УДК 519.257

Р.І. Дзінько, О.І. Лісовиченко, А.М. Гордійчук

МЕТОДИ ЗАХИСТУ ІНФОРМАЦІЇ ВІД ПІДМІНИ ДЖЕРЕЛА НА ВЕБ САЙТАХ

Вступ

Веб сайт являє собою додаток, архітектура якого є клієнт-серверною з тонким клієнтом, в ролі якого виступає веб оглядач або інша програма, частиною якої є алгоритми створення запитів та обробки відповідей від сервера через HTTP протокол. Інформація в даній системі рухається в двох напрямках: від клієнта до сервера, і від сервера до клієнта.

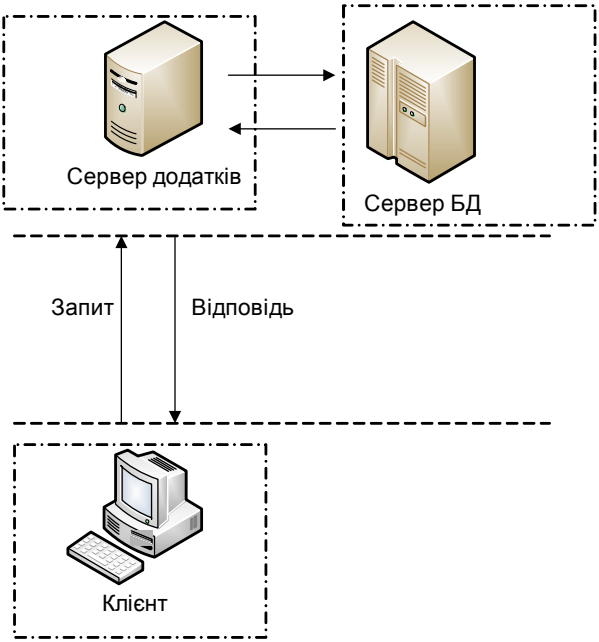

Рис. 1 – Схема руху та зони доступу до інформації.

Як бачимо з рис. 1 єдиною зоною доступу до інформації є клієнт. Інформація, до якої має доступ відвідувач сайту доступна зловмисникам у точно такому ж вигляді.

Постановка задачі

Через доступність інформації на сьогоднішній день багато веб сайтів зустрічаються з проблемою підміни джерела інформації, тобто інформа-

 ISSN 1562-9945

ція з сайта першоджерела з'являється на іншому сайті без дозволу адміністрації сайту-першоджерела. Такі випадки особливо актуальні для інформаційних сайтів (інформаціні агентства, дошки оголошень). Дані з таких сайтів зловмисники отримують шляхом використання роботів, які в автоматичному режимі читають інформацію з сайтів, аналізують її, структурують і в незміненому, або частково зміненому вигляді, розміщують на інших сайтах.

Такі дії призводять до того, що відвідуваність сайта знижується через те, що користувачі можуть отримати цю ж інформацію з альтернативного джерела. Оскільки, відвідуваність є джерелом прибутку (наприклад, доходом сайту є розміщення реклами), дана задача захисту інформації від автоматичного чи автоматизованого аналізу та структурування є актуальною і потребує вирішення.

Враховуючи те, що ці дані є публічно доступні і передаються через протокол передачі даних HTTP, важко визначити, є отримувач інформації відвідувачем сайту чи "роботом-павуком".

Проблема також ускладнюється тим, що інформація, яку розміщують на первинному сайті (первинне джерело) повинна бути доступною для роботів, які запускаються з метою аналізу пошуковими системами (Google, Yandex, Yahoo, Bing тощо), які по суті нічим не відрізняються від роботів зловмисників.

Терміни і поняття

- *Пошукова система* – програмно-апаратний комплекс з інтерфейсом користувача, що надає змогу виконувати пошук інформації в мережі Інтернет [1];
- *Пошуковий робот* – програма, складова частина пошукової системи і призначена для перебору інтеренет-сторінок з ціллю занесення інформації про них в базу даних системи [1];
- *User Agent* – текстовий рядок, що є частиною HTTP запиту, що починається з "User-agent:" та, зазвичай, включає таку інформацію як назва та версія програми клієнта, операційну систему комп'ютера та мову [2];
- *Робот-аналізатор* – програма-клієнт, що отримує інформацію з веб сайту, аналізує та класифікує її;
- *Робот-зловмисник* – робот-аналізатор, що застосовується з ціллю витягування інформації з сайтів-першоджерел інформації з ціллю розміщення її на сайтах конкурентів.

Введення обмежень

Необхідно ввести методи захисту інформації від підміни джерела таким чином, щоб не порушувались наступні умови:

1. *Зниження швидкості опрацювання запитів* до веб сайту не повинне бути настільки великим, щоб сайт став незручним у користуванні (опрацювання запиту повинно відбуватись не більше 5-ти секунд).

2. *Доступність інформації пошуковим системам* (тобто пошукові роботи повинні безперешкодно отримувати доступ до інформації на веб сайті).

Задачу можна вважати вирішеною в наступних часткових випадках:

1. *На сайті зловмисника відсутня інформація, що вказує на первинне джерело* (водяні знаки на зображеннях, відео, . . .).
2. *Інформація з’явилась на сайті зловмисника зі значною втратою якості* (розмиті зображення, стиснене відео. . .).

Задача вважається вирішеною, якщо виконуються наступна умова:

$$K \cdot C_p - \boldsymbol{C}_n < 0, \tag{1}$$

де C_p – вартість створення робота-аналізатора для зловмисника, $\boldsymbol{C}_n$ – максимальний прибуток, отриманий від використання інформації, отриманої і опрацьованої роботом аналізатором, K – деякий коефіцієнт, що знижує максимальний прибуток. При розв’язанні поставленої задачі візьмемо ідеальні умови:

- сайт-першоджерело та сайт зловмисника є однаково доступними, тобто, якщо на них обох одночасно знаходится певна інформація, то з N користувачів N_n користувачів отримають цю інформацію з першоджерела, а $N_з$ – із сайту зловмисника;
- розподіл відвідувань сайтів користувачами в середньому за добу рівномірний.

Розглянемо складові формули (1) на прикладі і більш детальніше.

Вартість створення робота-аналізатора та підрахунок максимального прибутку

Вартість створення робота-аналізатора визначається як грошові ресурси витрачені замовником на апаратно-програмні засоби та виплати програмістам в процесі створення робота-аналізатора. $\boldsymbol{C}_n$ визначається як прибуток, отриманий від відвідання сторінки сайту зловмисника, що містить дану інфрмацію, одним користувачем (в даному випадку джерело доходу не конкретизується, оскільки статей доходів може бути багато, наприклад, розміщення реклами), помножений на кількість унікальних користувачів, що відвідали цю сторінку, до того моменту, як інформація вже стала неактуальною:

$$\boldsymbol{C}_n = \boldsymbol{C}_1 \cdot N \tag{2}$$

Проте дана формула є правильною тільки у випадку, коли інформація є актуальною протягом безмежного періоду часу. Проте, це тільки частковий випадок, який трапляється вкрай рідко. Тому необхідно враховувати актуальність розміщуваної інформації.

 ISSN 1562-9945

Розрахунок втрат прибутку, що викликані обмеженням на актуальність інформації по часу

Інформація є актуальною протягом певного періоду часу. Отже, якщо вона буде розміщена на сайті зловмисника після стікання цього періоду, вона не принесе прибутку. Для оперування поняттям актуальності K введемо показники: A – актуальність виконання підміни джерела.

$$A = \text{sign}((T_Д + t_a) - T_н), \tag{3}$$

де $T_Д$ – момент часу, коли інформація стала доступна на першоджерельному сайті, $T_н$ – момент часу, коли інформація з'явилась на іншому сайті, t_a – проміжок часу, протягом якої інформація є актуальною.

$$K = \frac{(T_Д + t_a) - T_н}{t_a},\ A < 0 \tag{4}$$

Як видно з формул, коефіцієнт актуальності показує, яку частину часу, протягом якого інформація є актуальною, тобто сайт зловмисника відвідають не N користувачів, а $N \cdot K_A$. Тож, оскільки, прибуток зловмисника прямопропорційно залежить від відвідуваності, то він становитиме $C_n \cdot K_A$. Підставивши (4) в (1) та (2) в (1) отримаємо наступне розгорнуте співвідношення:

$$\frac{(T_Д + t_a) - T_н}{t_a} \cdot C_1 \cdot N - \boldsymbol{C_p} < 0 \tag{5}$$

З вищесказаного очевидно, що для вирішення задачі необхідно виконати наступне:

- збільшити вартість створення робота аналізатора C_p, шляхом використання методів захисту, які ускладнюють алгоритм аналізу інформації, що підвищить кількість людиногодин, що були затрачені на розробку робота;
- зменшити коефіцієнт актуальності інформації, що розміщується на сайті зловмисника, використовуючи методи захисту інформації, які вимагають від алгоритму аналізу довготривалого виконання.

Слід зазначити, що на сьогоднішній день інформація в мережі Інтернет подається в різних формах, тому розглянемо різні методи захисту інформації в залежності від форм представлення інформації.

Форми представленя інформації в Інтернет

На сьогоднішній день веб сайти, що містять інфомрацію, яка доступна більшості відвідувачів (не потребує додаткового програмного забезпечення, окрім веб оглядача) можна розділити на наступні групи (рис. 2):

- текстова;
- графічна (зображення в форматах JPEG, GIF, PNG, SVG..);
- відео (формати FLV, MMPEG, OGG, AVI);

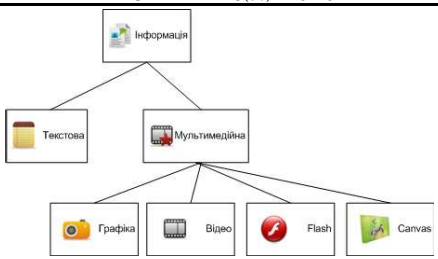

Рис. 2 – Класифікація груп представлення інформації на веб сайтах

- canvas (зображення, що створюються за допомогою мови JavaScript на клієнтській частині додатку);
- f lash (формат інтерактивних додатків SWF компанії Adobe).

Розглянемо способи захисту мультимедійної інформації.

Способи захисту мультимедійної інформації

Мультимедійна інформація може бути повністю захищена для виконання умов вирішення поставленої задачі. Графічну та відео-інформацію можна захистити шляхом додавання на зображення водяних знаків, що однозначно ідентифікують первинне джерело інформації.

Додатки, зображення, та анімація у форматі f lash піддаються декомпіляції, проте створення системи автоматичної обробки f lash є задачою надзвичайно складною, довготривалою, а в більшості випадків, неможливою. Крім того, додатки у цьому форматі являють собою клієнтську частину і не можуть бути використані без серверної частини. Зображення, створені за допомогою Canvas, зберігаються в оперативній пам’яті клієнта, і рідко використовують кешування на жорсткому диску, тому не є доступними для швидкого аналізу.

Розглянуті засоби захисту мультимедійної інформації - це часткові випадки вирішення даної задачі і не гарантують 100% захисту.

Розглянемо шляхи та способи захисту текстової інформації більш детальніше.

Захист текстової інформації

Захист текстової інформації є задачею набагато складнішою, адже така інформація доступна у відкритому вигляді, і легко піддається аналізу шляхом застосування регулярних виразів. В найкоротші строки така інформація у великих обсягах може бути структурована і розміщена на інформаційному ресурсі зловмисників.

В цьому випадку є декілька шляхів, які не дають повного захисту тексту від аналізу, проте дозволяють подати інформацію в такому вигляді, щоб швидкість аналізу інформації роботами була настільки низькою, що в результаті аналізу роботи отримали неактуальну інформацію або інформацію, коефіцієнт актуальності якої задовільняє умову (1).

 ISSN 1562-9945

Визначити чи є клієнт пошуковою машиною, веб огядачем, чи зловмисник, неможливо, тому що робот зловмисника може мати User-Agent будь-якого пошуковика чи веб-оглядача. Також важливою перешкодою є використання роботами таких сервісів типу "Хмара", як "Tor", які дозволяють йому при кожному новому запиті до сервера мати нову IP-адресу та User-Agent.

Можна застосовувати наступні способи захисту тексту, які базуються на таких представленнях текстової інформації, які не можна відрізнити від текстової людським оком, проте важко піддаються аналізу:

1. генерація окремих слів, словосполучень чи речень, що не містять ключові слова для пошуковиків, за допомогою клієнтських скриптів мовою JavasScript;
2. періодична (наприклад, 1 раз в добу) зміна шрифтів тексту;
3. конвертація окремих букв, слів, чи словосполучень, що не містять ключових слів в зображення чи f lash;
4. використання каскадних таблиць стилів для імітації тексту;
5. використання Canvas для виводу тексту.

Кожен з даних методів захисту текстової інформації ускладнює розробку та підвищує тривалість виконання алгоритму. Для кожного з методів можна обчислити тривалість виконання:

$$t_в = n_i \cdot t_i, \tag{6}$$

де n_i – кількість символів, t_i – час, що затрачається на вичленення та розпізнавання одного символу, що захищений i-м способом захисту. Загальний час, потрачений на виконання алгоритмів розпізнавання зменшує коефіцієнт актуальності інформації. Для обчислення загального часу, скористаємося слідуючим виразом:

$$T_в = \sum_{i=1}^{k} n_i \cdot t_i, \tag{7}$$

де n_i – кількість символів, що захищені i-м способом, t_i – час на розпізнавання символу, захищеного i-способом. Очевидно, що даний час являє собою величину:

$$T_в = T_н - T_Д \tag{8}$$

Підставивши вираз (8) в (5) отримаємо остаточний вираз (9) для розв'язання задачі захищенності інформації сайтів-першоджерел.

$$\frac{t_a - \sum\limits_{i=1}^{k} n_i \cdot t_i}{t_a} \cdot C_1 \cdot N - C_p > 0, \tag{9}$$

Висновок

Захищеність інформації на сайті від підміни джерела залежить від методів захисту інформації на сайті-першоджерелі. Задача вважається вирішеною при виконанні нерівності (9), яка може виконуватись завдяки використання методів захисту. Мультимедійна інформація може бути повністю захищеною, проте для текстової інформації можливі тільки ті методи захисту, які знижують актуальність інформації, що розміщується на сайті зловмисника. Для виконання нерівності (9) не обов'язково використовувати всі запропоновані методи захисту, достатньо використати тільки мінімальну кількість.

Література

1. Доповідь "Пошукові машини". П'ятнадцята науково-технічна конференція "Корпоративні бази даних-2010" , Москва, Calaf ia Consulting, 2010. http://www.citf orum.ru/internet/search/ips.shtml.
2. Рекомендація розробникам веб браузерів: "User Agent Accessibility Guidelines 1.0", W3C Consortium, 2002. http://www.w3.org/TR/WAI-USERAGENT/intro.html#user-control.
3. Довідковий центр компанії Google для вебмайстрів: http://www.google.com/support/webmasters/.
4. Дзінько Р.І., Лісовиченко О.І., Гордійчук А.М. Вибір технології для створення web-систем різного рівня складності // Адаптивні системи автоматичного управління. - 2009.- 15(35).- С. 22-30

Отримано 17.02.2010 р.