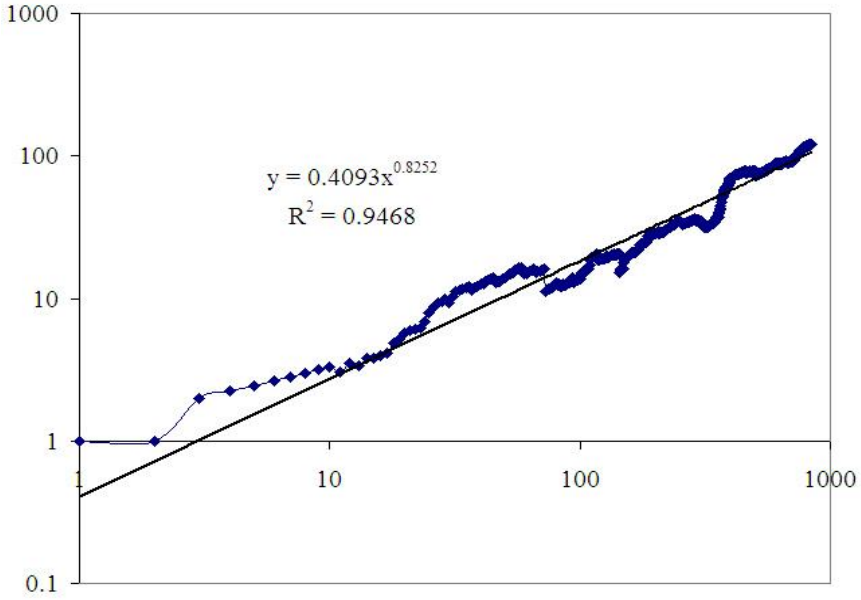


Рис. 4. Показник нормованого розкиду R/S в логарифмічній шкалі для всього періоду спостережень

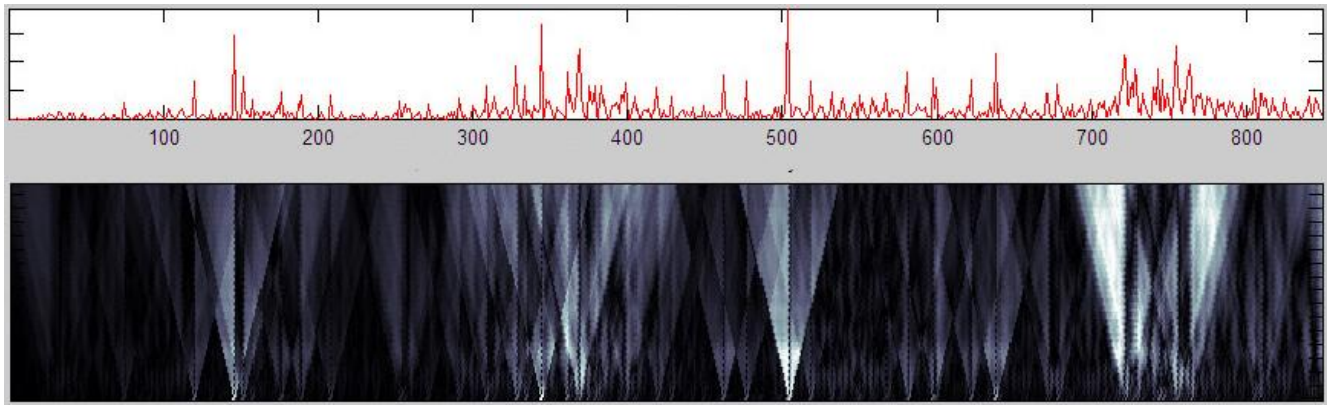
Значення Херста для досліджуваних інформаційних потоків відповідає величині ~ 0,83, що підтверджує припущення щодо самоподібності та ітерактивності публікацій щодо захисту персональних даних в інформаційному просторі. Це означає, що деякі резонансні публікації багаторазово дублюються, переказуються, обговорюються. При цьому загальна інформаційна напруженість залишається на великому рівні. Як тільки зникає “шлейф” одного сюжету щодо заданої теми, йому на зміну виникає новий сюжет, інколи, більш інтенсивний.

До кола поширених засобів оцінки рядів спостережень відноситься вейвлет-аналіз [8, 9], який особливо ефективний в тих випадках, коли необхідно виявляти локальні особливості поведінки досліджуваного процесу.

Головна ідея вейвлет-перетворення полягає в тому, що нестаціонарний часовий ряд розподіляється на окремі “вікна спостереження” і на кожному з них виконується обчислення величини, що показує ступінь близькості закономірностей досліджуваних даних з різними зрушеннями до деякого вейвлета (спеціальної функції) на різних масштабах. Вейвлет-перетворення генерує набір коефіцієнтів, що є функціями двох змінних: часу і частоти, і тому утворюють поверхню у тривимірному просторі. Вейвлет-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 (наприклад, чим більше коефіцієнт, тим яскравіше кольори). Отримані зображення називають скейлограмами.

Технологія використання вейвлетів дозволяє виявляти цикли, одиничні та нерегулярні “сплески”, різкі зміни кількісних показників у різні періоди часу, зокрема, обсягів тематичних публікацій в Інтернеті. Також виявляються моменти, коли за періодами регулярної динаміки настають хаотичні коливання [10, 11].

а)



б)

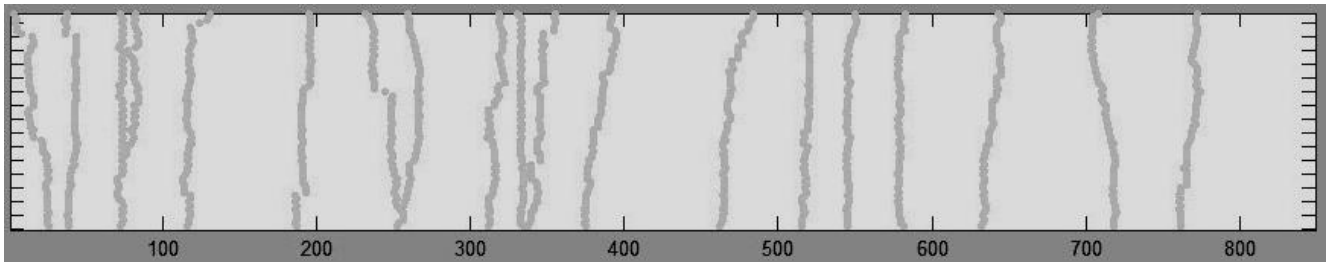


Рис. 5. Результат вейвлет-аналізу (неперервне вейвлет-перетворення):
а) вейвлет-скейлограма; *б)* лінії локальних максимумів (скелетон)

Кожний з основних факторів динаміки вихідного процесу має свій характерний відбиток на скейлограмі, при цьому інформація представляється в наочному й зручному для вивчення виді. На Рис. 5 наведена скейлограма – результат неперервного вейвлет-аналізу (вейвлет Хаара) часового ряду, що відповідає процесу, який досліджується.

До висновків.

Вейвлет-аналіз дозволяє виявляти не тільки очевидні аномалії в досліджуваному ряді, а й критичні значення, які приховані за відносно невеликими абсолютними значеннями елементів ряду. Наприклад, на скелетоні на більшості частот відмічено не тільки 344-й, 504-й, 755-й дні, а й неявні екстремуми (208-й, 253-й, 550-й дні тощо).

З метою візуалізації та аналізу часових рядів, пов’язаних із публікаціями в інформаційному просторі мережі Інтернет, був розроблений новий метод дисперсійного аналізу – Δ*L*-метод [12]. У відповідності з цим методом візуалізується відхилення точок ряду накопичення від локальної лінійної апроксимації. Така візуалізація у вигляді “рельєфної” діаграми становить певний інтерес для вивчення особливостей процесів, що відповідають вихідним часовим рядам (Рис. 6).

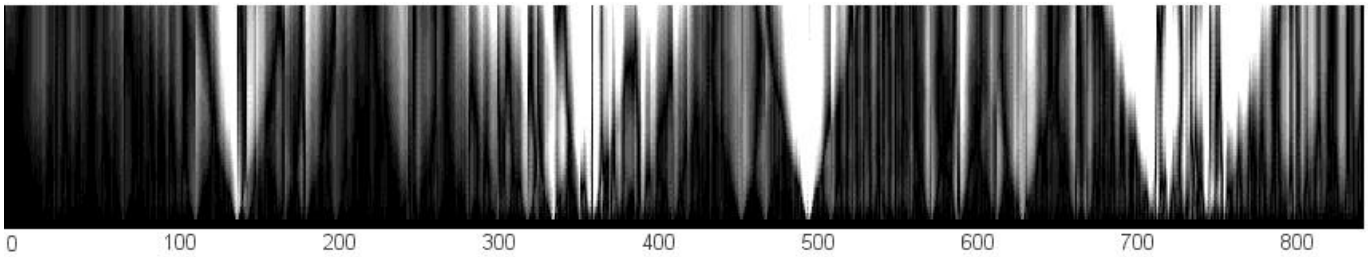


Рис. 6. Δ*L*-діаграма часового ряду інтенсивності тематичних публікацій

Запропонований метод візуалізації абсолютних відхилень Δ*L*, як і метод вейвлет-перетворень, дає змогу (і як показано на прикладі – не гірше) виявляти одиничні та нерегулярні “сплески”, різкі зміни значень кількісних показників у різні періоди часу.

На діаграмі, Рис. 6, додатково до вже розглянутих чітко видні аномалії у 120-й (ЄС передав Україні список з вимогами для скасування візового режиму, 2010.04.30), 328-й (ЄС висунув до України 60 вимог – для безвізового режиму, 2010.11.24), 622-й (МВС відмовляється повідомляти, за ким закріплені автомобілі з серією ВР, 2011.09.14) дні тощо.

Передрукування, цитування, повтори через певний час породжують високий рівень статистичної автокореляції в інформаційних потоках на досить тривалих часових інтервалах. Зокрема, на розглянутому прикладі висока персистентність процесу свідчить про загальну тенденцію високого рівня відображення у веб-просторі інформації щодо тематики захисту персональних даних.

Враховуючи тенденції розвитку сучасного світу, інформаційного суспільства, застосування та подальший розвиток запропонованих підходів та механізмів на базі статистично-семантичного аналізу контенту веб-простору може стати дієвим засобом аналітичної діяльності щодо оцінки процесів, явищ та подій.

Використана література

1. Про захист персональних даних : Закон України // Відомості Верховної Ради України. –2010. – № 34. – С. 481.

2. Фурашев В.М., Ланде Д.В. Інформаційні операції крізь призму системи моніторингу та інтеграції Інтернет-ресурсів // Правова інформатика. – 2009. – № 2(22). – С. 49 – 57.

3. Ланде Д.В., Фурашев В.М. Основи інформаційного і соціально-правового моделювання : монографія. – К. : ТОВ “ПанТот”, 2012. – 144 с.

4. Григорьев А.Н. InfoStream. Мониторинг новостей из Інтернет: технология, система, сервис : научно-методическое пособие / А.Н. Григорьев, Д.В. Ландэ, С.А. Бороденков и др. – К. : ООО “Старт-98”, 2007. – 40 с.

5. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет // Реєстрація, зберігання і обробка даних. – 2006. – № 2. – С. 93 – 99.

6. Van Raan A.F.J. Fractal Geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. – 1991. – 20. – № 3. – P. 439 – 449.

7. Федер Е. Фракталы. – М. : Мир, 1991. – 254 с.

8. Чуи К. Введение в вэйвлеты. – М. : Мир, 2001.

9. Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. – 1996. – 166. – № 11. – С. 1145 – 1170.

10. Давыдов А.А. Вейвлет-анализ социальных процессов // Социологические исследования. – 2003. – № 11. – С. 97 – 103.

11. Давыдов А.А. Системная социология. – М. : КомКнига, 2006. – 192 с.

12. Ландэ Д.В., Снарский А.А. Динамика отклонения элементов ряда измерений от локальных линейных аппроксимаций // Реєстрація, зберігання і оброб. даних. – 2009. – № 1. – С. 27 – 32.

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~