джерел дозволяє конкурентній розвідці діяти в рамках правового поля, але, при цьому, мати високу ефективність.

Можна констатувати, що чим швидше росте веб-простір, тим гірше воно охоплюється традиційними каталогами і пошуковим машинам. Через зростання кількості веб-сайтів і порталів, що використовують бази даних, динамічні системи керування контентом, появи нових версій форматів представлення інформації глибинний веб зростає дуже інтенсивно. З одного боку, Інтернет як величезне сховище збільшує об'єми інформації, доступної «в принципі», але з іншого боку – зростає інформаційний хаос, збільшується ентропія мережевого Інформаційного простору. Все менша частина інформаційних ресурсів стає доступною користувачам реально.

Провідні пошукові системи як і раніше намагаються знайти технічні можливості для індексації вмісту баз даних і отримати доступ до приватних веб-сайтів, проте, їх завдання об'єктивно розходяться із завданнями бізенс-аналітиків – орієнтація традиційних пошукових служб на масовий сервіс в даному випадку виправдана. Таким чином, ніша для систем пошуку в глибинному веб стає все ширшою.

УДК 004.02; 004.6
Ковальська М. – ст. гр. СКмз-61
Тернопільський національний технічний університет імені Івана Пулюя

ПРО ПОНЯТТЯ «ГЛИБИННОГО» ВЕБУ

Науковий керівник: ст. викладач Маєвський О.В.

Koval's'ka M.
Ternopil Ivan Pul`uy National Technical University

ON THE CONCEPT OF "DEEP" WEB

Supervisor: Majevskiy A.

Ключові слова: Веб-простір, «глибинний» веб
Keywords: Web space, "deep" web

Останні дослідження веб-простору показали, що доступні через традиційні інформаційно-пошукові системи більше трильйона веб-сторінок – це лише «видима частина айсберга».

Важливою проблемою є пошук інформації в «прихованому» або «глибинному» веб-просторі, де міститься незрівнянно більша кількість даних, потенціально цікавих для конкурентної розвідки, ніж у відкритій частині Інтернету.

Це, перш за все, динамічні веб-сторінки, інформація з численних баз даних, які можуть представляти великий інтерес для аналітичної роботи. До розряду «прихованого» веб відносяться і повнотекстові інформаційні системи типу LexisNexis або Factiva.

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

Відомо, що необхідної (в тому числі і для конкурентної розвідки) інформації в мережі Інтернет значно більше, ніж її охоплюють універсальні пошукові машини.

Передбачається, що на відміну від «відкритої» частини мережі Інтернет, «прихована» частина виявляється в сотні разів більш об'ємною.

Бізнес-аналітик часто стикається з ситуацією, коли йому відомо про існування в веб-просторі певного документу, але не може знайти його за допомогою традиційних пошукових систем, якими сьогодні можна вважати такі системи, як Google, Yahoo!, Bing, Яндекс, Рамблер або Мета. Однак, згадавши або знайшовши в закладках веб-адресу цього документа, він без проблем виходить на нього. Тобто у веб-просторі цей документ є, а знайти його звичайним способом не можна. Користувач зіткнувся з невидимим (invisible) для пошукових систем ресурсом.

Сукупність джерел у веб-просторі, недоступних користувачам традиційних пошукових систем, утворює так званий «глибинний веб» – поняття, введене Джил Ілсвортом (Jill Ellsworth) у 1994 році. Тобто під «глибинним вебом» (invisible web, deep web, hidden web) прийнято розуміти ту частину веб-простору, яка не індексується роботами (web-crawlers) пошукових систем. Використовуючи аналогію, інформація, будучи недоступною для пошуку, знаходиться «в глибині» (англ. – deep). При цьому не варто плутати «глибинний веб» з ресурсами, зовсім недоступними з мережі Інтернет – це «темний веб» (dark web). Деякі ресурси, доступ до яких відкритий лише для зареєстрованиххх користувачів, також відносяться до «глибинного вебу».

У 2000 році американська компанія BrightPlanet [1] опублікувала сенсаційну доповідь, в якому стверджувалося, що у веб-просторі в сотні разів більше сторінок, ніж їх вдалося проіндексувати найпопулярнішими на той час пошуковими системами. Компанія розробила програму LexiBot, яка дозволяє сканувати деякі динамічні веб-сторінки, що формуються з баз даних, і, запустивши її, отримала неочікувані дані. З'ясувалося, що в «глибинному» веб знаходиться в 500 разів більше документів, ніж доступно через пошукові системи. Звичайно, ці цифри неточні. Крім того, стало відомо, що середня сторінка «глибинного» веб на 27% компактніша середньої сторінки з видимої частини веб-простору.

Сьогодні ситуація змінилася, наприклад, провідні пошукові системи можуть індексувати документи, представлені у форматах, що містять текст. Звичайно, це, перш за все, .pdf, .rtf і .doc. У 2006 році Google запатентувала спосіб пошуку в «глибинному» веб: «Searching through content which is accessible through web-based forms» [2, 3]. На думку фахівців до «видимого» вебу відноситься лише 20-30% веб-простору.

Література.

1. DO ANYTHING WITH DATA // BrigthPlanet | Deep Web Intelligence. – URL: http://www.brightplanet.com/. – Last access: april 2016 y. – Title from screen.

2. Елена Салагаева. Deep web (invisible web, hidden web). / Салагаева Елена // On-line кабинет Александра Болдачева. – Режим доступа: http://www.boldachev.com/internet/deep-web-invisible-web-hidden-web. – Дата доступа: 2 апреля 2016 года. – Заглавие с экрана.

3. Searching through content which is accessible through web-based forms // Патент WO2006108069A3 – Searching through content which is accessible through web-based forms – Патенти Google. – Режим доступу: https://www.google.com/patents/WO2006108069A3?cl=ru. – Дата доступу: 2 квітня 2016 року. – Заголовок з екрану.