РАДА НАЦІОНАЛЬНОЇ БЕЗПЕКИ І ОБОРОНИ УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ НАЦІОНАЛЬНОЇ БЕЗПЕКИ

НАЦІОНАЛЬНА БЕЗПЕКА: УКРАЇНСЬКИЙ ВИМІР

Щоквартальний науковий збірник

Вип. 1-2 (20-21)

Київ 2008

УДК

ББК 66.2(4УКР)я5
Н35

Національна безпека: український вимір: щокв. наук. зб. / Рада нац. безпеки і оборони України, Ін-т пробл. нац. безпеки; редкол.: Горбулін В.П. (голов. ред.)[та ін.]. – К., 2008. – Вип. 1-2 (20-21). – 160 с.

У першому розділі збірника друкується низка статей з проблем державотворення і зміцнення національної безпеки передусім у сферах, що відносяться до так званого силового блоку уряду й держави. У другому розділі представлені статті, що розкривають актуальні проблеми економічної, енергетичної, інформаційної, техногенної та екологічної безпеки України. Друкуються також проект Концепції розвитку земельних відносин в Україні на 2008-2015 роки, матеріали «круглого столу» з теми: «Енергетична безпека України. Розвиток вітчизняних елементів ядерно-паливного циклу», інформаційні повідомлення до 5-річчя Інституту проблем національної безпеки та про нові книги з питань національної безпеки України.

ББК 66.2(4УКР)я5
Н35

Рекомендовано до друку Вченою радою Інституту проблем національної безпеки (прот. № 4 від 19 вересня 2008 р.)

Зареєстровано у Міністерстві юстиції 26.05.2008 р. (Свідоцтво про державну реєстрацію друкованого засобу масової інформації, Серія КВ № 13990-2963 ПР).

Автори опублікованих матеріалів несуть відповідальність за добір і точність наведених цитат, формул, власних імен та інших відомостей.

Адреса редакції:
03186, м. Київ, Чоколівський бульвар, 13, Інститут проблем національної безпеки
Тел.: (044) 245-48-44, факс: (044) 245-88-11, E-mail: rubanetz@nbu.gov.ua

УДК 342.4

Д.В. Ланде,
д-р техн. наук

НОВІТНІ ПІДХОДИ Й ТЕХНОЛОГІЇ ІНФОРМАЦІЙНО-АНАЛІТИЧНОЇ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

Сучасні засоби інформаційно-аналітичної підтримки прийняття рішень забезпечують вирішення цілого комплексу проблем, серед яких збір інформації про об'єкти, визначення зв'язків об'єктів, виявлення тенденцій, прогнозування. Функціональні можливості таких систем дозволяють виконувати діагностику та прогнозування розвитку ситуації. На додаток до можливостей глибинного аналізу даних і тексту у таких системах використовується також людський досвід, знання експертів. Нині вже очевидно, що реальний прорив у сфері інтенсифікації інформаційно-аналітичної роботи, як і в науці, можливий лише в результаті агрегування різних напрямків. Викладені в статті підходи з декількох, конфліктних раніше точок зору, сьогодні можуть розглядатися як шляхи рішення відкритої проблеми навігації в сучасному інформаційному просторі.

Ключові слова: *Інтернет, прийняття рішення, інформаційні потоки, складні мережі, масив мережевих публікацій, комп'ютерна вірусологія.*

Для інформаційно-аналітичної підтримки при вирішенні задач у сфері національної безпеки мають застосовуватися найсучасніші підходи, що реалізують найпродуктивніші технологічні ідеї та наукові концепції. З поміж них, на думку автора, нині викристалізувалися три основні: це, по-перше, глибинний аналіз текстів (Text Mining) [1], по-друге, концепція складних мереж (Complexes Networks) [2] і, по-третє, застосування методів нелінійної динаміки до аналізу інформаційних потоків і прогнозування [3, 4]. І звичайно ж, важливе значення в інформаційно-аналітичній роботі має доступ до джерел даних, інформації та знань.

Нагадаємо, що під даними розуміють «сирі», неопрацьовані відомості, засновані на фактах. Це можуть бути статистичні дані, факти з біографій або, наприклад, відомості зі звітності окремих компаній. Інформація - вже певним чином вибрані, оброблені та проаналізовані дані. Кінцевим же інформаційним продуктом будь-якої аналітичної роботи повинні бути знання - синтезовані висновки, рекомендації для прийняття рішень.

Процес перетворення даних у знання й доведення їх до кінцевих споживачів прийнято називати розвідувальним циклом [5, 6]. Його прийнято ділити на п'ять основних етапів:

- вибір цілі, планування, визначення джерел інформації;
- збір, добування даних;
- перетворення даних в інформацію шляхом обробки;
- аналіз і синтез інформації, перетворення її в знання;
- доведення знань до кінцевих споживачів.

За визнанням адмірала Захаріаса, заступника начальника розвідки ВМС США в роки Другої світової війни, 95% інформації розвідка військово-морських частин черпала з відкритих джерел, 4% - з офіційних і тільки 1% - з конфіденційних. На думку колишнього директора ЦРУ Хілленкерта, «80% розвідувальної інформації виходить з таких джерел, як книги, журнали, науково-технічні огляди тощо». Тобто інформація може бути отримана з офіційних джерел, неофіційних відкритих джерел, ЗМІ, оголошень, реклами, внутріфірмових, банківських, урядових звітів, баз даних, від експертів, шляхом аналізу або спеціальної обробки даних, текстів за прямими чи непрямими ознаками. Сьогодні, за оцінками експертів, саме мережа Інтернет за кількістю інформації посідає перше місце. При цьому у відкритих джерелах і спеціалізованих базах даних, доступних в Інтернет, міститься більша частина інформації, яка необхідна для проведення аналітичних досліджень, однак залишається відкритим питання її знаходження та ефективного використання.

Останні дослідження інформаційного веб-простору показали, що доступні через традиційні інформаційно-пошукові системи 20 млрд. вебсторінок - це лише «поверхнева видима частина айсберга». Кількість веб-сайтів в Інтернет збільшується зі швидкістю понад мільйона за місяць. Прихованих і невидимих (deep, invisible) ресурсів значно більше [7]. Це насамперед динамічні веб-сторінки, інформація із численних баз даних, які можуть становити великий інтерес для аналітичної роботи. До розряду «прихованого» веб відноситься, наприклад, і найбільша у світі повнотекстова онлайнова інформаційна система Lexis-Nexis, що містить понад 2 мільярди документів з ретроспективою понад 30 років. Щотижня в архіви цієї служби додається 14 млн. документів. На відміну від неструктурованих масивів традиційного «поверхневого» веб, користувачі Lexis-Nexis можуть використати потужний інструмент пошуку для доступу до достовірної та класифікованої інформації.

До «прихованих» ресурсів Інтернет можна віднести також пірінгові мережі, такі як BitTorrent, EDonkey, EMule, Gnutella, Kazaa.

Відомо, що при здійсненні інформаційно-аналітичної діяльності нині виникає ряд проблем.

Першою та найістотнішою є те, що колосальні обсяги інформації в Інтернет утрудняють пошук і вибір дійсно потрібних відомостей. Адже самі по собі неопрацьовані дані не можуть служити підмогою для прийняття рішень. Так, за словами Еріка Шмідта - голови Google, із п'яти мільйонів терабайтів існуючої сьогодні інформації в електронному вигляді проіндексовано лише 170 терабайтів і навіть така потужна пошукова система як Google зможе проіндексувати всю наявну інформацію лише приблизно через 300 років.

Друга проблема зумовлюється тим, що інформація в Інтернет має явно виражений динамічний характер: процес її розміщення, модифікації та вилучення цьому середовищі є безперервним. Наявні засоби узагальнення цієї інформації, природно, мають затримку у часі, що суттєво впливає на оперативність аналітичної роботи. Лише часткове вирішення цієї проблеми можливе при застосуванні систем контент-моніторингу інформаційних потоків.

Третя проблема пов'язана з необхідністю автоматичного витягу понять із формалізованих масивів інформації (таблиць, баз даних) і неструктурованих текстів.

Четверта проблема зумовлена необхідністю виявлення неочевидних закономірностей і зв'язків з наявних у користувачів даних. Нині відомо декілька шляхів вирішення цієї проблеми, наприклад, шляхом застосування концепції складних мереж.

П'ята проблема пов'язана з необхідністю пошуку інформації в «прихованому» веб-просторі, де міститься величезна кількість даних, потенційно більш цікавих для аналітичної роботі, ніж у відкритій частині Інтернету.

Повне вирішення названих проблем здається недосяжною метою. Разом з тим інформаційно-аналітичні системи підтримки прийняття важливих рішень мають забезпечувати реалізацію наступних трьох принципів:

– єдиний інформаційний простір взаємозалежних об'єктів і фактів незалежно від типу їх джерел або змісту;

– забезпечення зв'язку об'єктів і фактів з релевантними даними та джерелами інформації;

– наявність історично-просторової моделі банку даних системи, що припускає врахування атрибутів часу та місця для всіх об'єктів, а також неможливість їх безповоротного вилучення із системи.

Задля справедливості зазначимо, що згідно зі звітом Fuld's Intellgense Software Report 2006, відомих комерційних версій повноцінних інтегрованих систем, які дозволяють вирішувати весь комплекс названих проблем, поки не існує, принаймні на Заході.

Традиційно аналітичні дослідження спираються на інформаційні джерела, що містять інформацію про нові технології, утворення економічних і політичних об'єднань, злиття і придбання, оголошення про вакансії, виставки та конференції тощо. Тому останнім часом великої популярності набувають бази даних на основі архівів мас-медіа, в тому числі й мережних. У Росії, наприклад, великою популярністю користуються такі бази даних, як «Інтегрум», «Медіалогія». Так, агентство «Інтегрум» забезпечує збір електронних версій інформаційних продуктів різноманітних інформаційних джерел та інтегрує їх у єдиний масив. «Інтегрум» сьогодні - це найбільша в Росії інтегрована служба інформаційних ресурсів, що містить новини, комерційну та статистичну інформацію. Технологічною основою служби «Інтегрум» є інформаційно-пошукова система «Артефакт», у базах даних якої зібрано понад 400 млн. документів з 7500 джерел. В Україні цю нішу займає система інтеграції та моніторингу інформаційних ресурсів Інтернет InfoStream, яка орієнтована виключно на мережні засоби масової інформації та містить понад 60 млн. документів з понад 3500 джерел.

Концепція глибинного аналізу даних та тексту

За зовнішнім хаосом інформації, доступної як з Інтернет, так і з інших джерел, приховуються глибинні зв'язки, достовірна інформація, знання. Розібратися із цими даними можна, застосовуючи сучасні засоби аналізу змісту текстів, зокрема контент-аналізу глибинного аналізу даних і текстів (Data Mining, Text Mining). Зазначимо, що вони створювалися вченими в інтересах спецслужб протягом багатьох років, як на Заході, так і в колишньому Радянському Союзі. Переведення за останні 10-20 років значного обсягу інформації в електронну форму, широке використання та різке розширення мережі Інтернет, нові технології зробили аналітичну роботу в Інтернет однією з найперспективніших, а той факт, що це вже звична практика всіх спецслужб світу, лише підтверджує перспективність цього напрямку.

Важливе завдання технології Text Mining пов'язане з витягом із тексту його характерних елементів або властивостей, які можуть використатися як метадані документа, ключових слів, анотацій. Інше важливе завдання полягає у віднесенні документа до деяких категорій з наперед заданої схеми класифікації. Text Mining також забезпечує новий рівень семантичного пошуку документів.

Відповідно до сформованої нині методології до основних елементів Text Mining відносяться [8]: класифікація (classification), кластеризація (clustering), побудова семантичних мереж, витяг фактів, понять (feature extraction), реферування (summarization), відповіді на запити (question answering), тематичне індексування (thematic indexing) і пошук за ключо-

вими словами (keyword searching). Також у деяких випадках цей набір доповнюється засобами підтримки та створення таксономії (oftaxonomies), тезаурусів (thesauri) та онтологій (ontology).

При класифікації текстів використовуються статистичні кореляції для створення правил розміщення документів у визначені категорії. Завдання класифікації - це класичне завдання розпізнавання, де за деякою контрольною вибіркою система відносить новий об'єкт до тієї або іншої категорії. Особливість же концепції Text Mining полягає в тому, що кількість об'єктів та їх атрибутів може бути дуже великою - передбачається застосування інтелектуальних механізмів оптимізації процесу класифікації.

Кластеризація базується на ознаках документів, застосуванні лінгвістичних і математичних методів без використання заданих наперед категорій. Результат кластеризації - це таксономія або візуальна карта, що забезпечує ефективне охоплення великих обсягів даних. Кластеризація в Text Mining розглядається як процес виділення компактних підгруп об'єктів із близькими властивостями. Засоби кластеризації дозволяють саме знаходити ознаки й розділяти об'єкти по підгрупах на базі цих ознак. Кластеризація, як правило, передує класифікації, оскільки дозволяє визначити групи об'єктів [9].

Побудова семантичних мереж передбачає аналіз зв'язків, які визначаються, наприклад, появою певних дескрипторів (ключових фраз) у документах, або статистикою спільної появи певних понять у різних документах.

Витяг або екстрагування фактів (понять) призначено для одержання деяких фактів із тексту з метою поліпшення класифікації, пошуку, кластеризації та побудови семантичних мереж.

Автоматичне реферування (Automatic Text Summarization) [10] - це складання коротких викладів матеріалів, анотацій або дайджестів, тобто витяг найбільш важливих відомостей з одного або декількох документів і генерація на їх основі лаконічних, зрозумілих та інформаційно-наповнених звітів.

Існує багато шляхів здійснення автоматичного реферування, які досить чітко діляться на два напрямки - квазіреферування та короткий виклад змісту документів, що базується на застосуванні семантичних методів. Квазіреферування засноване на екстрагуванні фрагментів документів - виділенні найбільш інформативних фраз і формуванні з них квазірефератів.

У рамках квазіреферування виділяють три основних напрямки:

– статистичні методи, засновані на оцінці інформативності різних елементів тексту за частотою появи;

– позиційні методи, які спираються на припущення про те, що інформативність елемента тексту є залежною від його позиції в документі;

–індикаторні методи, засновані на оцінці елементів тексту, виходячи з наявності в них спеціальних слів і словосполучень - маркерів важливості.

Семантичні методи формування рефератів-викладів допускають два основних підходи: метод синтаксичного розбору речень і методи, що базуються на розумінні природної мови. Короткий виклад вихідного матеріалу ґрунтується на виділенні з текстів за допомогою підходів штучного інтелекту та спеціальних інформаційних мов найважливішої інформації та породженні нових текстів, змістовно узагальнюючих первинні документи.

На основі методів автоматичного реферування можливе формування пошукових образів документів. За автоматично побудованими анотаціями великих текстів - пошуковими образами документів - може проводитися пошук, що характеризується високою точністю (природно, за рахунок повноти). Тобто, замість пошуку у повних текстах масиву великих за розміром документів у деяких випадках виявляється доцільним пошук у масиві спеціально створених анотацій. Хоча пошукові образи документів часто виявляються утворенням, що лише віддалено нагадують вихідний текст і не завжди сприймаються людиною, але за рахунок входження найбільш вагомих ключових слів і фраз вони допомагають досягти цілком адекватних результатів при проведенні повнотекстового пошуку.

Теорія складних мереж

Останнім часом усе більшу популярність отримує область дискретної математики, що має назву «теорія складних мереж» (complex networks) [2]. Вона вивчає характеристики мереж з огляду не тільки на їхню топологію, а й статистичні феномени, розподіл ваг окремих вузлів і ребер, ефекти протікання та провідності в таких мережах.

Незважаючи на те, що в розгляд теорії складних мереж потрапляють різні мережі – електричні, транспортні, інформаційні, найбільший внесок у розвиток цієї теорії зробили дослідження соціальних мереж [11]. Термін «соціальна мережа» позначає зосередження соціальних об’єктів, які можна розглядати як мережу, вузли якої - об’єкти, а зв’язки - соціальні відносини. Цей термін було введено в 1954 році соціологом «Манчестерской школи» Дж. Барнсом. У другій половині XX ст. поняття «соціальна мережа» стало популярним серед західних дослідників, при цьому вузлами соціальних мереж почали розглядати не тільки представників соціуму, а й інші об’єкти. Тому сьогодні термін «соціальна мережа» набув широкого змісту. Він включає, наприклад, багато інформаційних мереж, у тому числі

й WWW. Розглядають не тільки статистичні, а й динамічні мережі - для розуміння їх структури необхідно враховувати принципи їх еволюції.

У теорії складних мереж виділяють три основних напрямки:

– дослідження статистичних властивостей, які характеризують поведінку мереж;

– створення моделей мереж;

– прогнозування поведінки мереж при зміні структурних властивостей.

У прикладних дослідженнях застосовують такі типові для мережного аналізу характеристики, як розмір мережі, мережна щільність, ступінь центральності тощо. При цьому досліджуються:

– параметри окремих вузлів;

– параметри мережі в цілому;

– мережні підструктури.

Для окремих вузлів виділяють такі параметри:

– вхідний ступінь вузла - кількість ребер графа, які входять у вузол;

– вихідний ступінь вузла - кількість ребер графа, які виходять із вузла;

– відстань від даного вузла до кожного з інших;

– середня відстань від даного вузла до інших;

– ексцентричність (eccentricity) - найбільша з геодезичних відстаней (мінімальних відстань між вузлами) від даного вузла до інших ;

– посередництво (betwetnness), що показує, скільки найкоротших шляхів проходить через даний вузол;

– центральність (загальна кількість зв'язків даного вузла відносно інших).

Для розрахунку характеристик мережі в цілому використовують такі параметри:

– число вузлів;

– число ребер;

– геодезична відстань між вузлами (мінімальна відстань між вузлами мережі);

– середня відстань від одного вузла до інших;

– щільність (обчислюється як відношення кількості ребер у мережі до можливої максимальної кількості ребер при даній кількості вузлів);

– кількість симетричних, транзитивних і циклічних тріад;

– діаметр мережі (найбільша геодезична відстань у мережі).

Існує декілька актуальних завдань дослідження складних мереж:

– визначення клік у мережі. Кліки - це підгрупи або кластери, у яких вузли зв'язані між собою сильніше, ніж зі членами інших клік;

– виділення компонент (частин мережі), які зв'язані усередині й не зв'язані між собою;

– знаходження блоків і перемичок. Вузол називається перемичкою, якщо при його вилученні мережа розпадається на незв'язані частини;

виділення групувань - груп еквівалентних вузлів (які мають максимально подібні профілі зв'язків).

Важливою характеристикою мережі є функція розподілу ступенів вузлів $P(k)$, яка визначається як ймовірність того, що вузол i має ступінь $k_i = k$. Мережі, що характеризуються різними $P(k)$, демонструють різноманітну поведінку. $P(k)$ у деяких випадках може бути розподілом Пуассона ($P(k) = e^{-m} m^k / k!$, де m – математичне очікування), експонентним ($P(k) = e^{-k/m}$) або степеневим розподілом ($P(k) \sim 1/k^{\gamma}$, $k \neq 0$, $\gamma > 0$).

Відстань між вузлами визначається як кількість кроків, які необхідно зробити, щоб по існуючих ребрах дістатися від одного вузла до іншого. Природно, вузли можуть бути з'єднані прямо або опосередковано. Шляхом між вузлами l назвемо найкоротшу відстань між ними. Для всієї мережі можна ввести поняття середнього шляху - як середню по всіх парах вузлів найкоротшу відстань між ними:

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

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

Угорськими математиками П. Ердьошем та А. Реньі було показано, що середня відстань між двома вершинами у випадковому графі зростає як логарифм від числа вершин [12, 13].

Деякі мережі можуть виявитися незв'язними, тобто знайдуться вузли, відстань між якими виявиться нескінченною. Для врахування таких випадків уводиться 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, який дорівнює максимальному значенню з усіх d_{ij}.

Д. Уаттс і С. Строгатц у 1998 р. визначили такий параметр мереж, як коефіцієнт кластерності [14], що відповідає рівню зв'язності вузлів у мережі. Цей коефіцієнт характеризує тенденцію до утворення груп взаємозалежних вузлів, так званих клік (clique). Крім того, для конкретного вузла коефіцієнт кластерності показує, скільки найближчих сусідів даного вузла є також найближчими сусідами один для одного.

Коефіцієнт кластерності може визначатися як для кожного вузла, так і для всієї мережі. Відповідно, рівень кластерності всієї мережі визначається як нормована за кількістю вузлів сума відповідних коефіцієнтів для окремих вузлів.

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

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

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.

Властивість еластичності мереж відноситься до розподілу відстаней між вузлами при вилученні окремих вузлів. Еластичність мережі залежить від її зв'язності, тобто від існування шляхів між парами вузлів. Якщо вузол буде вилучений з мережі, типова довжина цих шляхів збільшиться. Якщо цей процес продовжувати досить довго, мережа перестане бути зв'язаною. Р. Альберт (Réka Albert) з університету штату Пенсільванія, США, при дослідженні атак на інтернет-сервери вивчала ефект вилучення вузла мережі, що представляє собою підмножину WWW з 326000 сторінок [15]. З'ясувалося, що середня відстань між двома вузлами, як функція від кількості вилучених вузлів, майже не змінилася при випадковому видаленні вузлів (висока еластичність). Разом з тим цілеспрямоване видалення вузлів з найбільшою кількістю зв'язків призводить до руйнування мережі. Таким чином, Інтернет є високоеластичною мережею відносно випадкової відмови окремих вузлів, але високочутливою до навмисної атаки на вузли з високими ступенями зв'язків.

Про «структуру співтовариства» можна говорити тоді, коли існують групи вузлів, які мають високу щільність ребер між собою, при тому, що щільність ребер між окремими групами - низька. Традиційний метод для виявлення структури співтовариств - кластерний аналіз. Для великих складних мереж наявність структури співтовариств виявилася невід'ємною властивістю. Водночас до таких властивостей відносяться й так називані «слабкі» зв'язки. Аналогом слабких соціальних зв'язків є, наприклад, відносини з далекими знайомими й колегами. У деяких випадках ці зв'язки виявляються ефективнішими, ніж зв'язки «сильні». Так, нещодавно був отриманий концептуальний висновок в області мобільного зв'язку, що «слабкі» соціальні зв'язки між індивідуумами виявляються найважливішими для існування соціальної мережі [16]. Для дослідження були проаналізовані дзвінки 4,6 млн. абонентів мобільного зв'язку, що становить близько 20% населення однієї європейської країни. Це був перший випадок у світовій практиці, коли вдалося одержати й проаналізувати таку велику вибірку даних, що відносяться до міжособистісної комунікації. У цій соціальній мережі було виявлено 7 млн. соціальних зв'язків, тобто взаємних дзвінків від одного абонента іншому й назад, якщо зворотні дзвінки були зроблені протягом 18 тижнів. Частота та тривалість розмов використовувалися для того, щоб визначити вагу кожного зв'язку. Було виявлено, що саме слабкі зв'язки (один-два зворотних дзвінки протягом 18 тижнів) зв'язують воєдино велику соціальну мережу. Якщо ці зв'язки проігнорувати, то мережа розпадеться на окремі фрагменти. Якщо ж не враховувати сильних зв'язків, то зв'язність мережі не порушиться (рис. 1). На підставі проведених досліджень був зроблений висновок, що саме слабкі зв'язки є тим феноменом, що зв'язує мережу в єдине ціле.

Незважаючи на величезні розміри деяких реальних складних мереж, у багатьох з них (і в WWW зокрема) існує порівняно короткий шлях між двома будь-якими вузлами – геодезична відстань. У 1967 р. психолог С. Мілгран у результаті проведених масштабних експериментів обчислив, що існує ланцюжок знайомств, у середньому довжиною шість практично між двома будь-якими громадянами США [17]. Д. Уаттс і С. Строгатц виявили феномен, характерний для багатьох реальних мереж, названий ефектом малих світів (Small Worlds) [14]. При дослідженні цього феномена ними була запропонована процедура побудови наочної моделі мережі, якій властивий цей феномен. Три стани

1 2 3

Рис. 1. Структура мережі: 1) повна карта мережі соціальних комунікацій; 2) соціальна мережа, з якої вилучені слабкі зв'язки; 3) мережа, з якої вилучені сильні зв'язки: структура зберігає зв'язність

цієї мережі представлені на рис. 2: регулярна мережа - кожен вузол якої з'єднаний із чотирма сусідніми, та ж сама мережа, у якої деякі «ближні» зв'язки випадковим чином замінені «далекими» (саме в цьому випадку виникає феномен «малих світів»), а також випадкова мережа, в якій кількість подібних замін перевищила деякий поріг.

На практиці виявилося, що саме ті мережі, вузли яких мають одночасно деяку кількість локальних і випадкових «далеких» зв'язків, демонструють одночасно ефект малого світу й високий рівень кластерності.

На рис. 3 наведені графіки зміни середньої довжини шляху та коефіцієнта кластерності штучної мережі Д. Уаттса й С. Строгатца від імовірності встановлення «далеких зв'язків». WWW є мережею, для якої також підтверджений феномен малих світів. Аналіз топології Мережі, проведений Ши Жоу та Р. Дж. Мондрагоном з Лондонського університету, показав, що вузли з великим ступенем вихідних гіперпосилань мають більше зв'язків між собою, ніж із вузлами з малим ступенем, тоді як останні мають більше зв'язків із вузлами з великим ступенем, ніж між собою. Цей феномен був названий «клубом багатіїв» (rich-club phenomenon). Дослідження показало, що 27% усіх з'єднань мають місце між усього 5% найбільших вузлів, 60% доводиться на з'єднання інших 95% вузлів з 5% найбільших і тільки 13% - це з'єднання між вузлами, які не входять у лідируючі 5%. Ці дослідження дають підстави вважати, що залежність мережі WWW від великих вузлів значно істотніша, ніж передбачалося раніше, тобто вона ще більш чутлива до зловмисних атак. З концепцією «малих світів» пов'язаний також практичний підхід, що отримав назву «мережна мобілізація», яка реалізується над структурою «малих світів». Зокрема, швидкість поширення інформації завдяки ефекту «малих світів» у реальних мережах зростає на порядки в порівняно з випадковими мережами, адже більшість пар вузлів реальних мереж з'єднані короткими шляхами. Експертами з безпеки ефект «малих світів» останнім часом все частіше пов'язується з мережами терористичних організацій, надбудованими поверх Інтернет.

Аналізуючи зв’язки в мережі, можна довідатися про її важливі властивості, наприклад, виявити наявність кластерів, визначити їх склад, розходження у зв’язності усередині та між кластерами, ідентифікувати ключові елементи, що зв'язують кластери між собою тощо. Разом з тим серйозною перешкодою при аналізі є неповна інформація про зв'язки між окремими вузлами мережі. Нещодавно група дослідників з Інституту Санта Фе (Santa Fe Institute) представила алгоритм, за допомогою якого стає можливим автоматичне отримання інформації про ієрархічну структуру подібних мереж [18]. Новий метод відновлення мереж можуть взяти на озброєння різні спецслужби. Так, знаючи, наприклад, лише про половину зв'язків між терористами, можна буде з високою ймовірністю відновити відсутні ланки всього ланцюжка. Навіть не маючи повного опису системи, можна одержувати репрезентативну вибірку зв'язків і по ній намагатися добудовувати всю мережу. Аналіз графа, що вийшов, дозволяє виявити потенційно важливі зв'язки, які не вдалося виявити в реальній системі. Маючи інформацію лише про половину контактів терористів між собою, можна з імовірністю 0,8 прогнозувати ті зв'язки, про які спочатку нічого не було відомо. Очевидно, що даний метод може надати важливу допомогу в справі виявлення прихованих мережних організацій, і таким чином поставити справу забезпечення державної й міжнародної безпеки на якісно новий рівень.

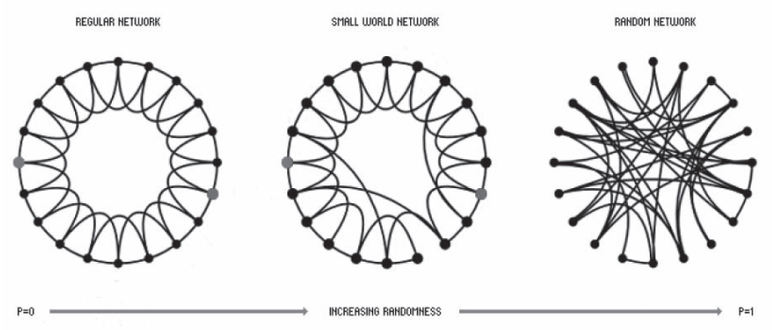

Рис. 2. Модель Д. Уаттса та С. Строгатца

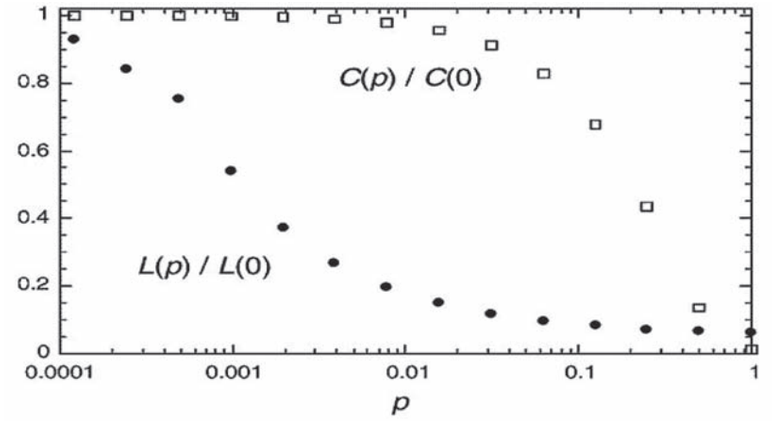

Рис. 3. Динаміка зміни довжини шляху та коефіцієнта кластерності

Нещодавно було показано, що найбільшу інформаційну провідність має особливий клас мереж, називаних «заплутаними» (entangled networks). Вони характеризуються максимальною однорідністю, мінімальною від-

станню між будь-якими двома вузлами та дуже вузьким спектром основних статистичних параметрів.

При вивченні «малих світів» визначився цікавий підхід, логічно пов'язаний з поняттям перколяції (протікання) [19, 20]. Виявляється, що багато питань, що виникають при аналізі мережної безпеки в Інтернет, безпосередньо відносяться до цієї теорії. Найпростіше очищене від усіх фізичних і математичних нашарувань формулювання задачі теорії перколяції має такий вигляд: «Дана сітка зі зв'язків, випадкова частина якої проводить сигнал, а інша частина його не проводить. Основне питання - чому дорівнює мінімальна концентрація провідних зв'язків, при якій ще існує шлях через усю сітку?». До задач, які розв'язуються у рамках теорії перколяції та аналізу мереж відносяться такі, як визначення граничного рівня провідності, зміни довжини шляху та його траєкторії при наближенні до граничного рівня провідності, кількості вузлів, які необхідно вивести з ладу, щоб порушити зв'вязність мережі.

Системи Text Mining

Унікальними особливостями концепції й технологій Text Mining є те, що з їхньою допомогою можна добувати з «сирих» даних неочевидні, корисні на практиці та доступні для інтерпретації знання, необхідні для прийняття рішень у різних сферах діяльності, в тому числі у сфері національної безпеки.

Одним із перших розсекречених комплексів, що реалізує підхід Text Mining, була французька система «TAIGA» (Traitement automatique d'information geopolitique d'actualite - автоматична система обробки актуальної геополітичної інформації). Цей програмний комплекс протягом 11 років застосовувався в інтересах французької розвідки, після чого був замінений на новіший, розсекречений і дозволений до комерційного використання. Новий комплекс «Noemic», поставлений на озброєння французької розвідки може обробляти інформацію зі швидкістю понад 1 млрд. знаків за секунду. Американський аналог цих програмних комплексів «Topic» також розсекречений.

Аналогічні системи створювались і в колишньому Радянському Союзі. Досить згадати такі системи як «Барометр», «Ельбрус», були й інші. На сучасному ринку представлений ряд як західних продуктів, так і систем виробництва пострадянських країн, здатних у тому або іншому обсязі здійснювати глибинний аналіз текстів.

Останнім часом всі основні західні бренди, що спеціалізуються на розробці інформаційних сховищ і баз даних, корпоративних системах керування, розширили свої лінійки продуктів системами або модулями Text Mining. Про наявність таких модулів заявляють SAP, Oracle, SAS, IBM та інші компанії.

Нині добре відома система Lotus Discovery Server фірми IBM – програмний продукт, призначений для керування знаннями в корпоративних порталах. Система знаходить та ідентифікує зв'язки між інформацією, людьми та їхньою діяльністю.

Російські розробники з «Інфорус» створили систему Avalance, що у процесі пошуку формує модель предметної області у вигляді набору «розумних папок», кожна з яких «знає», що до неї має потрапити. Наповненням папок займається спеціалізований робот, що запускається з комп'ютера «хазяїна» і «приносить» тільки те, що просили.

Серед найрозвиненіших систем керування знаннями можна назвати також систему Hummingbird Enterprise™ канадської компанії Hummingbird. Серед компонентів системи необхідно виділити Hummingbird Portal - платформу, що дозволяє інтегрувати інформацію з інформаційного сховища та застосування в єдиному Web-інтерфейсі.

Нижче наведено перелік деяких відомих систем глибинного аналізу тексту.

Intelligent Miner for Text (IBM)

Система є одним із кращих інструментів глибинного аналізу текстів, яка складається з окремих утиліт:

– Language Identification Tool - утиліта визначення мови, якою складений документ.

– Categorisation Tool - утиліта класифікації - автоматичного віднесення тексту до деякої категорії (вхідною інформацією на навчальній фазі роботи цього інструмента може бути результат роботи наступної утиліти - Clusterisation Tool).

– Clusterisation Tool - утиліта кластеризації - розбивки великої множини документів на групи за близькістю стилю, форми, різних частотних характеристик ключових слів, що виявляються.

– Feature Extraction Tool - утиліта визначення нового - виявлення в документі нових ключових слів (власні імена, назви, скорочення) на основі аналізу заданого заздалегідь словника.

– Annotation Tool - утиліта «виявлення змісту» текстів і складання рефератів - анотацій.

PolyAnalyst, WebAnalyst (Мегап'ютер Інтелідженс)

PolyAnalyst може застосовуватися для автоматизованого аналізу числових і текстових даних з метою виявлення раніше невідомих, нетривіальних, корисних і доступних розумінню закономірностей.

PolyAnalyst є клієнт-серверним застосуванням. При цьому користувач працює із програмою PolyAnalyst Workplace.

Математичні модулі виділені в серверну частину - PolyAnalyst Knowledge Server. PolyAnalyst працює з різними типами даних. Це - числа, логічні змінні, текстові рядки, дати, а також вільний текст.

PolyAnalyst може обробляти вихідні дані з різних джерел, приміром, файли Microsoft Excel, ODBC-сумісних СУБД, SAS data files, Oracle Express, IBM Visual Warehouse. До складу PolyAnalyst входить система TextAnalyst, що вирішує такі задачі Text Mining: створення семантичної мережі великого тексту, підготовка резюме тексту, пошук по тексту та автоматична класифікація і кластеризація текстів. Побудова семантичної мережі - це пошук ключових понять тексту й установлення взаємин між ними.

Text Miner (SAS)

Система SAS Text Miner може працювати з текстовими документами різних форматів із баз даних, файлових систем та Web. Text Miner забезпечує логічну обробку тексту в середовищі потужного пакета SAS Enterprise Miner. Це дозволяє інтегрувати текстову інформацію зі структурованими даними.

SemioMap (Semio Corp.)

SemioMap 2.0 - це перша система Text Mining, що працювала в архітектурі клієнт-сервер. Система SemioMap складається із двох основних компонентів - сервера SemioMap і клієнта SemioMap. Працює система у три фази:

– індексування: сервер SemioMap автоматично зчитує масиви неструктурованого тексту, витягає ключові фрази (поняття) і створює з них індекс;

– кластеризація понять: сервер SemioMap виявляє зв'язки між витягнутими фразами та на основі спільної появи будує з них лексичну мережу («понятійну карту»);

– графічне відображення та навігація: візуалізація карт зв'язків, що забезпечує швидку навігацію за ключовими фразами та зв'язкам між ними, а також можливість швидкого звернення до конкретних документів.

Oracle Text (Oracle)

Засоби Text Mining, починаючи з Text Server у складі СУБД Oracle 7.3.3 і картриджу interMedia Text в Oracle8i, є невід'ємною частиною продуктів Oracle. В Oracle9i ці засоби розвилися та одержали нову назву - Oracle Text. Основною задачею, на вирішення якої вони націлені, є пошук документів за їхнім змістом - словами або фразами, які при необхідності комбінуються з використанням булевих операцій. Результати пошуку ранжируються за релевантністю з урахуванням частоти появи слів запиту в знайдених документах. Система Oracle Text забезпечує проведення тематичного аналізу текстів, представлених англійською мовою. Під час обробки текст кожного документа піддається процедурам лінгвістичного та статистичного аналізу, у результаті чого визначаються його ключові теми та будується резюме.

Autonomy

Основна перевага системи Autonomy - інтелектуальні алгоритми, засновані на статистичній обробці, інформаційній теорії Шенона, байєсовських ймовірностях і нейронних мережах.

Про функціональні можливості пропозицій від Autonomy можна судити з переліку основних напрямків, що сформувався в результаті розвитку програмних продуктів власної розробки й тих, що були отримані разом із придбаними компаніями: IDOL 7 Enterprise Search (корпоративний пошук); ZANTAZ IDOL for Managing Risk (керування ризиками); Varage IDOL forb Audio & Video Search (пошук в аудіо- та відеофайлах); etalk IDOL for Call Center CRM (підтримка контакт-центрів і систем керування відносинами із клієнтами); CARDIFF IDOL for Business Process Management (керування бізнес-процесами); meridio IDOL for Record Management (керування записами).

Autonomy включає такі основні можливості, як автоматична класифікація, кластеризація, автоматичне реферування, автоматичне встановлення гіперпосилань, автоматичне створення профайлів (інформаційних портретів), генерація таксонометричних дерев, створення й маніпулювання метаданими, інтелектуальна обробка XML-даних, пошук тощо.

RetrievalWare (Convera)

RetrievaWare - система повнотекстового й атрибутивного пошуку. До документів, з якими здатна працювати ця система, відносяться тексти в різних кодуваннях і форматах (понад 200). Позиціонується як система видобутку знань (Knowledge Mining).

Galaktika-ZOOM (корпорація «Галактика»)

Основне призначення Galaktika-ZOOM - інтелектуальний пошук за ключовими словами з урахуванням морфології, а також формування інформаційних портретів конкретних понять. Орієнтація на великі інформаційні об'єкти. Система містить інструментарій для аналізу змістовних зв'язків і формування «образу» проблеми - багатомірної моделі у формі списку значимих словосполучень. Система містить інструментарій для виявлення тенденцій та динаміки появи понять.

InfoStream
(Інформаційний центр «Електронні вісті»).

Система забезпечує:

– доступ до оперативної інформації (понад 3000 джерел) з єдиного інтерфейсу в пошукових режимах з урахуванням можливого дублювання та семантичної близькості документів, мовних версій, розмірів документів, їхньої цифрової насиченості тощо;

– доступ до унікального ретроспективного фонду, що перевищує 60 млн. записів;

– підтримку аналітичної роботи у режимі реального часу: побудову сюжетних ланцюжків, дайджестів, діаграм появи й таблиць взаємозв'язків понять, медіа-рейтингів тощо.

Запити до пошукових систем

На прикладі застосування системи інтеграції та моніторингу інформаційних ресурсів InfoStream [21] покажемо, як формуються запити до пошукових систем з елементами Text Mining для підтримки аналітичної діяльності в галузі економічної безпеки.

Як приклад назвемо ряд проблем, а потім поставимо їм у відповідність фрагменти запитів і розглянемо фрагменти текстів, що публікуються різними джерелами, які надалі можна використовувати при побудові різного роду аналітичних довідок. Нижче наведені уточнюючі запити, що відносяться до фінансового становища окремих компаній:

(Статутний~капітал)грн.
(Статутний~фонд)дол.
Фінансове~становище
належить~/2/акцій

(У системі InfoStream "~" - оператор контекстної близькості, "~/2/" - близькість на відстані не більше двох слів).

У результаті обробки запиту були отримані тексти, що містять такі фрагменти:

По данным НБУ, к началу 2008 года доля иностранного капитала в совокупном **уставном капитале** украинских банков составляла 35%. Общее количество действующих в стране банков с иностранным капиталом в апреле не изменилось, и к началу мая их насчитывалось 47. Количество банков со 100%-ным иностранным капиталом в апреле также не изменилось: к началу мая действовало 17 таких финучреждений. «Украинский банковский портал» 2008.06.20.

Акціонери Авдєєвського коксохімічного заводу схвалили допемісію на 2,566 млн. грн. Таке рішення ухвалене сьогодні на загальних зборах акціонерів підприємства, повідомили в прес-службі «Метінвест-холдингу». Під час допемісії будуть випущені 1,458 млн. простих іменних акцій номінальною вартістю 1,76 грн. Акції розміщуватимуться у формі закритого розміщення виключно серед акціонерів ОАО «АКХЗ» у два етапи. При цьому **статутний фонд** Авдіївського КХЗ збільшиться до 343,31 млн. грн. «РБК-Україна» 2008.06.19.

Президент «Енергокомпанії України» вважає, що для поліпшення **фінансового становища** в енергокомпанії міська влада повинна підвищувати тарифи на опалення й гаряче водопостачання, які забезпечують столичні ТЕЦ. «Енергокомпанії України» належить 50%+1 акція «Київенерго». Держхолдинг також володіє 100% акцій ВАТ «Дніпродзержинська ТЕЦ», ВАТ «Миколаївська ТЕЦ», ВАТ «Харківська ТЕЦ-5», ВАТ «Херсонська ТЕЦ» і ВАТ «Одеська ТЕЦ». «Економічна правда» 2008.06.18.

Группа SCM объединяет более 90 компаний. Доминирующее положение в группе, как по объемам производства, так и по объемам прибыли, занимают предприятия тяжелой промышленности, однако осуществляемая стратегия реструктуризации бизнеса группы предполагает увеличение доли постиндустриальных направлений - финансового, телекоммуникационного, медиа и других - в общей структуре бизнеса группы SCM. Бизнесмену Ринату Ахметову **принадлежит 90% акций** ЗАО SCM. Fin.org.ua 2008.06.20

Інформація про злиття та придбання в тій або іншій сфері бізнесу може бути отримана в результаті відпрацьовування таких запитів:

Придбав~/2/акцій
приобрел~/2/пакет~акций
продал~/2/пакет~акций
слияни & компан & (акци, актив)

Отримано:

ОАО «Шумский маслозавод» (Тернопольская обл.) **приобрело 88,1% акций** ОАО «Городоцкая молочная компания «Белая роса» (Львовская обл.). Ранее этот пакет акций принадлежал физлицу. По данным ГКЦБФР, 90,8% акций ОАО «Шумский маслозавод» принадлежит ООО «Галмолторг» (Львов). ОАО «Городоцкая молочная компания «Белая роса» (ранее - Городоцкий молочный завод) закончило 2007 г. с чистым убытком 582 тыс. грн. против 489 тыс. грн. в 2006 г. «Экономические известия» 2008.06.20.

Переможцем конкурсу з **продажу 61,32% акцій** ВАТ «Львівський завод радіоелектронної медичної апаратури» («ЛЗ РЕМА», Львів) стало ВАТ «Іскра» (м. Львів). Як повідомили в прес-службі регіонального відділення Фонду державного майна по Львівській області, пакет акцій проданий за 1,892 млн грн при стартовій ціні 1,85 млн грн. «Економічні новини-Львів» 2008.06.18.

В январе 2006 года Луцкий автомобильный завод **продал 8% своих акций** иностранным инвесторам за $16 млн. на Франкфуртской фондовой бирже. InvestfundS.UA 2008.06.19

Вступ України до СОТ суттєво вплине на обсяги **злиття та поглинань**, прогнозують експерти. Причому хвиля концентрацій охопить не тільки іноземних покупців, а й українські компанії, які об'єднуватимуться, щоб захистити своє місце на ринку. У зв'язку з цим зростатиме попит на фінансових директорів, аудиторів та інших представни-

ків подібних професій. «Український Бізнес Ресурс» 2008.06.19.

Для виявлення публікацій про зміну фінансового стану та банкрутства можна використати такі уточнюючі запити:

выпуск~/2/акций
(увеличить~уставной)&(фонд,капитал)
продать~/2/акций
оголосити~/2/банкрутство

Відпрацьовування цих запитів дозволило знайти такі документи:

Акціонери Авдіївського коксохімічного заводу (ВАТ АКХЗ) на Загальних зборах акціонерів, що відбулися 19 червня, вирішили **збільшити статутний фонд** підприємства на 2 566 080 гривень до 343 310 000 гривень шляхом додаткового випуску акцій. Під час додаткової емісії буде випущено 1 458 000 простих іменних акцій існуючою номінальною вартістю 1,76 гривень. Акції будуть розміщатися у формі закритого (приватного) розміщення виняткова серед акціонерів ВАТ АКХЗ у два етапи. «Систем Кепітал Менеджмент» 2008.06.20.

Накопичені запитання щодо стану справ в Ощадбанку «ДТ» адресувало його голові правління Анатолію Гулею. - Анатолію Івановичу, чи вважаєте ви кошти, закладені в бюджеті на **збільшення статутного капіталу** Ощадбанку, достатніми для забезпечення його належного розвитку? - У нинішньому варіанті бюджету на ці цілі передбачено 200 млн. грн., причому відповідно до бюджетного розпису вони мають бути виділені до кінця другого кварталу. Тому ми сподіваємося до 1 липня ці двісті мільйонів одержати. «Дзеркало тижня» 2008.06.21.

Політ із Варшави до країн Середземного моря подорожчав на 100-200 злотих (45-90 доларів). У «чорному списку» дискаунтерів, які підвищують ціни, - «Центральвінгс» (доплата за пальне від 30 до 100 злотих), «Джерманвінгс» (запровадила доплату за багаж), світовий лідер «лоу кост» перевезень «Райнейр» (30 відсотків - ріст цін за багаж і відправлення). Експерт польського повітряного ринку Ерик Клопотовський заявив: «Епоха польотів за один злотий уже закінчилася». До речі, «Райнейр» та британський дискаунтер «Ізі Джет» відзвітували про збитки в першому кварталі 2008 року. Ще чотири «лоу кости» у квітні **оголосили про власне банкрутство**. Газета «Україна молода» 2008.06.18

Понятия в динамике :
(газ~криз) & geo.UA*

Формат: Обычный | Группировать по: дням | - позитив | - негатив

Дата	Всего	Понятие
2008.02.01	51292	2
2008.02.02	14524	0
2008.02.03	10231	1
2008.02.04	52725	7
2008.02.05	52531	2
2008.02.06	53253	4
2008.02.07	54365	2
2008.02.08	51774	90
2008.02.09	15007	13
2008.02.10	10593	7
2008.02.11	52705	27
2008.02.12	53869	67
2008.02.13	54111	41
2008.02.14	55341	19
2008.02.15	53170	21
2008.02.16	14700	1
2008.02.17	11106	0
2008.02.18	53614	14
2008.02.19	54873	9
2008.02.20	55544	15
2008.02.21	55753	7
2008.02.22	52129	4
2008.02.23	13922	1
2008.02.24	10354	1
2008.02.25	34472	11
2008.02.26	54835	17
2008.02.27	55186	30
2008.02.28	55352	15
2008.02.29	53239	11
2008.03.01	15300	3
2008.03.02	18098	6
2008.03.03	54420	131
2008.03.04	55242	272
2008.03.05	55418	184
2008.03.06	55675	97
2008.03.07	48387	56

Рис. 4. Динаміка появи понять

Методи контент-моніторингу

Методи контент-моніторингу - це адаптація класичних методів контент-аналізу до умов динамічних інформаційних масивів, наприклад, потоків інформації з Інтернет. Типове завдання контент-моніторингу - побудова діаграм динаміки появи понять у часі. Розглянемо, як у системі InfoStream відслідковувалися публікації щодо газової кризи у лютому-березні 2008 року. Для цього був складений запит ***«газ~криз & geo.UA»***, який було введено через Web-інтерфейс системи (рис. 4).

На прикладі ринку нафтопродуктів розглянемо, як з масивів текстової інформації з Інтернет можуть бути виявлені сюжетні ланцюжки з документів, що містять максимальну кількість цінової інформації з даного ринку.

Для одержання основних сюжетів, що відносяться до ринку нафтопродуктів, було введено запит ***«(нафтопродукти | бензин) & ціни»***, який було уточнено спеціальними ознаками ***«numb.medium | numb.large»***, що в системі InfoStream означає середній або високий рівень присутності в документах цифрової інформації (рис. 5).

Після цього можна перейти до режиму «Сюжети» і проаналізувати документи та посилання, видані системою. Режим «Сюжети» не тільки передбачає обробку запитів в рамках булевої моделі пошуку, а й додає врахування вагових критеріїв, видаючи лише найбільш вагомі ланцюжки документів. Тому забезпечується досить високий рівень відповідності документів та інформаційної потреби, вираженої запитом.

Перспективними напрямками технологій Text Mining є автоматичний витяг понять із неструктурованих текстів, а також побудова таблиць взаємозв'язків і гістограм розподілу понять.

Аналіз та візуалізації складних мереж

Одним із напрямків аналізу соціальних мереж є візуалізація, яка має важливе значення, оскільки найчастіше дозволяє робити важливі висновки щодо характеру взаємодії вузлів, не вдаючись до точних методів аналізу. При відображенні моделі соціальної мережі доцільним може бути:

- розміщення вузлів мережі у двох вимірах;
- просторове впорядкування об'єктів в одному вимірі відповідно до деякої кількісної властивості;
- використання загальних для всіх мережних діаграм методів для відображення кількісних і якісних властивостей об'єктів і відносин.

Як приклади візуалізації мереж можна розглянути деякі розробки компанії TouchGraph. Так, TouchGraph Amazon відображає мережу, породжену книгами та зв'язками між ними (за тематиками, авторами, видавництвами). TouchGraph також реалізувала інтерфейс для побудови соціомережі на основі Livejournal - TouchGraph LiveJournal Browser. У випадку візуалізації WWW засобами TouchGraph Google Browser (http://www.touchgraph.com/TG GoogleBrowser.html) ребрами виступають не гіперпосилання, а відносини подібності. Google Browser представляє собою Java-аплет, що дозволяє візуалізувати зв'язки подібності між веб-сайтами, що розраховуються в пошуковій системі Google. У цьому інтерфейсі можна побачити всі сайти, зв'язані відношенням подібності з вихідним заданим, при цьому користувач може задавати глибину зв'язків і відображати взаємозв'язки різних сайтів. TouchGraph Google Browser - досить корисний інструмент також при пошуку сайтів, пов'язаних з вихідним загальною тематикою (рис. 6).

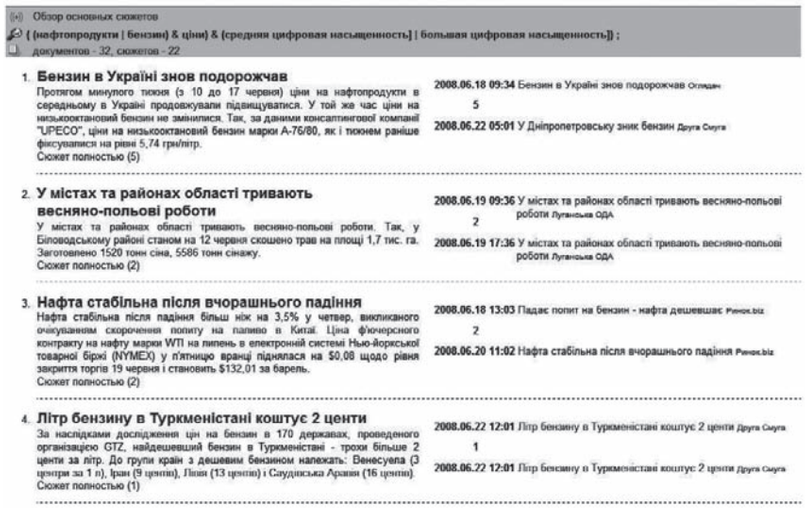

Рис. 5. Ланцюжок основних сюжетів

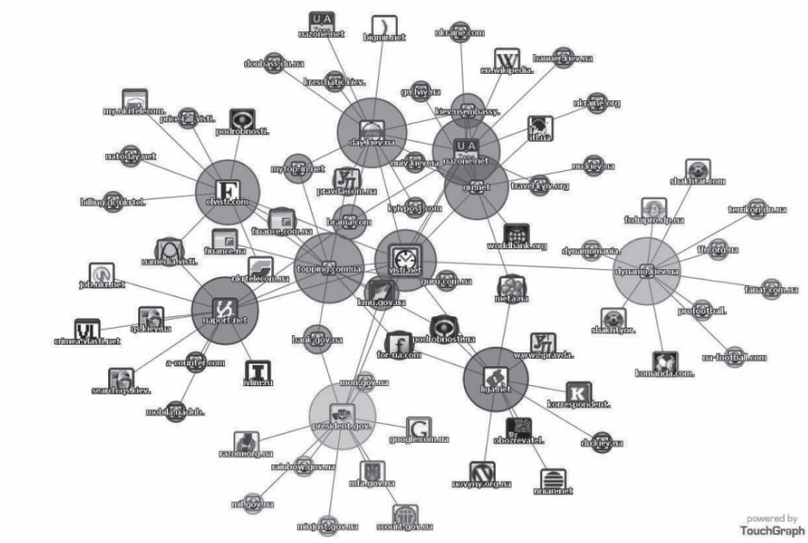

Рис. 6. Відображення зв'язків веб-серверів (від TouchGraph)

У якості інструмента для аналізу та візуалізації соціальних мереж можна навести програму NetVis (http://www.netvis.org), що використовує online-дані та імпортовані файли. Також широко відомі програми візуалізації та аналізу соціальних/організаційних мереж InFlow (поточна версія 3.1 доступна за адресою http://www.orgnet.com/inflow3.html) і система аналізу соціальних мереж UCINET (http://www.analytictech.com/ucinet/ucinet.htm) з інтегрованою до неї вільно розповсюджуваною програмою візуалізації NetDraw.

Як приклад застосування можливостей теорії складних мереж наведемо фрагмент дослідження мережі зв'язків понять (прізвищ персон), що екстрагуються з корпусів неструктурованих текстів. Як такі корпуси текстів використовувалися масиви документів, що скануються з Інтернет системою InfoStream.

При побудові мережі понять використовувалися алгоритми автоматичного витягу понять із неструктурованих текстів. Слід зазначити, що підходи до витягу різних типів понять із текстів істотно відрізняються як за контекстом їхнього подання, так і за структурними ознаками. Так, для виявлення приналежності документа до тематичної рубрики можуть використовуватися спеціальним чином складені запити, що включають логічні й контекстні оператори, дужки тощо. Виявлення географічних назв припускає використання таблиць, у яких, крім шаблонів написання цих назв, використовуються коди країн, назви регіонів і населених пунктів. Ще один вид понять, такий як «персони», екстрагується з текстів на підставі правил, що враховують таблиці припустимих імен і прізвищ, шаблони ініціалів, можливі варіанти спільного написання ініціалів/імен і прізвищ. Слід зазначити, що система InfoStream включає засоби екстрагування понять та, серед іншого, надає результати користувачам у вигляді «інформаційних портретів», які включають такі поняття, як ключові слова, географічні назви, прізвища персон, назви фірм тощо. Надалі описуватимуть властивості мереж, утворених поняттями, які зв'язані один з одним згадуваннями у тих самих документах. Більш конкретно досліджувалася мережа, утворена прізвищами персон, що витягаються з повідомлень українських інтернет-ЗМІ за загальнополітичною тематикою за 1 місяць обсягом 55 тис. документів. Усього в текстах згадувалося понад 19 тис. персон.

Мережа, утворена поняттями, що витягаються з потоків текстів, не є статичною, а залежить від обсягів документів, з яких витягаються відповідні поняття. Отже, для розуміння структури такої мережі необхідно враховувати її еволюцію.

Ребрам вихідної мережі приписуються вагові значення, рівні кількості документів, у яких зустрічаються згадування персон, що відповідають вузлам. Для запобігання «шуму» ребра з вагою, меншою ніж 2, не враховувалися. При розвитку мережі з фіксованої кількості персон, при збільшенні кількості розглянутих документів, середня відстань між вузлами відповідно зменшується, досягаючи свого логічного насичення.

Одним із результатів, отриманих у рамках даного дослідження, є встановлення того факту, що вузли розглянутої мережі персон із максимальною кількістю вихідних ребер переважно мають найбільший рівень посередництва (рис. 7), що не дозволяє розглядати їх як основи для побудови кластерів при автоматичному групуванні, а скоріше як елементи, що з'єднують окремі групи персон.

У розглянутій мережі при фіксованій кількості вузлів, що відповідають персонам, і зростаючій кількості документів розподіл ступенів вузлів виявився спочатку близьким до степеневого, а потім - до пуасонівського (рис. 8). Це зумовлюється тим, що спочатку ступені вузлів мають систематичний характер, який відповідає реальним зв'язкам, а потім за раху-

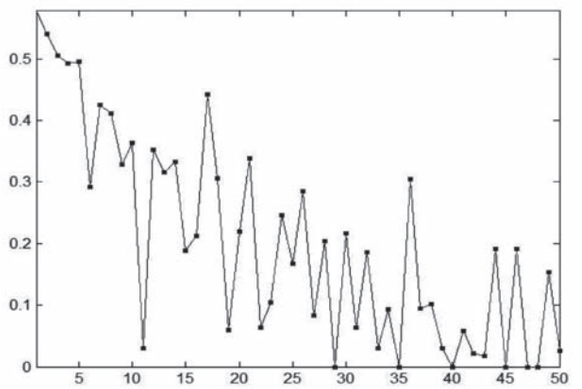

Рис. 7. Коефіцієнти посередництва (вісь ординат) для вузлів, що ранжирувані за ступенем (кількістю зв'язків)

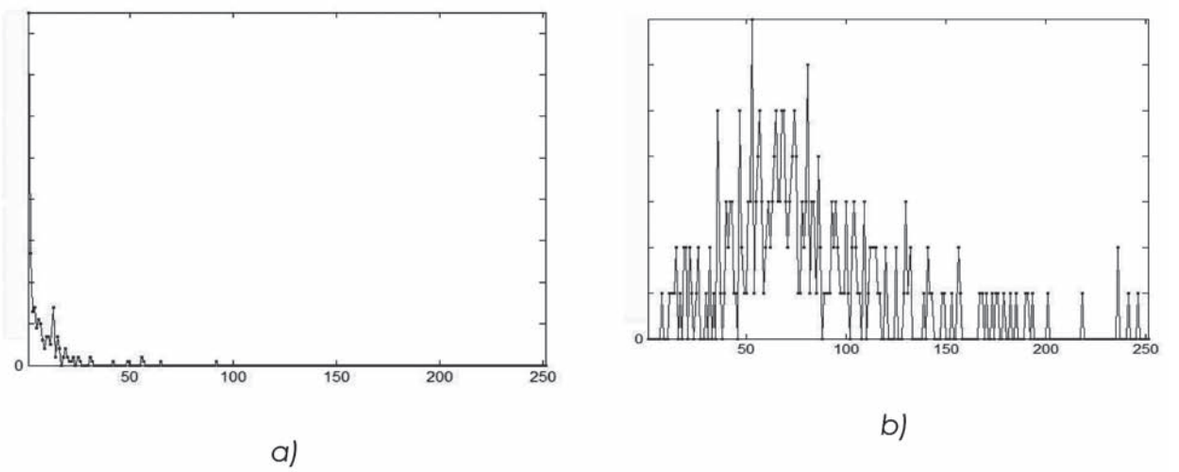

Рис. 8. Розподіл ступенів мережі:
a) при малому співвідношенні обсягу текстового корпусу до кількості персон (1000:250);
b) при великому співвідношенні (50000:250)

нок великої кількості «випадкових зв'язків», що проявляються при великих обсягах документів, мережа стає близькою до випадкової, в якій більша частина вузлів з'єднана з багатьма іншими (рис. 9, 10).

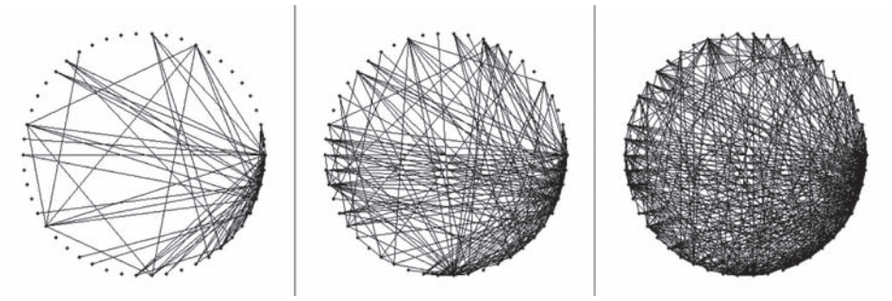

Рис. 9. Динаміка розвитку мережі при зростанні кількості документів у текстовому корпусі

У розглянутій мережі при фіксованій кількості вузлів, що відповідають персонам, і зростаючій кількості документів розподіл ступенів вузлів виявився спочатку близьким до степеневого, а потім - до пуасонівського (рис. 8). Це зумовлюється тим, що спочатку ступені вузлів мають систематичний характер, який відповідає реальним зв'язкам, а потім за рахунок великої кількості «випадкових зв'язків», що проявляються при великих обсягах документів, мережа стає близькою до випадкової, в якій більша частина вузлів з'єднана з багатьма іншими (рис. 9, 10).

Щоб відійти від явної деградації розглянутої мережі, пов'язаної з нагромадженням «випадкових зв'язків», було визначено накладену мережу, що відповідає вихідній, але із зміненими значеннями ваг ребер, причому вага ребер у новій мережі визначалася в такий спосіб:

$$v' = \begin{cases} 1, & v \geq \varepsilon v_{\max}; \\ 0, & v < \varepsilon v_{\max}, \end{cases}$$

де v' - вага ребра накладеної мережі, v - вага ребра вихідної мережі персон, $v_{\max}$ - максимальне значення ваги ребер вихідної мережі, ε - коефіцієнт огрубіння.

Відповідно до представленої моделі, де ε є граничним значенням при умові огрубіння накладеної моделі, сумарне значення всіх v' (кількість ребер у накладеній мережі) виявляється величиною постійною. Дослідження реальних даних показали, що постійними виявилися також значення середнього шляху та кластерності. Цей ефект дозволяє говорити про стабілізацію накладеної мережі та її відносної незалежності від обсягів вхідних документів. Зокрема, для значення $\varepsilon = 0.001$, 50 персон і кількості документів від 1000 до 50000, коефіцієнт кластерності склав 0.78 ± 0.01, а середня інверсна відстань - 0.65 ± 0.02.

Отримані емпіричні результати можуть бути корисними, наприклад, при теоретичному описі та моделюванні соціальних процесів [22], виявленні й візуалізації неявних зв'язків окремих об'єктів або суб'єктів. Феномен стабілізації накладеної мережі на практиці дозволяє шляхом аналізу невеликого масиву документів виявляти стійкі зв'язки, знижувати вплив шумових факторів. Разом з тим поки залишається відкритим питання оцінки кореляції отриманих інформаційних взаємозв'язків персон, що розраховуються шляхом підрахунку частоти документів, у яких персони згадуються спільно, та взаємозв'язків реальних.

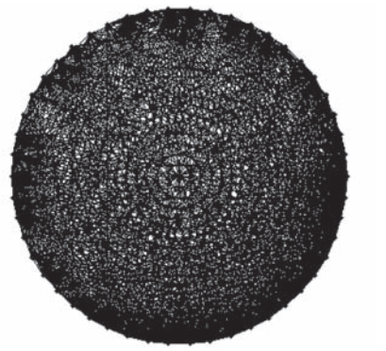

Рис. 10. Мережа, яка близька до стану деградації: 50 персон, 50000 документів

Засоби аналізу динаміки інформаційних потоків

Динаміка та обсяги представлення інформації в Інтернет створюють потужний інформаційний потік [3]. Причому потік досить неоднорідний, який може характеризуватися великою кількістю параметрів, серед яких виділяються такі, як джерела інформації (наприклад, Web-сайти) й тематики.

Обсяги повідомлень у тематичних інформаційних потоках утворюють часові ряди. Для дослідження часових рядів сьогодні все ширше використовується теорія фракталів [4, 23, 24], традиційна область застосування якої - фрактальна геометрія, обробка зображень та інше. Разом з тим часові ряди, породжувані тематичними інформаційними потоками, також мають фрактальні властивості й можуть розглядатися як стохастичні фрактали [25, 26]. Цей підхід розширює область застосування теорії фракталів на інформаційні потоки.

Якщо для традиційних засобів наукової комунікації підходи до статистичних досліджень інформаційних масивів з точки зору теорії фракталів були вперше досліджені Ван Рааном [25], що аналізував масиви статей і зв'язки, утворені цитуванням, то інформаційні потоки повідомлень із Інтернет до останнього часу не асоціювалися із фракта-

лами. Це пов'язане 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.

Сьогодні 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. Однією з основних властивостей фракталів є самоподібність (скейлінг). Як показано в роботах С.А. Іванова, для послідовності повідомлень тематичних інформаційних потоків відповідно до скейлінгового принципу, кількість повідомлень, резонансів на події реального світу 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. Крім того, для більшості тематичних інформаційних потоків спостерігається збільшення їхніх обсягів, причому на коротких часових інтервалах - лінійнийне зростання, а на тривалих - експонентне.

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

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

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

Сьогодні у зв'язку з розвитком теорії стохастичних фракталів стає популярною така характеристика часових рядів, як показник Хьорста (H). У книзі Е. Федера [23] показано, що він пов'язаний із традиційною «клітинною» фрактальною розмірністю (ρ) простим співвідношенням:

$$\rho = 2 - H.$$

Умова, за якої показник Хьорста пов'язаний із фрактальною «клітинною» розмірністю відповідно до вищенаведеної формули, визначена Е. Федером у такий спосіб: «...розглядаються клітки, розміри яких малі порівняно 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». Не вдаючись у подробиці відзначимо, що для інформаційних потоків ця властивість інтерпретується як самоподібність, яка виникає в результаті процесів їхнього формування. Можна помітити, що зазначені властивості притаманні не всім інформаційним потокам, а лише тим, які характеризуються достатньою потужністю й ітеративністю при формуванні.

Відомо, що показник Хьорста є мірою персистентності - схильності процесу до трендів (на відміну від звичайного броунівського руху). Значення $H > 1/2$ означає, що спрямована в певну сторону динаміка процесу в минулому, найімовірніше, спричинить продовження руху в тому ж напрямку. Якщо $H < 1/2$, то прогнозується, що процес змінить спрямованість. $H = 1/2$ означає невизначеність - броунівський рух.

Показник Хьорста зв'язують із коефіцієнтом нормованого розмаху (R/S), де R – обчислений певним чином «розмах» відповідного часового ряду, а S – стандартне відхилення за наступним алгоритмом. Спочатку обчислюється середнє значення вимірюваної змінної (у нашому випадку кількість повідомлень в інформаційному потоці) за N днів - $<\xi>_N$, потім розраховуються відхилення, що накопичуються, ряду вимірів $\xi(t)$ від середнього :

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

Після цього розраховується різниця максимального та мінімального відхилень, що накопичилися, яка називається "розмахом":

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

Хьорст експериментально виявив, що для багатьох часових рядів виконується:

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

де H і є показником Хьорста.

У цій статті дослідимо, як поводять себе часові ряди, що відповідають кількості публікацій у мережі Інтернет з визначеної проблематики. На

підставі обробки даних спостережень отримані значення різних статистичних показників відповідних рядів, а також показано, що вони мають фрактальну природу. Дослідження проводилися на наборі документальних корпусів, що містять повідомлення з відкритих веб-сайтів, сформовані системою InfoStream. Під час досліджень оброблявся інформаційний корпус, що містить повідомлення онлайнових ЗМІ — масив з 4317 документів, опублікованих за 486 діб з 1 січня 2006 р. по 30 квітня 2007 р., за тематикою комп'ютерної вірусології, що задовольняють запиту «комп'ютерний вірус» OR «вірусна атака» (рис. 11).

На рис. 12 зображена залежність нормованого розмаху (R/S) ряду спостережень від розміру ряду спостережень у подвійному логарифмічному масштабі. При цьому вона добре апроксимується прямою, показник Хьорста становить ~ 0,67, що відповідає фрактальній розмірності ~ 1,33.

Ще один з підходів до виявлення самоподібності часових рядів ґрунтується на методі DFA (Detrended Fluctuation Analysis) [27] - досить універсальному методі обробки тимчасових рядів. Цей підхід є варіантом дисперсійного аналізу одномірних випадкових блукань, що дозволяє досліджувати ефекти тривалих кореляцій у нестаціонарних часових рядах. У рамках алгоритму DFA аналізується середньоквадратична похибка лінійної апроксимації залежно від розміру ділянки апроксимації. Цей метод був застосований до ряду значень кількості публікацій з теми комп'ютерної вірусології в розрізі дат. У методі DFA для різних ділянок ряду спостережень однакової довжини k досліджуваної послідовності будується лінійна апроксимація, для якої потім обчислюється середньоквадратична помилка $D(k)$.

На рис. 13 представлена залежність середньоквадратичної помилки апроксимації від довжини ділянок апроксимації в подвійному логарифмічному масштабі. Наявність лінійного тренда на цьому графіку дозволяє говорити про наявність локального скейлінгу.

Проведені дослідження тематичних інформаційних потоків підтвердили припущення про самоподібність та ітеративність процесів в інформаційному просторі [28]. Републікації, цитування, прямі посилання тощо породжують самоподібність, що проявляється у стійких статистичних розподілах і відомих емпіричних законах.

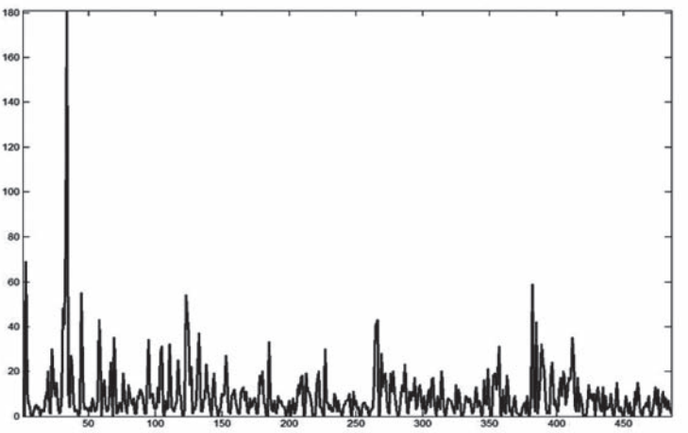

Рис. 11. Кількість публікацій (вісь ординат) у розрізі дат (вісь абсцис)

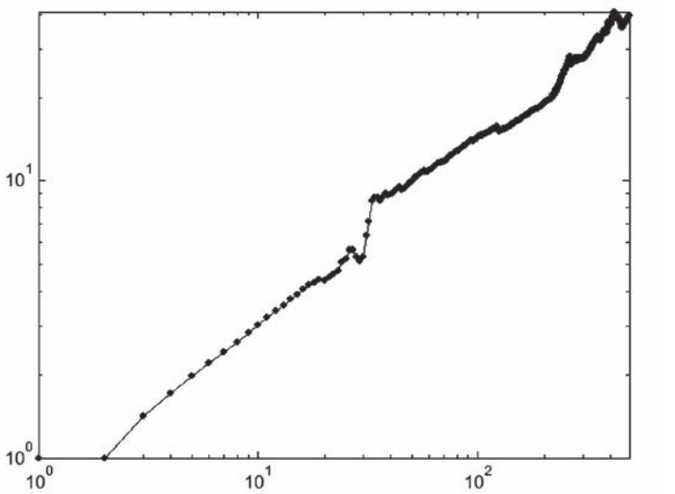

Рис. 12. Значення R/S (вісь ординат) залежно від розміру ряду спостере

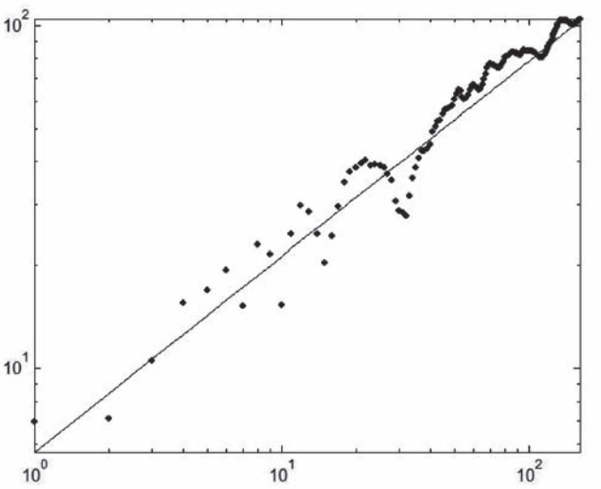

Рис. 13. Залежність D(k) ряду спостережень (вісь ординат) від довжини ділянки апроксимації (вісь абсцис)

Аналіз самоподібності інформаційних масивів, таким чином, може розглядатися як технологія, призначена для здійснення аналітичних досліджень із елементами прогнозування, здатна до екстраполяції отриманих залежностей.

Природно, описані результати досліджень можуть використовуватися не тільки для наведеного тематичного інформаційного потоку.

Об'єктно-статистичний аналіз інформаційних потоків

Аналіз процесів, які мають значні часові рамки, все ще чекає свого інструментарію. Як спробу створення засобів аналізу та візуалізації об'єктного розподілу відібраних інформаційних масивів великих обсягів наведемо результати досліджень, що проводилися під керівництвом автора [29], щодо виявлення об'єктного розподілу відібраних інформаційних масивів на прикладі аналізу динаміки публікацій в Інтернет-просторі про діяльність системи виборчих комісій в Україні по виборах Президента України й народних депутатів України за 2004-2006 роки. Ця динаміка відображає реальний інтерес громадськості, через електронні засоби інформації, до виборчих процедур, а також процеси, що відбувалися під час виборчих кампаній.

Система контент-моніторингу InfoStream дозволила побудувати залежність добових обсягів тематичних публікацій за 3 роки (1096 діб, загальна кількість – понад 320 тис.). Піки на графіку (рис. 14) дозволяють оцінити інтенсивність висвітлення в пресі як президентської виборчої кампанії 2004 р., так і виборів до Верховної Ради України в 2006 р.

Необхідно визначити, що для детального аналізу процесів загальноприйнятими методиками використовують аналіз Фур'є та вейвлет-аналіз [30, 31]. Технологія використання вейвлетів (маленьких хвиль) 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 [32]. Метод вейвлет-аналізу використовується також для декомпозиції, виділення сигналу з «шуму», вивчення динаміки різних процесів, у тому числі економічних і соціальних. На рис. 15 наведена спектрограма - результат вейвлет-аналізу часового ряду, що відповідає процесу, який досліджується.

Прекрасно відображаючи спектральні характеристики сигналів, вейвлет-аналіз однак, за своєю природою, не може використовуватися, коли інфор-

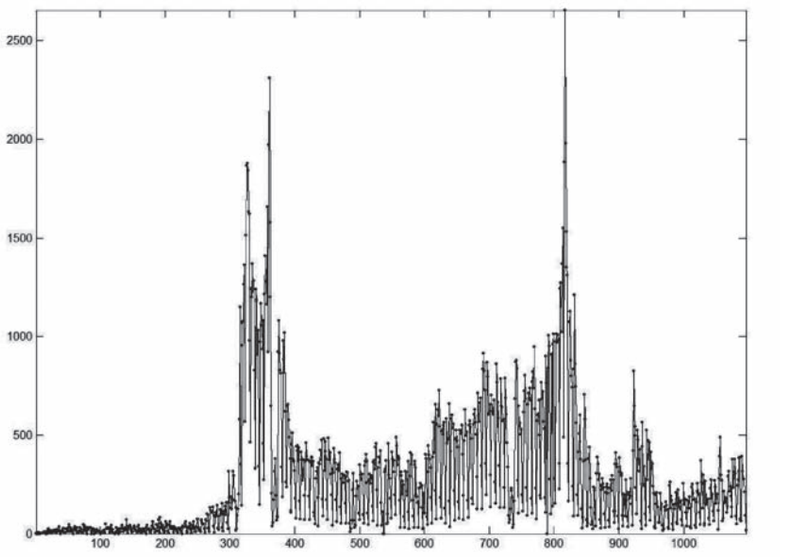

Рис. 14. Кількість тематичних публікацій (вісь ординат) по днях (вісь абсцис)

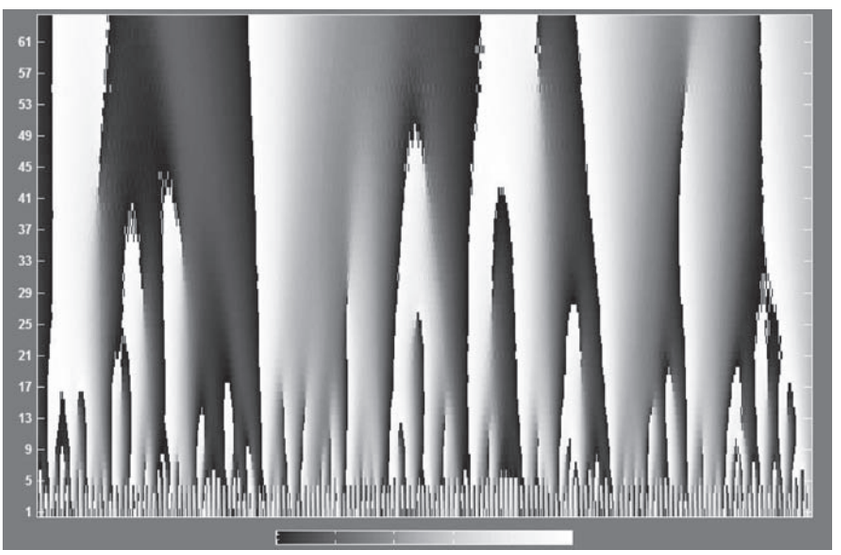

Рис. 15. Вейвлет-спектограма динаміки тематичного інформаційного потоку (одномірне безперервне вейвлет-перетворення, вейвлет Гауса), вісь абсцис - дні, вісь ординат - частоти

маційний потік варто розглядати з об'єктної точки зору. У випадку, що розглядається, такими об'єктами виступали окремі особи, згадувані в публікаціях за своїми прізвищами й посадами. Зокрема, за допомогою засобів екстрагування інформації системи InfoStream з розглянутого потоку було виявлено згадування про більш ніж 40 тис. осіб, що тією чи іншою мірою мали відношення до виборчого процесу.

В екстрагованому вигляді кожна персона представлялася одним дескриптором. Для забезпечення врахування й аналізу розподілу інформаційних потоків у розрізі персон, що цікавлять, був запропонований оригінальний метод так званих вордлет-діаграм. Вони є формою візуального відображення інформаційного потоку в розрізі об'єктів і дат, яка представляє собою прямокутну таблицю. Стовпцям цієї таблиці відповідають дати, а рядкам - об'єкти, що є своєрідними змістовними фільтрами інформаційного потоку, що досліджується. Об'єктам у розглянутому випадку відповідають визначені персони. Природно, для візуального відображення з безлічі персон вибираються лише кілька десятків з тих, що цікавлять дослідника.

Візуально вордлет-діаграма є таблицею, клітинки якої зафарбовані відтінками сірих кольорів, залежно від значень об’ємів публікацій за обраним об'єктом у відповідний день.

Вордлет-діаграми для відносно невеликої кількості рядків дозволяють візуально виявляти групи найбільш зв'язаних за датами та інтенсивністю публікацій щодо об'єктів. Для великої кількості об'єктів у процесі побудови відповідної діаграми пропонується її кластеризація шляхом перестановки рядків (перегрупування об'єктів) відповідно до алгоритму k-means [8].

На рис. 16 наведена діаграма першого рівня (прев'ю), що дозволяє візуально виявляти аномальні кореляції. На цій діаграмі, що охоплює інформацію про 49 персон, чітко помітні цикли святкових днів, а також кореляції окремих об'єктів. За допомогою уточнюючої діаграми можна точно вказати на виявлені кореляції.

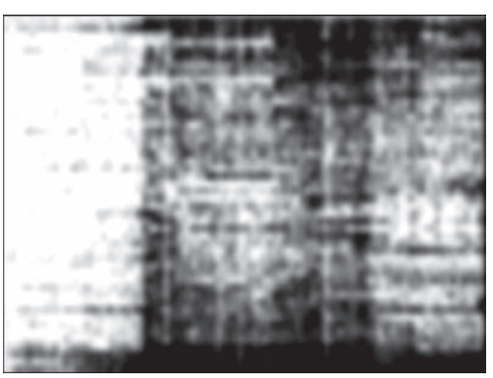

Рис. 16. Вордлет-діаграма-прев’ю

У результаті проведених експериментів є підстави припустити, що використання таких засобів візуалізації, як вордлет-діаграми, дозволяє «розкладати» вихідні часові ряди відповідно до об'єктів, виявляти медіа-активність по обраних об'єктах, виявляти взаємозв'язки об'єктів у розрізі дат, визначати деталі медіа-активності кожного об'єкта або групи об'єктів. Наведені засоби дають змогу адекватніше аналізувати динаміку публікацій в Інтернет у розрізі об'єктів, надаючи наочно важливу інформацію щодо динаміки реальних процесів. Використання вордлет-діаграм є важливим доповненням до вже визнаних методів досліджень, таких як аналіз Фур'є, кореляційний і фрактальний аналіз, а також вейвлет-аналіз.Крім того, необхідно відзначити, що представленому підходу до аналізу інформаційних потоків притаманний об'єктно-статистичний характер, що, у свою чергу, представляється як істотна складова методологічної бази прогнозно-емпіричного аналізу.

Висновки

Для прийняття грунтовних рішень у галузі національної безпеки держави необхідне використання комплексних систем, які дозволяють збирати, обробляти та узагальнювати інформацію про об’єкт досліджень, отриману з різних джерел із застосуванням різних технологій.

Сьогодні складні завдання інформаційно-аналітичної підтримки прийняття рішень, з одного боку, стимулюють розвиток систем керування знаннями, глибинного аналізу даних і текстів, а з іншого - найбільш розвинені із цих систем у явному виді містять адаптовані й готові до використання аналітичні блоки. Тому вже є широкий вибір засобів автоматизації аналітичної діяльності. Причому рівні функціональності таких систем можуть бути дуже різноманітним - від простих інформаційно-пошукових програм, необхідних на етапі становлення аналітичних систем, до дорогих і ресурсномістких систем керування знаннями та глибинного аналізу даних і текстів.

Обсяги інформації, необхідної для аналітичної роботи, бувають настільки величезні, що навіть спеціалізована пошукова система не завжди здатна швидко відшукати необхідний документ. Низка досліджень у США показали, що співробітники компаній можуть витрачати до трьох годин на день для пошуку потрібної інформації. Внаслідок цього багато найбільших фірм щорічно втрачають $2,5 млрд. Саме для вирішення даної проблеми вже існують портали знань, що представляють середовище для ефективного пошуку та обміну знаннями. По суті, портали є рішеннями, що виконують одночасно функції зберігання, класифікації, знаходження та обробки знань.

***

Современные средства информационно-аналитической поддержки принятия решений обеспечивают решение целого комплекса проблем, среди которых сбор информации об объектах, определении связей объектов, выявление тенденций, прогнозирование. Функциональные возможности таких систем разрешают выполнять диагностику и прогнозирование развития ситуации. В дополнение к возможностям глубинного анализа данных и текста в таких системах используется также человеческий опыт, знание экспертов.Сейчас уже очевидно, что реальный прорыв в сфере интенсификации информационно-аналитической работы, как и в науке, возможный лишь в результате агрегирования разных направлений. Изложенные в статье подходы с нескольких, конфликтных прежде точек зрения, сегодня могут рассматриваться как пути решения открытой проблемы навигации в современном информационном пространстве.

***

Modern information and analytical support tools of decision-making provide solution of whole complex of problems as: collection of information about object, identification of object relations, tendency detection and prognosticating. Functional abilities of such systems will enable to make diagnostics and prognostication of situation development. Deep data and text analysis is amplifies with human experience and knowledge of experts. It's obviously that real breakthrough in information and analytical work intensification is feasible only by aggregating of different concepts. Given approaches to some conflicting before standpoints now could be considered as the way of solution for open problem of navigation in modern information space.

***

1. Michael W. Berry. Survey of Text Mining. Clustering, Classification, and Retrieval. - Springer-Verlag, 2004. - 244 p.

2. Newman M.E.J. The structure and function of complex networks // SIAM Review. - 2003. - Vol. 45. - pp. 167–256.

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

4. Гринченко В.Т., Мацыпура В.Т., Снарский А.А. Введение в нелинейную динамику. Хаос и фракталы: Изд. 2. – М.: УРСС, 2007. - 263 с.

5. Плэтт В. Информационная работа стратегической разведки. – М.,1958. - 510 с.

6. Ландэ Д.В., Прищепа В.В. Школа веб-разведки // Журнал "Телеком". – К., 2007. - № 6. - С. 40-45.

7. Ландэ Д.В. Затерянный веб // Журнал "Телеком". – К., 2005. - № 1-2. - С. 46-51.

8 *Chakrabarti Soumen*. Mining the web. Discovery knowledge from hypertext data. - Publisher: Morgan Kaufmann, 2002. - 344 p.

9. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. - М.: "Вильямс", 2005. - 272 с.

10. Хан Удо, Мани Индерджиет. Системы автоматического реферирования // Открытые системы. - 2000. - № 12. (http://www.osp.ru/os/2000/12/067.htm).

11. Ландэ Д.В., Снарский А.А. Попытки объять необъятное, или World Wide Web под прицелом // Журнал "Сети и бизнес". - Киев, 2007. - № 4 (35). - С. 18-24.

12. Erdős, P., Rényi A. On Random Graphs. I. // Publicationes Mathematicae 6, pp. 290-297. -1959.

13. Erdős P., Rényi A. On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci. 5. - pp. 17-61. - 1960.

14. Watts D.J., Strogatz S.H. Collective dynamics of ″small-world″ networks // Nature. - 1998. - Vol. 393. - pp. 440–442.

15. Albert R., Jeong H., Barabasi A. Attack and error tolerance of complex networks // Nature. - 2000. - Vol. 406. - pp. 378-382.

16. Bjorneborn, L., Ingwersen, P. Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 55(14): 1216-1227. – 2004.

17. Milgram S. The small world problem, Psychology Today, 1967, - Vol. 2. - pp. 60-67.

18. Clauset, A., Moore, C., Newman, M.E.J. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98-101 (1 May 2008).

19. Broadbent S.R., Hammersley J.M. Percolation processes // I. Crystals and mazes, Proc Cambridge Philos. Soc. – pp. 629-641. – 1957.

20. Снарский А.А., Безсуднов И.В., Севрюков В.А. Процессы переноса в макроскопических неупорядоченных средах: От теории среднего поля до перколяции. – М.: УРСС, Изд-во ЛКИ, 2007. - 304 с.

21. Григорьев А.Н., Ландэ Д.В. и др. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: ООО «Старт-98», 2007. – 40 с.

22. Фурашев В.Н., Ландэ Д.В., Брайчевский С.М. Моделирование информационно-электоральных процессов: Монография. - К.: НИЦПИ АпрН Украины, 2007. - 182 с.

23. Федер Е. Фракталы. - М.: Мир, 1991. - 254 с.

24. Ландэ Д.В. Фрактальные свойства тематических информацийнных потоков из Интернет // Реєстрація, зберігання і обробка даних. - 2006. - Т. 8. - № 2. – С. 93-99.

25. Van Raan A. F. J. Fractal geometry of information space as represented by cocitation

clustering // Scientometrics. - 1991. – Vol. 20. - № 3. – P. 439-449.

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

27. 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.

28. Додонов А.Г., Ландэ Д.В. Самоподобие массивов сетевых публикаций по компьютерной вирусологии // Реєстрація, зберігання і обробка даних. – 2007. - Т. 9. - № 2. – С. 53- 60.

29. Ландэ Д.В., Григорьев А.Н., Брайчевский С.М., Дармохвал А.Т., Снарский А.А. Объектная визуализация тематических информационных массивов // Труды 9-ой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL'2007. - Переславль-Залесский, Россия, 2007. – С. 148–150.

30. Чуи К. Введение в вэйвлеты. - М.: Мир, 2001.

31. Давыдов А.А. Вейвлет-анализ социальных процес сов // Социолог. исслед. – 2003. - №11. - С. 97-103.

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

Інститут проблем національної безпеки