УДК 351.745.7+681.142

Гавловський Владислав Данилович — начальник відділу Міжвідомчого науково-дослідного центру з проблем боротьби з організованою злочинністю при Раді національної безпеки і оборони, кандидат юридичних наук, старший науковий співробітник

Окремі питання отримання інформації з відкритих джерел для правоохоронних органів

У статті розглядаються можливості та проблеми отримання інформації, необхідної для оперативної і службової діяльності правоохоронних органів, з відкритих джерел, зокрема мережі Інтернет.

Ключові слова: інформаційно-пошукові системи, мережа Інтернет, злочинність, правоохоронні органи.

В умовах реформування України проблема забезпечення кримінологічної безпеки особи, суспільства, держави набула глобального виміру. Негативний соціальний фон, що склався в країні, обумовив виникнення складної криміногенної ситуації. Сьогодні злочинність є одним із основних чинників, що перешкоджають здійсненню соціальної реформи, породжують у громадян почуття тривоги за своє життя і благополуччя, знижують довіру до органів влади та управління, до проведеної державної політики, безпосередньо загрожують національній безпеці держави.

З огляду на це, захист громадян і суспільства від злочинних посягань на їхні права, законні інтереси та майно з боку злочинців сьогодні не мислимий без широкого та активного використання в діяльності органів внутрішніх справ досягнень сучасної науки і техніки, нових технологій, в тому числі інформаційних.

Україна активно прямує до інформаційного суспільства, що характеризується розвитком та інтенсивним упровадженням сучасних інформаційних технологій в різні сфери діяльності. Це, в свою чергу, обумовлює зростання злочинних проявів у цих сферах. Особливо це відчувається з приєднанням до міжнародних систем телекомунікації та

підвищення інтелектуального рівня зловмисників, які використовують мережу Інтернет для вчинення не лише комп'ютерних, а й традиційних злочинів новими засобами. Через мережу Інтернет злочинці обмінюються інформацією щодо готування та вчинення злочинів. Мережа також використовується, зокрема, для нелегальної торгівлі наркотиками, людськими органами, порнографічною продукцією тощо.

Тому співробітники правоохоронних органів повинні вміти виявляти, розкривати, розслідувати та попереджати злочини, які вчиняються з використанням інформаційних технологій. З цією метою потрібно знати і вміти знаходити потрібну (розвідувальну, оперативно значущу та ін.) інформацію і в відкритих джерелах.

Метою статті є окреслення окремих проблем щодо моніторингу Інтернет простору на предмет виявлення інформації, яка представляє інтерес для правоохоронних органів.

Зі збільшенням кількості відкритих джерел, з яких отримується найбільше інформації, збільшується об'єктивність здобутої інформації. Між тим, різко зростають і трудовитрати на знаходження такої інформації. Проте, не вся потенційно відкрита "нетаємна" інформація є добре доступною, швидше – навпаки. Знаходження необхідної в кожному конкретному випадку інформації – складне завдання. Експерти переконані, що 85—90 % інформації в мережі Інтернет можна отримати в результаті порівняння, інтеграції, аналізу численних розрізнених даних і лише 10–15 % потрібної інформації є в готовому вигляді [1].

На думку колишнього директора ЦРУ Р. Хілленкерта, "80 % розвідувальної інформації отримується із таких джерел, як книги, журнали, науково-технічні огляди, фотографії, комерційні аналітичні звіти, теле- і радіопередачі" [2].

Необхідно зазначити, що при пошуку потрібної інформації в мережі Інтернет існує низка проблем, зокрема:

— необхідність охоплення великих обсягів інформації;

— велика і складна динаміка інформаційних потоків;

— багатократне дублювання інформації та надлишок шумової інформації;

— наявність закритого веб-простору, який недоступний сучасним інформаційно-пошуковим системам;

— відсутність реального веб-простору, ефективних алгоритмів пошуку в розподілених мережах (наприклад, пірінтових, соціальних), засобів смислового пошуку;

— пошук мультимедійної інформації, мультимовних засобів пошуку;

— відсутність у вільному доступі універсальних пошукових служб, що забезпечують пошук персональних інформаційних потреб користувачів, тощо [3, с. 37].

Розглянемо деякі з цих проблем більш детально.

Однією з особливостей мережі Інтернет є постійне зростання темпів продукування інформації. На початку існування Всесвітньої павутини лише на невеликій кількості веб-сайтів публікували інформацію окремих авторів для відносно невеликої кількості відвідувачів. Сьогодні активна участь відвідувачів у створенні контенту призвела до різкого зростання обсягу і динаміки інформаційного простору. За підрахунками американських дослідників Берклі Пітера Лаймана та Хола Верієна, обсяг усієї інформації, яка виробляється людством, за три роки подвоюється [4, с. 136]. Сьогодні для доступу традиційними інформаційно-пошуковими системами, наприклад пошуковою системою Cuil, доступні понад 120 млрд веб-сторінок. Проте, це незначна частина інформаційних ресурсів, тому що прихованих і невидимих ресурсів глобальної мережі значно більше. Ерік Шмідт, керівник пошукової системи Google, зауважив, що на сьогодні проіндексовано лише 170 Тбайт інформації. Навіть такій потужній пошуковій системі, як Google, знадобиться близько 300 років для індексації всієї інформації [5].

Інформація в мережі Інтернет має явно виражений динамічний характер: вона розміщується на сайтах, модифікується і з часом знищується. Інформаційний простір складається зі стабільної і динамічної частин, які з часом змінюють свою наповненість: деякі документи направляються в стабільну частину у вигляді архівів, інші зникають. Тобто, користувач при зверненні до пошукових систем знайде релевантні запиту документи із стабільної частини, але жодного документа не отримає із оновленої динамічної системи. Частково цю проблему можливо вирішити використовуючи систему контент-моніторингу інформаційних потоків у мережі Інтернет, яка збільшує кількість знайдених релевантних документів у середньому в 20 разів. Між тим, контент-моніторинг не робить повної вибірки документів із стабільної частини Інтернет, для цього мають бути використані два інструменти — традиційні інформаційно-пошукові системи для стабільної частини веб-простору і системи контент-моніторингу інформаційних потоків.

Крім того, важливі повідомлення в мережі Інтернет нерідко дублюються в експоненціально зростаючій кількості веб-сайтів, у той час, як кількість джерел, що заслуговують на увагу, зростає лінійно [3, с. 138]. Існуючі на сьогодні алгоритми виявлення дублів у сучасних інформаційних потоках вимагають великих обчислювальних потужностей. Важливим у цьому аспекті є використання лінгвістично-математичних алгоритмів.

Необхідно зазначити, що інтенсивність росту шумової інформації (спаму – масової розсилки кореспонденції рекламного чи іншого характеру людям, які не висловили бажання її одержувати), яка нерідко нав'язується користувачам, багаторазово перевищує інтенсивність росту корисної інфо-

рмації. Найбільший потік спаму поширюється через електронну пошту. Сьогодні частка вірусів і спаму в загальному трафіку електронної пошти становить за різними оцінками від 70 до 95 % [6].

Наступна проблема – пошук інформації в "невидимому" веб-просторі. Це відкриті веб-ресурси з відкритою інформацією, яка знаходиться у вільному доступі, але недоступна для традиційних інформаційно-пошукових систем. Тут зберігається набагато більше інформації, яка потенційно цікавіша для багатьох користувачів, і, в перше чергу, співробітників правоохоронних органів, ніж у відкритій частині глобальної мережі. Наприклад, мережа Freenet дозволяє користувачам бути абсолютно анонімними навіть тоді, коли вони поширюють віруси, розширюють кримінальні зв'язки та виставляють дитячу порнографію. У невидимій мережі, куди не дістають пошукові системи, "ховаються" збоченці, "громадські активісти" та злочинці.

Інтегрований доступ до таких ресурсів нині залишається ще відкритою проблемою, хоча часткове її вирішення існує, наприклад за допомогою спеціальних каталогів і систем. Так, американський журнал Business 2.0 надрукував список із семи технологій, які повинні "змінити світ" у найближчому майбутньому. Серед них є і пошук у невидимих мережах ("невидимому" веб-просторі) [7]. Тому немає потреби говорити про те, наскільки ця технологія буде корисною для правоохоронних органів.

Кріс Герман і Гері Прайс виокремлюють чотири типи невидимості вмісту Всесвітньої павутини [8, 63].

1. Невидимість, обумовлена налаштуваннями пошукових систем і їх природними особливостями, це так званий сірий Інтернет, який має ряд варіантів:

— обмеження глибини проникнення пошукових систем на сайт, налаштоване власниками пошукових машин;

— зміна сторінок, що відбулася після відвідування пошуковою системою, тобто після чергового індексування;

— обмеження максимальної кількості показаних до видачі сторінок;

— веб-сторінки не прописані в формі "Додати сторінку" та не мають посилань з інших адрес.

Пошукові системи ніяк не зможуть дізнатися про існування таких веб-сторінок, тому ніколи їх не проіндексують. Ці сторінки, до речі, можуть представляти великий інтерес для правоохоронних органів тому, що на них може знаходитися інформація про готування до злочинів, про самі злочини, особливо у фінансовій сфері та економіці, пропозиції щодо продажу наркотиків, зброї, дитячої порнопродукції тощо. Посилання на такі веб-сторінки знаходяться на інших веб-

сторінках, які відомі певній категорії користувачів, тому інформація на цих сайтах орієнтована на певні групи клієнтів.

За підрахунками IBM кількість таких веб-сторінок, невідомих пошуковим системам, становить близько 20 % від загальної кількості адрес, які могли б бути проіндексовані з технічної точки зору.

2. Веб-сторінки, які навмисно виключені веб-майстром по індексації. Це ресурси, захищені паролем або включені в файл robots. txt чи прикриті тегом <noindex>. У цьому випадку сам власник сайту визначає, чи можна індексувати окремі файли, групу файлів, окремі веб-сторінки чи весь веб-сайт. Метод обмеження індексування за допомогою файла robots. txt чи мегатега <noindex> отримав широке розповсюдження тому, що він не "заважає" роботі пошукових систем і без обмежень дозволяє переглядати зміст сторінки. Варто зазначити, що більшість пошукових машин з повагою ставляться до подібного способу захисту інформації.

3. Веб-сторінки, які потребують реєстрації. Це безкоштовні ресурси, але при вході на веб-сторінку потрібно натиснути кнопку "я згоден" чи ввести якусь інформацію, що висвітлюється поряд у графічному вигляді.

4. Дійсно невидима мережа Інтернет:

— на веб-сторінці знаходяться дані в недоступному для пошукової машини форматі;

— веб-сторінки, які навмисно не обслуговуються з тих чи інших причин;

— інформація зберігається в базі даних і доступ до неї можливий лише за умови заповнення певної форми.

До "невидимих" вебів також відносяться чисельні системи інтерактивної взаємодії з користувачами – консультації, навчання, які вимагають участі людей для формування динамічної відповіді від серверів. До них можна віднести також закриту (частково чи повністю) інформацію, яка доступна користувачам, лише з певних адрес чи груп адрес, інколи міст чи країн [9].

Особливість більшості "невидимих" ресурсів – у їх вузькій спеціалізації. Для пошуку в них використовуються ті ж механізми, що і для відкритого web, однак, найчастіше, пошукові системи для невидимого web включають унікальні для кожного такого ресурсу модулі доступу до даних.

Традиційна пошукова система частіше за все може видати адресу бази даних, але не відомо, які конкретно документи будуть в ній міститися. Типовим прикладом є база даних "Законодавство України" Верховної Ради України, тисячі документів з якої стають доступні тільки після входу в систему, а стандартні пошукові системи не в змозі проіндексувати контент баз даних.

Значний обсяг інформаційних ресурсів в мережі Інтернет знаходиться в "пірінгових" мережах (P2P). У таких мережах відсутні виділені сервери, а кожен вузол є як клієнтом, так і сервером. Існує ряд областей, де застосовуються пірінгові мережі. Серед них: обмін файлами, обмін повідомленнями, Інтернет-телефонія, розподільні обчислення, групова робота.

Такі мережі на сьогодні відносяться до числа найбільш популярних у мережі Інтернет з точки зору обміну файлами. Вони були створені, в першу чергу, для легальних цілей, але нині посилено використовуються для обміну забороненими файлами, тобто з метою вчинення противоправних діянь. За допомогою цих мереж відбувається обмін інформацією між територіально розподіленими злочинними елементами, файлами з дитячою порнографією, співпрацюють члени терористичних організацій, які займаються рекрутуванням, розповсюдженням стратегічних вказівок, поточних наказів тощо.

Питання ефективного пошуку в таких мережах залишається відкритим, існують лише спеціальні пошукові програми, які допомагають вирішувати цю проблему. Стандартні дії щодо захисту інтересів правовласників, у вигляді накладення різних заборон і обмежень, призводять до того, що вже зараз починають створюватися анонімні, повністю децентралізовані пірінгові мережі з системою шифрування даних, що може призвести до дуже негативних наслідків, пов'язаних із виникненням мереж, які дозволять абсолютно безперешкодно поширювати матеріали будь-якого змісту.

Над вирішенням проблеми смислового, змістовного пошуку працюють чисельні колективи вчених і фахівців у всьому світі. Вирішенням цієї проблеми займаються і науковці України. Так, на факультеті кібернетики розроблено технологію комп'ютерно-лінгвістичної обробки текстів на природній мові, яка базується на створених потужних лінгвістичних базах даних та інтелектуальних алгоритмах смислової обробки текстів.

Для створення ефективних методів та алгоритмів обробки текстової інформації необхідне представлення мовних знань у вигляді баз даних і баз знань, серед яких найбільш потужною є англомовна семантична мережа WordNet. Розроблено білінгвістичну семантичну базу знань UkrWordNet обсягом більш, ніж 120 000 смислових одиниць та асоціативних зв'язків між ними.

Використання цієї технології дозволило створити цілий ряд систем різнотипної обробки текстових даних:

система смислового пошуку текстів. Програма допомагає сучасним системам пошуку здійснювати визначення документів, релевантних запитам користувача, не за фактом входження ключових слів запи-

ту в тіло тексту, а за подібністю смислових структур тексту до семантичного образу запиту;

система семантичної фільтрації текстів. На вхід системи подається список тем, що цікавлять користувача та вхідний текст природною мовою. Система аналізує текст і визначає чи є документ семантично приналежним до заданих тем;

система смислової тематичної класифікації текстових корпусів і потоків, яка здійснює огляд корпусів текстів і текстових потоків, вибираючи з них документи, які відповідають профілю інтересів користувача. Дозволяє користувачу самому сформувати свій профіль тем, які його найбільш цікавлять, через створення потрібних йому тематичних каталогів і підкаталогів з подальшим заповненням їх еталонними текстами зазначених тем. Система аналізує вміст папок користувача, самонавчається та налаштовується на визначення текстів, що належать до улюблених тем користувача;

система "Рефератор" призначена для обробки текстів природної мови. За допомогою даної системи можна легко та зручно створювати реферати текстів і проводити їх індексацію (визначення за тематикою). Підтримка системи каталогізації надає можливість зберегти результати в базі даних;

система фільтрації Internet-повідомлень із використанням лінгвістичних методів аналізу текстів. Система аналізує потоки текстової інформації у комп'ютерній мережі (з можливістю заборони доступу до визначеного контенту).

Не дивлячись на те, що в мережі Інтернет знаходиться значна кількість інформації, яку співробітники правоохоронних органів можуть використовувати в оперативній і службовій діяльності, залишається відкритим питання щодо її знаходження та ефективного використання, зокрема, через недоліки мережі Інтернет, а саме непропорційний ріст рівня інформаційного шуму, слабкої структурованості та пов'язаності між собою інформації, динамічності й відсутності гарантій її цілісності, багатократного дублювання, обмеження доступу до "невидимого" Інтернет-простору тощо; майже повну відсутність у співробітників правоохоронних органів спеціальних інформаційно-пошукових систем, особливо контент-моніторингу, контент-аналізу; недостатню кількість спеціалістів, підготовлених у цьому напрямку.

Список використаних джерел

1. Доронин А. Аналитическая разведка средствами Інтернет [Електронний ресурс] / А. Доронин. — Режим доступу :

http://webcache. googleusercontent. com/search? hl=ru&q=cache: H1tsoNnwa RAJ: http://www.agentura. ru/dossier/russia/people/doronin/internet/.

2. Школа веб-разведки. Инструменты и источники / [Електронний ресурс]. — Режим доступу :

http://webcache. googleusercontent. com/search? hl=ru&q=cache: O_uG Hvm-BhUJ: http://scip. org. ua/2010/05/04.

3. Додонов А. Г. Современные поисковые технологии – проблемы и некоторые пути их решения / Додонов А. Г., Ланде Д. В., Путятин В. Г. // Реєстрація, зберігання і обробка даних. — 2010. — № 3. — Т. 12. — С. 46–55.

4. Фурашев В. М. Електронне інформаційне суспільство України : погляд у сьогодення і майбутнє : [монографія] / Фурашев В. М., Ланде Д. В., Григорєв О. М., Фурашев О. В. — К. : Інжиніринг, 2005. — 164 с.

5. Google : на индексацию всей информации уйдёт 300 лет / [Електронний ресурс]. — Режим доступу :

http://webcache. googleusercontent. com/search? hl=ru&q= cache:31s93bC3 bKUJ: http://virusinfo. info /.

6. Спам / [Електронний ресурс]. — Режим доступу :

http://webcache. googleusercontent. com/search? hl=ru&q=cache: FZnScQ2PM2gJ: http://ru. wikipedia. org/wiki /.

7 Ландэ Д. В. Конкурентная разведка в WEB [Електронний ресурс] / Ландэ Д. В., Прищепа В. В. — Режим доступу :

http://www. inter-biz. com. ua/fifth/index. php? razdel_id=139.

8. Ющук Е. Интернет-разведка. / Е. Ющук. — М. : Вершина, 2007. — 256 с.

9. Ландэ Д. Дорожная карта сетевого поискового бизнеса / [Електронний ресурс]. — Режим доступу :

http://webcache. googleusercontent. com.

В статье рассматриваются возможности и проблемы получения информации, необходимой для оперативной и служебной деятельности правоохранительных органов, из открытых источников, в частности сети Интернет.

The article deals with the examination of the possibilities and problems of the receipt of information, necessary for the operative and official activity of the law enforcement bodies, from the public sources, in particular the Internet networks.

Стаття надійшла до редакції журналу 21 грудня 2010 року.